ESPRESSO: Robust Concept Filtering in Text-to-Image Models

Anudeep Das*, Vasisht Duddu*, Rui Zhang†§, N. Asokan*‡
*University of Waterloo, †Zhejiang University, ‡Aalto University
{anudeep.das, vasisht.duddu}@uwaterloo.ca, zhangrui98@zju.edu.cn, contact@sebszyller.com, asokan@acm.org

***Abstract*—Diffusion based text-to-image (T2I) models generate high fidelity images for given textual prompts. They are trained on large datasets scraped from the Internet, potentially containing *unacceptable concepts* (e.g., copyright infringing or unsafe). Retraining T2I models after filtering out unacceptable concepts in the training data is inefficient and degrades utility. Hence, we need concept removal techniques (CRTs) which are i) *effective* in preventing the generation of images with unacceptable concepts, ii) *utility-preserving* on acceptable concepts, and, iii) *robust* against evasion with adversarial prompts. None of the prior CRTs satisfy all these requirements simultaneously.**

We introduce ESPRESSO, the first *robust concept filter* based on Contrastive Language-Image Pre-Training (CLIP). We configure CLIP to identify unacceptable concepts in generated images using the distance of their embeddings to the text embeddings of *both* unacceptable and acceptable concepts. This simple change significantly improves robustness by restricting the adversary to add noise *only* along the vector connecting the embeddings of unacceptable and acceptable concepts. Further, fine-tuning ESPRESSO to separate embeddings of unacceptable and acceptable concepts, while preserving their pairing with image embeddings, improves utility. We evaluate ESPRESSO on eleven concepts to show that it is more *effective* (CLIP accuracy on unacceptable concepts in [0.00, 0.20] and FNR in [0.14, 0.53]) and more *robust* (CLIP accuracy on adversarial prompts for unacceptable concepts in [0.00, 0.40] and FNR in [0.00, 0.40]) than prior CRTs, while retaining *utility* (normalized CLIP score on acceptable concepts in [0.59, 0.98] and FPR in [0.01, 0.08]). Finally, we present theoretical bounds for certified robustness of ESPRESSO against adversarial prompts, and an empirical analysis.

1. Introduction

Diffusion based text-to-image (T2I) models have demonstrated a remarkable ability to generate high quality images from textual prompts [49], [50], [47]. These models are trained on large datasets of unfiltered content from the Internet [46], [53]. Due to their large capacity, T2I models memorize specific *concepts*, as seen in the generated images [28], [55], [6]. Some of these concepts, may be *unacceptable* for various reasons, such as copyright infringement (e.g., a movie character or celebrity), or inappropriateness (e.g., "nudity" or "violence") [16], [52], [21]. Retraining T2I models after filtering out unacceptable concepts is inefficient, only partially effective [34], [51], and compromises utility [16]. Hence, there is a need for *concept removal techniques* (CRTs) to minimize unacceptable concepts in generated images.

Ideally, CRTs should be *effective* in reducing the generation of unacceptable concepts while preserving the *utility* on all others, and *robust* to evasion with adversarial prompts. As we show in Section 6, none of the existing CRTs *simultaneously* satisfy these requirements: i) *filtering CRTs*, which detect unacceptable concepts, lack robustness ([48] and Section 6.2), ii) *fine-tuning CRTs* which modify T2I models, trade-off effectiveness and utility [16], [30], [21], [52], and may lack robustness [56], [60], [42], [66]. Designing a CRT meeting all the requirements is an open problem. Our goal is to design such a CRT.

We opt to use a filter as it will not alter the T2I model, thus reducing its impact on utility. We construct our filter using the Contrastive Language-Image Pre-Training (CLIP) [46], an essential component of T2I models. CLIP co-locates the embeddings of textual prompts and their corresponding images within a unified embedding space where similar concepts are closer. CLIP is pre-trained on a vast dataset which encodes a broad spectrum of concepts [46], making it a versatile choice for a filter, unlike specialized classifiers (e.g., [67]). However, relying on CLIP to identify unacceptable images by solely measuring the distance between the embeddings of a generated image, and an unacceptable concept, was shown not to be robust ([48]). Subsequently, several prior works have identified filtering as a viable direction which needs further exploration [48], [34].

We present a **robust** (CLIP-based) content filter, ESPRESSO, by configuring CLIP to identify unacceptable concepts in generated images using the distance of their embeddings to the text embeddings of *both* unacceptable and acceptable concepts. This simple change makes it harder to generate effective adversarial prompts: the adversary is restricted to adding noise in the direction of the acceptable concept. To restore any drop in utility stemming from this change, we fine-tune ESPRESSO to increase the separation between the text embeddings of unacceptable and acceptable concepts, while maintaining their pairing with their corresponding image embeddings. This fine-tuning approach is further supported by fine-tuning CRTs which show that such an optimization leads to better utility [30], [56], [15]. We claim the following main contributions: we present

§. Work done while visiting the Secure Systems Group, University of Waterloo.

1) ESPRESSO[1], the **first robust** (CLIP-based) content filter which identifies unacceptable concepts in generated images by measuring the distance of their embeddings to the text embeddings of *both* unacceptable and acceptable concepts (Section 4),
2) a comprehensive comparative evaluation (Section 5) of the fine-tuned variant of ESPRESSO with six state-of-the-art fine-tuning CRTs, and one filtering CRT, showing that it is **more effective** (CLIP accuracy on unacceptable concepts in [0.00, 0.20] and FNR in [0.14, 0.53]) and **more robust** (CLIP accuracy on adversarial prompts for unacceptable concepts in [0.00, 0.40] and FNR in [0.00, 0.40]) than prior CRTs, while **retaining utility** (normalized CLIP score on acceptable concepts in [0.59, 0.98] and FPR in [0.01, 0.08]). (Section 6), and
3) theoretical bounds for **certifying the robustness** of ESPRESSO, assuming a hypothetically strong adversary, empirical analysis of these bounds, and making the case that ESPRESSO is likely to be more robust in practice. (Section 7).

2. Background

We describe T2I models (Section 2.1), different CRTs (Section 2.2), and attacks against them (Section 2.3).

2.1. Diffusion based T2I Models

A diffusion based T2I model is a function $f : \mathtt{p} \rightarrow x$ which generates an image x for a given a textual prompt $\mathtt{p}$. It comprises two key components: an encoder (ϕ) which is used to incorporate the textual prompt in the image generation process, and a diffusion model (ϵ_θ) which is responsible for the generation of the image.

A popular encoder is CLIP, trained on a large dataset of image-text pairs, to map the embeddings of images and their corresponding text closer together in a joint text-image embedding space [46]. Given N images $\{x_j\}_{j=1}^N$ and their corresponding text prompts $\{\mathtt{p}_j\}_{j=1}^N$, the training data is $\mathcal{D} = \{(x_j, \mathtt{p}_j)\}_{j=1}^N$.

CLIP is trained to maximize the cosine similarity between the embeddings of a prompt $\mathtt{p}_j$ and its corresponding image x_j while minimizing the similarity between $\mathtt{p}_j$ and any other x_k for a $k \neq j$. We denote the cosine similarity as $\cos(\phi_p(\mathtt{p}_j), \phi_x(x_j))$, where $\phi_x(x_j)$ is the CLIP image embedding of the image x_j, and $\phi_p(\mathtt{p}_j)$ is the CLIP text embedding of the prompt $\mathtt{p}_j$. To achieve this, CLIP is trained using a contrastive loss function [64], [57], [32]:

$$\mathcal{L}_{\text{Con}}(\mathcal{D}) = -\frac{1}{N}\sum_{j=1}^{N}\log\frac{\exp(\cos(\phi_x(x_j), \phi_p(\mathtt{p}_j))/\tau)}{\sum_{k=1}^{N}\exp(\cos(\phi_x(x_j), \phi_p(\mathtt{p}_k))/\tau)}$$

where τ is the temperature parameter for scaling the predictions [46], [35].

Given access to a pre-trained encoder ϕ, the actual images in T2I models are generated by a diffusion model, ϵ_θ, parameterized by θ. During training of ϵ_θ, Gaussian noise is added to an initial image x_0 for T time steps to produce x_T, in a process known as the *forward diffusion process*. The noise is then iteratively removed to approximate the initial image $\tilde{x_0}$ in the *reverse diffusion process*. During inference, the reverse diffusion process generates an image from noise. Further, ϵ_θ can be conditioned with a textual prompt $\mathtt{p}$ to guide the generation of $\tilde{x_0}$ to match the description in $\mathtt{p}$. After generating $\phi_p(\mathtt{p})$, ϵ_θ is trained by minimizing the following loss function: $\mathcal{L} = \mathbb{E}_{\epsilon,\phi_p(\mathtt{p}),t}[||\epsilon - \epsilon_\theta(x_t, \phi_p(\mathtt{p}), t)||_2^2]$ for each time step t, and random Gaussian noise $\epsilon \sim \mathcal{N}(0, 1)$.

Several prominent T2I models (e.g., Stable Diffusion v1.4 (SDv1.4)) improve the efficiency of the diffusion process by conducting it in the embedding space of a variational autoencoder (VAE), rather than on images [49]. For a VAE decoder D, VAE encoder $\mathcal{E}$, and $z_t \in \mathcal{E}(x)$ as the latent representation of x in the VAE's latent space, ϵ_θ is trained by minimizing the following objective:
$\mathcal{L} = \mathbb{E}_{\epsilon,z_t,\phi_p(\mathtt{p}),t}[||\epsilon - \epsilon_\theta(z_t, \phi_p(\mathtt{p}), t)||_2^2]$ where $\epsilon \sim \mathcal{N}(0, 1)$. The final image is generated from the approximation, $\tilde{z_0}$, by passing it through D: $\tilde{x_0} = D(\tilde{z_0})$. Table 11 in Appendix A summarizes frequently used notations.

2.2. Concept Removal Techniques

The textual phrase for an acceptable concept is c^a, and for an unacceptable concept is c^u. For a given c^u, c^a is either the opposite (e.g., c^u = *violence* vs. c^a = *peaceful*) or a general category of a specific character/object/person (e.g., c^u = *R2D2* vs. c^a = *robot*). We discuss the selection of c^a for a given c^u in Section 5.3. An image x generated from a T2I model may either contain an unacceptable concept (referred to as x^u) or an acceptable one (referred to as x^a). Similarly, a text prompt $\mathtt{p}$ may contain a phrase for an acceptable concept ($\mathtt{p}^a$) or an unacceptable concept ($\mathtt{p}^u$). An example of an unacceptable prompt $\mathtt{p}^u$ containing an unacceptable concept c^u = *Captain Marvel*, is "*Captain Marvel* soaring through the sky". CRTs seek to thwart the generation of x^u by either fine-tuning the T2I model to suppress x^u, or using a classifier as a filter to detect x^u, and serve a replacement image instead. We first present six state-of-the-art fine-tuning CRTs:

Concept Ablation (CA) [30] fine-tunes the T2I model to minimize the KL divergence between the model's output for $\mathtt{p}^u$ and $\mathtt{p}^a$ to force the generation of x^a instead of x^u. Formally, they optimize the following objective function:

$$\mathcal{L}_{\text{CA}} = \mathbb{E}_{\epsilon,z_t,c^u,c^a,t}[w_t||\epsilon_\theta(z_t, \phi_p(\mathtt{p}^a), t).sg() - \epsilon_\theta(z_t, \phi_p(\mathtt{p}^u), t)||_2^2]$$

where $\epsilon \sim \mathcal{N}(0, 1)$, $z_t \in \mathcal{E}$, and w_t is a time-dependent weight, and $.sg()$ is the stop-gradient operation.

Forget-Me-Not (FMN) [68] minimizes the activation maps for c^u by modifying ϵ_θ's cross-attention layers. Further, fine-tuning ϵ_θ, instead of just the cross-attention layers, results in degraded utility.

Selective Amnesia (SA) [21] fine-tunes T2I models by adapting continuous learning techniques (elastic weight consolidation and generative replay) for T2I models to forget

1. We will open-source the code upon publication.

a concept. They optimize $\mathbb{P}(x|\theta^*, c^u)$, the probability of generating x given θ^* and c^u, where θ^* are the frozen parameters of the original T2I model:

$$\mathcal{L}_{\texttt{SA}} = -\mathbb{E}_{\mathbb{P}(x|\texttt{p})\mathbb{P}_f(c^u)}[\log\mathbb{P}(x|\theta^*, c^u)] - \lambda \sum_i \frac{M_i}{2}(\theta_i^* - \theta_i)^2 + \mathbb{E}_{\mathbb{P}(x|\texttt{p})\mathbb{P}_r(c^a)}[\log\mathbb{P}(x|\theta, c^a)]$$

where M is the Fisher information matrix over c^a and c^u, $\mathbb{P}_r$ and $\mathbb{P}_f$ are probabilities taken over the distributions of c^a and c^u respectively, and λ is a regularization parameter.

Erased Stable Diffusion (ESD) [16] fine-tunes the T2I model by modifying the reverse diffusion process to reduce the probability of generating x^u:

$$\epsilon_\theta(x_t, \phi_p(c^u), t) = \epsilon_{\theta^*}(x_t, t) + \eta(\epsilon_{\theta^*}(x_t, \phi_p(c^u), t) - \epsilon_{\theta^*}(x_t, t))$$

where $\eta > 0$ encourages the noise conditioned on c^u to match the unconditioned noise.

Unified Concept Editing (UCE) [19] fine-tunes the T2I model's cross-attention layers with a language model to minimize the influence of c^u, while keeping the influence of remaining concepts unchanged. Their optimization is:

$$\mathcal{L}_{\texttt{UCE}} = \sum_{c^u \in \mathcal{C}^u, c^a \in \mathcal{C}^a} ||W \times c^u - W^* \times c^a||_2^2 + \sum_{c \in S} ||W \times c - W^* \times c||_2^2$$

where W, W^* are the parameters of the fine-tuned and original *cross-attention layers in* ϵ_θ, $\mathcal{C}^u$ and $\mathcal{C}^a$ are the space of pre-defined unacceptable and acceptable concepts, and S is a set of concepts for which to preserve utility.

Safe diffusion (SDD) [26] fine-tunes the T2I model by encouraging the diffusion model noise conditioned on c^u to match the unconditioned noise, while minimizing the utility drop using the following objective function:

$$\mathcal{L}_{\texttt{SDD}} = \mathbb{E}_{\epsilon, z_t, c^u, t}[||\epsilon_\theta(z_t, \phi_p(c^u), t) - \epsilon_\theta(z_t, t).sg()||_2^2]$$

where $\epsilon \sim \mathcal{N}(0, 1)$ and $z_t \in \mathcal{E}(x)$.

We now describe two filtering CRTs:

Stable Diffusion Filter (SD-Filter) [46] is black-box and the design of the filter is not publicly available. However, Rando et. al. [48] hypothesize that it involves computing the cosine similarity between the embeddings of a generated image, x, and a pre-defined set of c^u. If the cosine similarity is greater than some threshold (Γ), then x has c^u. Formally, their filter $\texttt{F}_{SD}$ can be described as

$$\texttt{F}_{SD}(x) = \begin{cases} 1 & \cos(\phi_x(x), \phi_p(c^u)) > \Gamma \\ 0 & \text{otherwise} \end{cases}$$

where $\cos(\phi_x(x), \phi_p(c^u)) = \overline{\phi_x(x)} \cdot \overline{\phi_p(c^u)}$, and $\bar{\cdot}$ denotes normalization. Here $\texttt{F}_{SD}(x) = 0$ indicates x^a and $\texttt{F}_{SD}(x) = 1$ indicates x^u. Note that Γ varies with c^u.

Unsafe Diffusion (UD) [67] is the current state-of-the-art filtering CRT and outperforms SD-Filter [67]. UD trains a multi-headed neural network classifier on top of CLIP to identify x^u where each head classifies different c^u: *nudity, violence, disturbing, hateful*, and *political*. Their objective is given as $\texttt{F}_{\texttt{UD}}(x) = \text{MLP}(\phi_x(x))$ where MLP is a multi-layer perceptron, and $\texttt{F}_{\texttt{UD}} \in \{0, 1\}$.

2.3. Evading Concept Removal Techniques

An adversary ($\mathcal{A}dv$) may construct adversarial prompts ($\texttt{p}^{adv}$) to evade CRTs and force the T2I model to generate unacceptable images. We denote a dataset containing adversarial prompts and their corresponding images (x^{adv}) as $\mathcal{D}^u_{adv} = \{(x^{adv}_j, \texttt{p}^{adv}_j)\}_{j=1}^N$ where ideally x^{adv} will contain c^u. $\mathcal{A}dv$ is assumed to account for the target T2I model f, using a CRT. A dumb $\mathcal{A}dv$ who does not account for this is *naïve*.

$\mathcal{A}dv$'s objective is to construct $\texttt{p}^{adv}$ which can force a T2I model to output images with c^u while being semantically closer to acceptable prompts so that x^{adv} is incorrectly identified as acceptable while triggering the generation of unacceptable images. The existing attacks formulate the construction of $\texttt{p}^{adv}$ as an optimization problem using some reference x^u (or the difference between $\texttt{p}^a$ and $\texttt{p}^u$) as the ground truth. Different attacks vary in ways to solve this optimization. We present four state-of-the-art attacks below:

PEZ [60] constructs $\texttt{p}^{adv}$ by identifying text tokens by minimizing: $\mathcal{L}_{PEZ} = 1 - \cos(\phi_p(\texttt{p}^{adv}), \phi_x(x^u))$.

RingBell [56] identifies tokens for $\texttt{p}^{adv}$ by first calculating the average difference between the embedding vectors of $\texttt{p}^u$ and $\texttt{p}^a$: $\phi_p(\hat{\texttt{p}}) = \frac{1}{N}\sum_{i=1}^N\{\phi_p(\texttt{p}^u_i) - \phi_p(\texttt{p}^a_i)\}$ from a set of N prompts. Then, RingBell computes $\texttt{p}^{adv}$ by minimizing the distance of its text embedding to $(\phi_p(\texttt{p}^{init}) + \eta \cdot \phi_p(\hat{\texttt{p}}))$, formulated as: $\min_{\texttt{p}^{adv}} ||\phi_p(\texttt{p}^{adv}) - (\phi_p(\texttt{p}^{init}) + \eta \cdot \phi_p(\hat{\texttt{p}}))||^2$ where $\texttt{p}^{init}$ is an initial unacceptable prompt which is censored by a CRT. They solve this using a genetic algorithm.

SneakyPrompt [66] uses reinforcement learning to construct $\texttt{p}^{adv}$ specifically against filtering CRTs. Given an initial prompt $\texttt{p}^{init}$, the attack searches for tokens to update $\texttt{p}^{init}$ to form $\texttt{p}^{adv}$. The reward function is the cosine similarity between $\phi_x(x^{adv})$ and $\phi_p(\texttt{p}^{adv})$. Typically, they use $\texttt{p}^u$ as $\texttt{p}^{init}$ for better effectiveness and faster convergence.

CCE [42] uses textual-inversion to construct $\texttt{p}^{adv}$ [14]. $\mathcal{A}dv$ updates CLIP's vocabulary to include a new token "<s>" which, when included in $\texttt{p}^{adv}$, generates $\mathcal{A}dv$'s desired image. To optimize <s>, we find v, an embedding for <s>, corresponding to least loss $\mathcal{L}_{CCE} = \mathcal{L}_{MSE}(\epsilon, \epsilon_\theta(z_t, v, t))$, where $\epsilon \sim \mathcal{N}(0, 1)$, and $\mathcal{L}_{MSE}$ is the mean-squared-error.

All of these attacks are naïve except for CCE and SneakyPrompt, which account for different fine-tuning CRTs. In Section 5.1, we describe how we modify these naïve attacks to account for CRTs.

3. Problem Statement

Our goal is to design a CRT which can effectively detect unacceptable concepts in images generated from T2I models. We describe an adversary model, requirements of an ideal CRT, and limitations of prior works according to the requirements.

Adversary Model. We assume (blackbox) access to a deployed target T2I model (*f*) to which a client can send an input ($\mathtt{p}$) to generate an image (x). Further, *f* uses some CRT. The goal of the adversary ($\mathcal{A}dv$) is to force *f* to generate x^u despite the presence of a CRT to suppress it. We give an advantage to $\mathcal{A}dv$ by allowing whitebox access to a local identical copy of *f* with the CRT for use in designing attacks. This is reasonable as ϵ_θ and CLIP are publicly available. For filtering CRTs, we assume that $\mathcal{A}dv$ has whitebox access to the filter to use its loss function in designing the attacks.

Requirements An ideal CRT should be: **R1** *Effective* in minimizing the generation of x^u; **R2** *Utility-preserving*, maintaining the quality of acceptable images (for fine-tuning CRTs) or not blocking them (for filtering CRTs); and **R3** *Robust* to resist evasion with adversarial prompts ($\mathtt{p}^{adv}$).

Limitations of Prior Works. In Section 6, we empirically show the limitations of prior works by evaluating **R1**, **R2**, and **R3** of different CRTs. We provide a brief summary here. *Fine-tuning CRTs* modify *f* thereby explicitly creating a trade-off between effectiveness (**R1**) and utility (**R2**) [16], [30], [68], [21], [21], [19], [26]. Further, these CRTs do not consider robustness (**R3**) in their design and are susceptible to evasion by $\mathcal{A}dv$ with $\mathtt{p}^{adv}$. *Filtering CRTs* [46], [67] detect unacceptable concepts either in $\mathtt{p}$ (aka prompt filter) or in x and block them (a.k.a image filter). Since, they do not modify *f*, they can maintain utility without impacting effectiveness. Prior filtering approaches may not be accurate in detecting unacceptable concepts (impacts **R1**) [34]. They can also be easily evaded (harms **R3**) ([48] and Section 6.2). Further, the current state-of-the-art filter, UD [67], trains specialized classifiers for each concept on substantial data, which limits their generalization to new concepts.

4. ESPRESSO: Robust Filtering CRT

We present ESPRESSO, a robust concept filtering CRT which aims to satisfy the requirements from Section 3 and achieve a better trade-off than prior CRTs. ESPRESSO uses a classifier F to detect images and filter unacceptable concepts in generated images. Following SDv1.4, on detecting an unacceptable concept, ESPRESSO generates a replacement image [46], [48]. We identify CLIP as the natural choice for such a classifier as it is 1) pre-trained on a large dataset covering a wide range of concepts, and 2) used across many T2I models, and encodes similar information as seen in them. Hence, CLIP is a better choice for a filter than training specialized classifiers for each concept (e.g., [67]).

However, simply using CLIP for ESPRESSO is not sufficient as seen in Stable Diffusion's filter ($\mathtt{F}_{SD}$) [46]. $\mathtt{F}_{SD}$ thresholds the cosine similarity between the embeddings of x and each pre-defined unacceptable concept to identify x^u. Rando et al. [48] design adversarial prompts ($\mathtt{p}^{adv}$) to evade $\mathtt{F}_{SD}$. Since $\mathtt{F}_{SD}$ uses only the cosine similarity to c^u, it gives $\mathcal{A}dv$ the freedom to modify the prompt embeddings in *any direction*, while constructing $\mathtt{p}^{adv}$, to force a misclassification. We address this in ESPRESSO by (a) configuring CLIP's classification objective for effectiveness and robustness, and (b) fine-tuning to further improve utility.

Configuring CLIP's Classification Objective. Instead of using the cosine similarity to only c^u as in $\mathtt{F}_{SD}$, we configure the objective function of ESPRESSO for filtering x^u by using the cosine similarity to *both* c^u and c^a. Further, jointly optimizing for two embeddings yields better utility as observed in prior fine-tuning CRTs [30], [56], [15].

Given x, ESPRESSO checks the cosine similarity of $\phi_x(x)$ to $\phi_p(c^u)$ and $\phi_p(c^a)$. Formally, we define ESPRESSO as $\mathtt{F}(x, c^u, c^a)$

$$= \text{argmax}_{i, i\in a,u} \left\{ \frac{\exp(\cos(\phi_x(x), \phi_p(c_i))/\tau)}{\sum_{j\in\{a,u\}} \exp(\cos(\phi_x(x), \phi_p(c_j))/\tau)} \right\} \tag{1}$$

where $\tau = \frac{1}{100}$ is the default temperature parameter used in CLIP. Further, the cosine similarity is equivalent to the scaled Euclidean distance between the embeddings [35]:

$$(\overline{\phi_x(x)} - \overline{\phi_p(c)})^T(\overline{\phi_x(x)} - \overline{\phi_p(c)}) = 2 - 2\overline{\phi_x(x)} \cdot \overline{\phi_p(c)}$$

$$\implies \cos(x, c) = -\frac{1}{2}(\overline{\phi_x(x)} - \overline{\phi_p(c)})^T(\overline{\phi_x(x)} - \overline{\phi_p(c)}) + 1$$

Hence, by using CLIP as a filter, we project the image embeddings onto the vector between c^u and c^a and classify it based on its distance to both. This simple configuration change in CLIP's classification objective restricts $\mathcal{A}dv$ to adding noise only along the low-dimensional vector joining c^u and c^a: $\phi_p(c^u) - \phi_p(c^a)$ in the direction of c^a. This is more challenging for $\mathcal{A}dv$ to evade than $\mathtt{F}_{SD}$, as projecting to a lower-dimensional representation can improve robustness [59], [4]. Hence, we conjecture that this can help satisfy **R1** and **R3** (as shown later in Section 6).

Fine-tuning. The above configuration change may impact the utility on some concepts due to the close proximity of some c^u and c^a. To restore any drop in utility, we use fine-tuning. We use two different fine-tuning variants depending on the group of concepts.

For the case where c^u and c^a are opposites (e.g., Group-1 concepts where c^u = *violence* and c^a = *peaceful*), the above objective function might have a low correlation between $\phi_p(\mathtt{p})$ and the corresponding $\phi_x(x)$. This might result in poor utility. We use following objective function:

$$\begin{aligned}\mathcal{L}_{\text{ESPRESSO}} = \alpha_{\text{aa}}\mathcal{L}_{\text{Con}}(\mathcal{D}_{aa}) - \alpha_{\text{ua}}\mathcal{L}_{\text{Con}}(\mathcal{D}_{ua}) \\ +\alpha_{\text{uu}}\mathcal{L}_{\text{Con}}(\mathcal{D}_{uu}) - \alpha_{\text{au}}\mathcal{L}_{\text{Con}}(\mathcal{D}_{au}) \\ +\alpha_{\text{uu-t}}\mathcal{L}_{MSE}(\phi_p(\mathcal{P}^u), \phi_p(\mathcal{P}^a))\end{aligned} \tag{2}$$

where $\mathcal{D}_{aa} = \{({x^a}_j, \mathtt{p}^a_j)\}_{j=1}^N$, $\mathcal{D}_{au} = \{({x^a}_j, \mathtt{p}^u_j)\}_{j=1}^N$, $\mathcal{D}_{ua} = \{({x^u}_j, \mathtt{p}^a_j)\}_{j=1}^N$, $\mathcal{D}_{uu} = \{({x^u}_j, \mathtt{p}^u_j)\}_{j=1}^N$, $\mathcal{P}^u = \{\mathtt{p}^u_j\}$, and $\mathcal{P}^a = \{\mathtt{p}^a_j\}$, and $\alpha_.$ are regularization hyperparameters. We assign equal weight to each of the loss terms, thus choosing $\alpha_{(.)} = 1$. The above objective function encourages the CLIP embeddings of x^u and $\mathtt{p}^u$, and x^a and $\mathtt{p}^a$, to be closer together while increasing the distance between x^u and $\mathtt{p}^a$, and x^a and $\mathtt{p}^u$, respectively.

In contrast, for concepts in Group-2 and 3, we fine-tune F to maximize the distance between $\phi_p(\mathtt{p}^u)$ and $\phi_p(\mathtt{p}^a)$, and hence, minimize the following loss:

$$L_{\text{ESPRESSO}} = -||\overline{\phi_p(\mathtt{p}^u) - \phi_p(\mathtt{p}^a)}||_2 \tag{3}$$

We use prompts ($\mathtt{p}^u$ and $\mathtt{p}^a$) for fine-tuning in Equation 3 and Equation 2 instead of only concepts (c^u and c^a): $\mathtt{p}^u$ and $\mathtt{p}^a$ already contain c^u and c^a, and provide more context.

During fine-tuning of ESPRESSO, there is a trade-off between effectiveness and utility which is inherent to all other fine-tuning CRTs as well. We subject our fine-tuning to the constraint that achieved the highest effectiveness for the least drop in utility. We empirically show the benefit of fine-tuning in Appendix B: Table 12).

5. Experimental Setup

We use Stable Diffusion v1.4 [49] and its default configuration as the target T2I model following prior CRTs [16], [30], [68], [21], [19], [26]. We describe the attack baselines against CRTs (Section 5.1), metrics (Section 5.2), different concepts used for evaluation (Section 5.3), and pipeline used for evaluating CRTs (Section 5.4).

5.1. Revisiting Attack Baselines

We consider different state-of-art attacks from literature (Section 2.3). We modify existing naïve attacks (indicated with "+"), Ring-a-Bell, and PEZ, to account for CRTs. All CRTs use some variant of the following optimization: detach c^u from $\mathtt{p}$ such that the generated image is far from c^u and closer to some c^a. On the other hand, attacks against T2I models described in Section 2.3, design $\mathtt{p}^{adv}$ such that x^{adv} is closer to c^u and far from c^a. Hence, to design an effective attack which accounts for such CRTs, in addition to the attacks' original objectives, we minimize the loss between the embeddings of $\mathtt{p}^{adv}$ and c^a while increasing the loss with the embeddings of c^u. This, in turn, will move $\phi_p(\mathtt{p}^{adv})$ closer to $\phi_p(c^a)$ and farther from $\phi_p(c^u)$. We modify the attacks to construct $\mathtt{p}^{adv}$ using following loss:

$$\mathcal{L}_{att+} = \mathcal{L}_{att} - \alpha_u \mathcal{L}_{MSE}(\phi_p(c^u), \phi_p(\mathtt{p}^{adv})) + \alpha_a \mathcal{L}_{MSE}(\phi_p(c^a), \phi_p(\mathtt{p}^{adv})) \quad (4)$$

where $att \in \{RingBell, PEZ\}$ when CRT $\in$ {CA, FMN, SA, ESD, UCE, SDD, UD, ESPRESSO }, and $\mathcal{L}_{att}$ is from Section 2.2. We assign equal weight to all loss terms and use $\alpha_u = \alpha_a = 1$. Recall from Section 2.3 that CCE already accounts for different fine-tuning CRTs. For filtering CRTs (UD and ESPRESSO), we modify CCE using Equation 4 and call this attack CCE+.

Finally, typographic attack [41] against CLIP is where text characters are superimposed onto an (unrelated) image to fool CLIP by forcing it to focus on this text instead of the image. We turn this into an attack against CRTs by superimposing c^a at the bottom of x^u. Using the resulting adversarial images, we use PEZ+ to find their corresponding $\mathtt{p}^{adv}$. We call this attack *Typo+*.

5.2. Metrics

We now describe the metrics to evaluate each of the requirements. We assume access to a *reference CLIP*, separate from f with the CRTs. The reference CLIP is a Stable Diffusion v1.4 model, a standard assumption in prior work [30], [16], [19], [26].

R1 (Effectiveness). Depending on the CRT, we use:

- **CLIP accuracy** [22], [30] for fine-tuning CRTs. This is the cosine similarity (divided by temperature parameter $\tau' = \frac{1}{100}$ of the reference CLIP) between the embeddings of the generated image x with the embeddings of c^u and c^a from the reference CLIP. This outputs the likelihood of predicting c^u. Hence, CLIP accuracy should be low (ideally zero) for effective concept removal. Formally, it is:

$$\frac{\exp(\cos(\tilde{\phi_x}(x), \tilde{\phi_p}(c^u)/\tau'))}{\exp(\cos(\tilde{\phi_x}(x), \tilde{\phi_p}(c^u)/\tau')) + \exp(\cos(\tilde{\phi_x}(x), \tilde{\phi_p}(c^a)/\tau'))}$$

where $\tilde{\phi_p}$ and $\tilde{\phi_x}$ are embeddings from the reference CLIP. For filtering CRT, if x^u is detected, a replacement image is generated. Here, we calculate the CLIP accuracy on the final set of images after filtering.

- **False Negative Rates (FNR) [67], [48]** for filtering CRTs is the fraction of images with c^u which are not blocked. It should be low (ideally zero).

R2 (Utility). Depending on the CRT, we use:

- **Normalized CLIP score** for fine-tuning CRTs is is the ratio of cosine similarity between $\phi_x(x)$ and $\phi_p(\mathtt{p})$ from f, compared to that from a reference CLIP as a baseline which is assumed to have the maximum achievable CLIP score. Formally, $\frac{\cos(\phi_x(x), \phi_p(\mathtt{p}))}{\cos(\tilde{\phi_x}(x), \tilde{\phi_p}(\mathtt{p}))}$. Normalized CLIP score should be high (ideally one) for high utility. For the normalized CLIP score to be greater than one, the fine-tuned T2I model should generate images that match their prompts more closely than the original T2I model. This is unlikely to occur in practice as seen in our evaluation (c.f. Table 6). This metric is different from standard CLIP score from prior work [22], [30], [26], [19] which measures the cosine similarity between $\phi_x(x)$ and $\phi_p(\mathtt{p})$ from f. We did this to compare the utility of the T2I with the CRT to that of the T2I without the CRT. Hence, a perfect match between them will give a normalized CLIP score of one.
- **False Positive Rates (FPR) [67], [48]** for filtering CRTs is the fraction of images without c^u which are blocked. It should be low (ideally zero).

R3 (Robustness). We use the same metrics as effectiveness: CLIP accuracy to evaluate fine-tuning CRTs and compare ESPRESSO with them, and FNR for filtering CRTs.

5.3. Concept Types

We use the same set of unacceptable concepts as in prior work [30], [42], [26], categorizing them into three groups:

- Group-1 covers inappropriate concepts such as *nudity* (e.g., female breasts or male/female genitalia), *violence* (e.g., bloody scenes, fighting, burning, hanging, weapons, and wars), *disturbing* (e.g., distorted faces, bodies, human flesh, or bodily fluids), and *hateful* (e.g., defamatory racial depiction, harmful stereotypes, or Holocaust scenes).
- Group-2 covers copyright-infringing concepts such as *Grumpy Cat, Nemo, Captain Marvel, Snoopy*, and *R2D2*.

• Group-3 covers (unauthorized) use of personal images for celebrities such as *Taylor Swift, Angelina Jolie, Brad Pitt*, and *Elon Musk*.

5.4. Pipeline for Evaluating CRTs

We now describe the pipeline which includes identifying acceptable concepts (step 1), generation of datasets (step 2), training filters or fine-tuning T2I models with CRTs (step 3), and validating acceptable concepts (step 4), and evaluating different CRTs (step 5). We present an overview of the pipeline in Figure 1 and describe each of the steps below.

Step 1: Identifying Acceptable Concepts. A good choice of c^a is one which effectively steers a T2I model away from generating x^u while maintaining utility on other concepts. Hence, the choice of c^a can impact the overall effectiveness and utility of CRTs. We select c^a for a given c^u such that it is either opposite to c^u (Group-1) or is a semantic generalization of c^u so as to avoid infringing copyrights (Group-2,3). For c^u in Group-1, we consider multiple alternative synonyms for c^a from which we choose the best possible candidate by measuring effectiveness and utility on a validation dataset (c.f. Step 4). For c^u in Group-2,3, we use c^a chosen by prior works [30], [21] for a fair comparison. We indicate them in the format "$c^u \rightarrow c^a$":

• **Group-1**: c^a is the opposite of c^u. We consider the following choices of c^a: *nudity* $\rightarrow$ {*clothed* and *clean*}; *violence* $\rightarrow$ {*peaceful*, *nonviolent*, and *gentle*}; *disturbing* $\rightarrow$ {*pleasing*, *calming*, and *soothing*}; *hateful* $\rightarrow$ {*loving*, *compassionate*, and *kind*}.

• **Group-2**: c^a is the type of c^u, following prior works [30], [21], [16]. For instance, *Grumpy Cat* $\rightarrow$ *cat*, *Nemo* $\rightarrow$ *fish*, *Captain Marvel* $\rightarrow$ *female superhero, Snoopy* $\rightarrow$ *dog*, and *R2D2* $\rightarrow$ *robot*.

• **Group-3**: c^a is the sex of c^u, following prior works [21], [68], [26]: {*Taylor Swift, Angelina Jolie*} $\rightarrow$ *woman* and {*Brad Pitt, Elon Musk*} $\rightarrow$ *man*.

Step 2: Generate and Split Datasets. We describe the generation of train, validation, and test datasets.

Train Datasets. We summarize the different training/fine-tuning dataset configurations across CRTs in Table 1.

Fine-tuning CRTs. Different fine-tuning CRTs use different dataset configurations to ensure that **R1** and **R2** are met. We use the exact same configuration as described in each of the prior works to ensure that they satisfy the requirements claimed in their papers. While the configuration across different CRTs vary, the configuration for a specific CRT across different concepts is the same, in accordance with the instructions in the respective works. We briefly summarize them in Table 1. CA uses acceptable prompts which are generated from ChatGPT such that they contain c^a, and 1 image per prompt. ESD and SDD both use an unacceptable concept, and SDD additionally uses 10 images corresponding to c^u. All images are generated using SD v1.4. FMN uses unacceptable prompts of the form "An image of $\{c^u\}$", and 1 image per prompt. For all the fine-tuning CRTs, we use their publicly available code for training to ensure the configuration is same as reported in their papers [31], [17], [69], [18], [20], [27], [45].

TABLE 1. TRAINING DATASET CONFIGURATIONS REQUIRED BY CRTS.

CRT	Configuration Requirements
CA [30]	200 unacceptable prompts ($\mathtt{p}^a$) from ChatGPT with one image/prompt from SDv1.4
ESD [16]	Unacceptable concept (c^u)
FMN [68]	8 unacceptabe prompts ($\mathtt{p}^u$) and one image/prompt from SDv1.4
SDD [26]	Unacceptable concept (c^u) and 10 corresponding images
SA [21]	Acceptable concept (c^a) and 1000 corresponding images, and 6 unacceptable prompts ($\mathtt{p}^u$)
UCE [19]	Unacceptable and acceptable concepts (c^u and c^a)
UD [67]	776 total images with 580 acceptable images (x^a), and 196 unacceptable images (x^u): nudity (48), violence (45), disturbing (68), and hateful (35)
ESPRESSO	10 unacceptable prompts ($\mathtt{p}^u$) and acceptable prompts ($\mathtt{p}^a$) using ChatGPT with one image/prompt from SDv1.4

Filtering CRTs. We train $\mathtt{F}_{\mathtt{UD}}$ using its exact dataset configuration (Table 1) and code [67]. For ESPRESSO, we follow the template by Kumari et al. [30] to generate 10 unacceptable and acceptable prompts using ChatGPT, with one image per prompt from SDv1.4. For Group-2 concepts, we randomly select 10 ChatGPT-generated prompts from the datasets provided by Kumari et al. [30]. Our choice for having a small amount of training data is inspired from prior works on poisoning attacks which show that little data is required to modify CLIP [64], [7]. Our results also show that using such little data is indeed sufficient to meet **R1**, **R2**, and **R3** (c.f. Section 6).

Validation Datasets. We consider the following validation datasets ($\mathcal{D}_{val}$): we denote dataset with acceptable prompts ($\mathtt{p}^a$) and corresponding images (x^a) as $\mathcal{D}^u_{val}$; and dataset with acceptable prompts ($\mathtt{p}^a$) and corresponding images (x^a) as $\mathcal{D}^a_{val}$. For **R1**, we use $\mathcal{D}^u_{val}$ by generating 10 unacceptable prompts using ChatGPT, and generate 5 images per prompt using SD v1.4. For **R2**, we use $\mathcal{D}^a_{val}$ by randomly choosing 100 non-overlapping acceptable prompts from the COCO 2014 dataset [36], and 1 image per prompt, generated from SD v1.4. We summarize them in Table 2.

TABLE 2. VALIDATION DATASET CONFIGURATION.

Concepts (Requirement)	Data	Configuration
Group-1 (**R1**)	$\mathcal{D}^u_{val}$	10 unacceptable prompts ($\mathtt{p}^u$) from ChatGPT and 5 images (x^u) per prompt from SDv1.4
Group-2 (**R1**)	$\mathcal{D}^u_{val}$	10 unacceptable prompts ($\mathtt{p}^u$) from ChatGPT and 5 images (x^u) per prompt from SDv1.4
Group-3 (**R1**)	$\mathcal{D}^u_{val}$	10 unacceptable prompts ($\mathtt{p}^u$) from ChatGPT and 5 images (x^u) per prompt from SDv1.4
All (**R2**)	$\mathcal{D}^a_{val}$	100 acceptable prompts ($\mathtt{p}^a$) and 1 image (x^a) per prompt from the COCO 2014 dataset

Test Datasets. We consider test datasets ($\mathcal{D}_{te}$): we denote the dataset with acceptable prompts ($\mathtt{p}^a$) and corresponding images (x^a) as $\mathcal{D}^a_{te}$; and with unacceptable prompts ($\mathtt{p}^u$) and corresponding images (x^u) as $\mathcal{D}^u_{te}$.

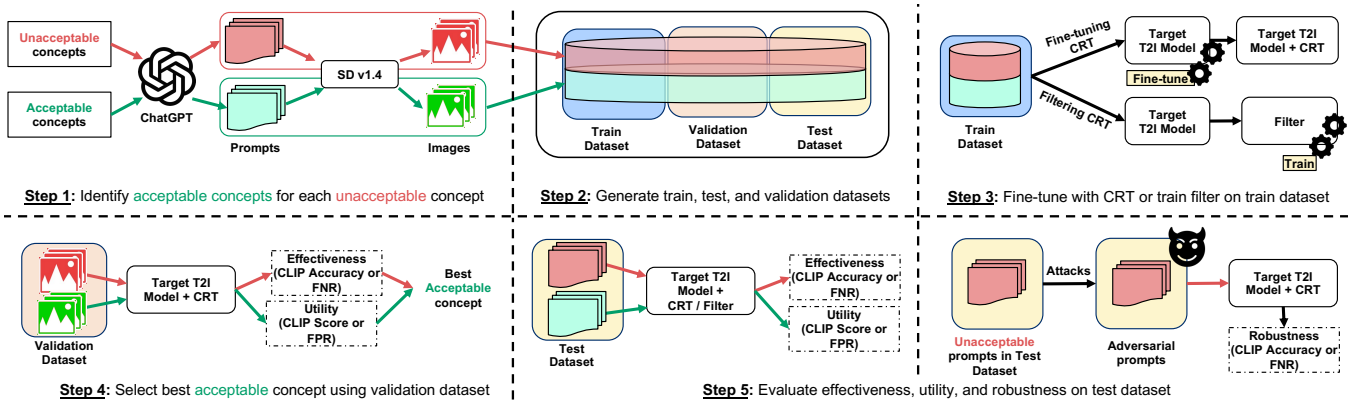


Figure 1. **Step 1** → Identify some possible c^a for each c^u. **Step 2** → Generate datasets containing prompts and corresponding images for both c^a and c^u for train ($\mathcal{D}_{tr}$), test ($\mathcal{D}_{te}$), and validation ($\mathcal{D}_{val}$). **Step 3** → Use $\mathcal{D}_{tr}$ to fine-tune T2I model with CRTs and train the filtering classifier. **Step 4** → Evaluate effectiveness and utility on $\mathcal{D}_{val}$ and choose the best candidate for c^a. **Step 5** → Evaluate using $\mathcal{D}_{te}$: for effectiveness (**R1**), use CLIP accuracy or FNR on unacceptable prompts ($\mathcal{D}^u_{te}$), utility (**R2**) on acceptable prompts ($\mathcal{D}^a_{te}$), and robustness (**R3**) using CLIP score or FPR on $\mathcal{D}^u_{adv}$ which includes $\mathtt{p}^{adv}$ constructed using attacks on $\mathcal{D}^u_{te}$. [Color Scheme and Component Descriptions] → Prompts, images and arrows corresponding to unacceptable (acceptable) are indicated in red (green). Metrics are indicated in dashed boxes.

TABLE 3. EVALUATION DATASET CONFIGURATION.

Concepts (Requirement)	Data	Configuration
Group-1 (**R1**, **R3**)	$\mathcal{D}^u_{te}$	I2P Dataset [52] of unacceptable prompts: for *nudity*: 449 prompts, *violence*: 758, *disturbing*: 857, *hateful*: 235
Group-2 (**R1**, **R3**)	$\mathcal{D}^u_{te}$	10 unacceptable prompts ($\mathtt{p}^u$) from ChatGPT and 20 images (x^u) per prompt from SDv1.4
Group-3 (**R1**, **R3**)	$\mathcal{D}^u_{te}$	10 unacceptable prompts ($\mathtt{p}^u$) from ChatGPT and 20 images (x^u) per prompt from SDv1.4
All (**R2**)	$\mathcal{D}^a_{te}$	200 acceptable prompts ($\mathtt{p}^a$) from COCO 2014 dataset

For evaluating effectiveness (**R1**) and robustness (**R3**), we use $\mathcal{D}^u_{te}$ which is generated as follows: For Group-1, we use an independent benchmark dataset, Inappropriate Image Prompts (I2P), containing prompts likely to generate unsafe images [52]. It covers all four Group-1 concepts. For *nudity*, I2P includes a "nudity percentage" attribute. We choose all unacceptable prompts with a nudity percentage > 10, resulting in a total of 300 unacceptable prompts, in accordance with prior works [16], [19]. UD [67] also used *political* as a concept, which is excluded from our evaluation since it is not a part of I2P. Note that I2P prompts do not explicitly contain c^u, *i.e., the words* nude, violent, hateful, *or* disturbing *are not explicitly included in the prompts*. For Group-2 and Group-3 concepts, there are no standard benchmark datasets. Hence, we generate the dataset by following prior works [30], [16] to obtain 200 unacceptable images generated from SDv1.4 from 10 unacceptable prompts generated from ChatGPT. For **R2**, we use $\mathcal{D}^a_{te}$ which includes the COCO 2014 test dataset with 200 randomly chosen acceptable prompts. We summarize the datasets in Table 3.

Step 3: Fine-tune T2I model with CRTs or train filter. Using $\mathcal{D}_{tr}$, we fine-tune T2I models using different fine-tuning CRTs whose optimizations are described in Section 2.3. We use six fine-tuning CRTs (CA, FMN, SA, ESD, UCE, and SDD). For filtering CRT, we consider one baseline (UD) and do not compare with SD-Filter as UD outperforms it. For UD, we train the filtering classifier on their dataset. We fine-tune ESPRESSO on $\mathcal{D}_{tr}$ and use the CLIP L-patch-14 [1] due to its popularity, which is also the default text encoder with Stable Diffusion v1.4. However, other variants of CLIP are also applicable.

TABLE 4. SUMMARY OF UNACCEPTABLE CONCEPTS (c^u), ACCEPTABLE CONCEPTS (c^a), AND WHETHER WE USE THEM TO EVALUATE FINE-TUNING CRTS, FILTERING CRTS, OR BOTH.

Type	$c^u \rightarrow c^a$	CRT (s)
Group-1	*Nudity* → *Clean*	Both
	Violence → *Peaceful*	Both
	Disturbing → *Pleasing*	Filtering
	Hateful → *Loving*	Filtering
Group-2	*Grumpy Cat* → *Cat*	Fine-tuning
	Nemo → *Fish*	Fine-tuning
	Captain Marvel → *Female Superhero*	Fine-tuning
	Snoopy → *Dog*	Fine-tuning
	R2D2 → *Robot*	Fine-tuning
Group-3	*Taylor Swift* → *Woman*	Fine-tuning
	Angelina Jolie → *Woman*	Fine-tuning
	Brad Pitt → *Man*	Fine-tuning
	Elon Musk → *Man*	Fine-tuning

Step 4: Select best c^a using $\mathcal{D}_{val}$. For concepts in Group-1, we evaluate the effectiveness (**R1**) and utility (**R2**) for different candidates of c^a on $\mathcal{D}^u_{val}$ and $\mathcal{D}^a_{val}$ respectively. We found that the following concepts gave the best results: *nudity* → *clean*, *violence* → *peaceful*, *disturbing* → *pleasing*, and *hateful* → *loving*. For *nudity*, we eliminated *clothed* as it blocked images with minimal exposed skin despite being acceptable [67], [52].

We use concepts in Group-1 following prior work [67] to evaluate filtering CRTs. We use *nudity* and *violence* from

TABLE 5. **R1** EFFECTIVENESS: COMPARISON WITH FINE-TUNING CRTS USING CLIP ACCURACY ON *unacceptable* PROMPTS (LOWER IS BETTER). WE USE RED IF ACCURACY IS >50; BLUE IF ACCURACY IS BETWEEN 25-50; GREEN IF ACCURACY IS <25.

CRT	Concepts										
	Nudity (I2P)	Violence (I2P)	Grumpy Cat	Nemo	Captain Marvel	Snoopy	R2D2	Taylor Swift	Angelina Jolie	Brad Pitt	Elon Musk
CA [30]	0.82 ± 0.01	0.78 ± 0.01	0.00 ± 0.00	0.02 ± 0.00	0.40 ± 0.05	0.06 ± 0.05	0.13 ± 0.02	0.73 ± 0.05	0.83 ± 0.02	0.86 ± 0.04	0 .64 ± 0.03
FMN [68]	0.83 ± 0.01	0.64 ± 0.04	0.34 ± 0.02	0.61 ± 0.01	0.82 ± 0.03	0.16 ± 0.00	0.89 ± 0.03	0.45 ± 0.02	0.59 ± 0.06	0.79 ± 0.04	0.56 ± 0.22
SA [21]	0.69 ± 0.09	0.69 ± 0.00	0.16 ± 0.00	0.87 ± 0.04	0.93 ± 0.02	0.55 ± 0.07	0.98 ± 0.01	0.82 ± 0.05	0.49 ± 0.04	0.63 ± 0.05	0.75 ± 0.04
ESD [16]	0.62 ± 0.06	0.63 ± 0.01	0.28 ± 0.06	0.64 ± 0.06	0.37 ± 0.04	0.20 ± 0.02	0.41 ± 0.04	0.11 ± 0.02	0.29 ± 0.05	0.17 ± 0.02	0.17 ± 0.02
UCE [19]	0.70 ± 0.01	0.71 ± 0.01	0.05 ± 0.00	0.43 ± 0.00	0.04 ± 0.00	0.03 ± 0.00	0.40 ± 0.01	0.02 ± 0.01	0.06 ± 0.00	0.05 ± 0.00	0.10 ± 0.01
SDD [26]	0.57 ± 0.02	0.55 ± 0.02	0.20 ± 0.02	0.20 ± 0.03	0.41 ± 0.03	0.37 ± 0.03	0.39 ± 0.02	0.05 ± 0.02	0.06 ± 0.01	0.04 ± 0.01	0.06 ± 0.01
ESPRESSO	0.15 ± 0.06	0.20 ± 0.05	0.00 ± 0.01	0.10 ± 0.02	0.03 ± 0.01	0.08 ± 0.02	0.00 ± 0.00	0.02 ± 0.00	0.03 ± 0.00	0.00 ± 0.00	0.03 ± 0.00

TABLE 6. **R2** (UTILITY): COMPARISON WITH FINE-TUNING CRTS USING NORMALIZED CLIP SCORES ON *acceptable* PROMPTS (HIGHER IS BETTER). WE USE RED IF SCORE IS BETWEEN 50-70, BLUE IF BETWEEN 70-90; GREEN IF >90.

CRT	Concepts										
	Nudity	Violence	Grumpy Cat	Nemo	Captain Marvel	Snoopy	R2D2	Taylor Swift	Angelina Jolie	Brad Pitt	Elon Musk
CA [30]	0.93 ± 0.00	0.93 ± 0.00	0.93 ± 0.00	0.93 ± 0.00	0.93 ± 0.00	0.93 ± 0.00	0.93 ± 0.00	0.93 ± 0.00	0.93 ± 0.00	0.93 ± 0.00	0.93 ± 0.00
FMN [68]	0.79 ± 0.00	0.79 ± 0.00	0.79 ± 0.00	0.79 ± 0.00	0.79 ± 0.00	0.79 ± 0.00	0.79 ± 0.00	0.79 ± 0.00	0.79 ± 0.00	0.79 ± 0.00	0.79 ± 0.00
SA [21]	0.79 ± 0.00	0.79 ± 0.00	0.79 ± 0.00	0.79 ± 0.00	0.79 ± 0.00	0.79 ± 0.00	0.79 ± 0.00	0.79 ± 0.00	0.79 ± 0.00	0.79 ± 0.01	0.79 ± 0.00
ESD [16]	0.82 ± 0.01	0.82 ± 0.00	0.82 ± 0.01	0.82 ± 0.00	0.82 ± 0.01	0.82 ± 0.01	0.82 ± 0.01	0.82 ± 0.01	0.82 ± 0.01	0.82 ± 0.01	0.82 ± 0.01
UCE [19]	0.96 ± 0.02	0.96 ± 0.00	0.97 ± 0.03	0.96 ± 0.00	0.96 ± 0.00	0.97 ± 0.02	0.96 ± 0.00	0.97 ± 0.02	0.97 ± 0.01	0.96 ± 0.03	0.97 ± 0.00
SDD [26]	0.86 ± 0.00	0.75 ± 0.00	0.93 ± 0.00	0.86 ± 0.00	0.82 ± 0.00	0.82 ± 0.00	0.64 ± 0.00	0.82 ± 0.00	0.82 ± 0.00	0.61 ± 0.00	0.89 ± 0.00
ESPRESSO	0.94 ± 0.08	0.59 ± 0.11	0.98 ± 0.04	0.37 ± 0.02	0.37 ± 0.03	0.66 ± 0.03	0.66 ± 0.02	0.98 ± 0.02	0.98 ± 0.03	0.98 ± 0.01	0.97 ± 0.01

Group-1, and all the concepts from Group-2 and Group-3 following prior works [30], [21], [68], [26] to evaluate fine-tuning CRTs. Hence, we compare ESPRESSO with each CRT category separately using concepts they were evaluated on. We summarize the final concepts for evaluation in Table 4.

Step 5: Evaluating R1, R2, and R3 on $\mathcal{D}_{te}$. We evaluate effectiveness (**R1**) on $\mathcal{D}^u_{te}$, utility (**R2** on $\mathcal{D}^a_{te}$, and robustness (**R3**) $\mathcal{D}^u_{adv}$ which is generated by running different attacks $\mathcal{D}^u_{te}$ to get $\mathtt{p}^{adv}$. In all cases, we only pass the prompts from the dataset to T2I models and compute different metrics described in Section 5.2.

6. Evaluation

We evaluate ESPRESSO with fine-tuning CRTs (Section 6.1), and then filtering CRT (Section 6.2).

6.1. Comparison with Fine-tuning CRTs

R1 Effectiveness. We report CLIP accuracy in Table 5 for all eleven concepts on $\mathcal{D}^u_{te}$. All CRTs show poor accuracy on *nudity* and *violence*. We attribute this to fine-tuning CRTs being sensitive to the input prompts [43], [65], [37]. Hence, they perform poorly for *nudity* and *violence* which do not explicitly include c^u. In comparison to other CRTs, ESPRESSO consistently maintains high accuracy on the I2P benchmark dataset as it classifies the generated images and does not depend on the prompts. Further, ESPRESSO satisfies **R1** using only 20 prompts and corresponding images for training. This is consistent with results on CLIP poisoning where changing the performance of CLIP only required a small number of poison samples [7], [64].

ESD, UCE, and SDD have better accuracy compared to the other three fine-tuning CRTs. This could be attributed to the similar optimizations used by ESD and SDD for fine-tuning the T2I model: both fine-tune ϵ_θ, conditioned on c^u, to match original ϵ_θ without any prompt to reduce the influence of c^u on the output. For UCE, higher effectiveness can be attributed to directly removing the influence of c^u from the T2I model parameters. *Overall,* ESPRESSO *is more effective than other fine-tuning CRTs.*

R2 Utility. We report normalized CLIP scores in Table 6 on $\mathcal{D}^a_{te}$. All the fine-tuning CRTs perform well across all concepts (either blue or green). This could be attributed to explicitly fine-tuning for utility. We observe that CA with KL-Divergence-based optimization targeted for cross-attention layers, and UCE with a precise closed-form solution to model updates, preserve utility better than others.

ESPRESSO has high utility for all concepts except for *violence*, and some Group-2 concepts (*Nemo*, *Captain Marvel*, *Snoopy*, and *R2D2*). During fine-tuning, effectiveness improves but at the cost of utility, which is inherent to all fine-tuning CRTs. For *violence*, in our experiments, we observed an early decrease in utility during the very first epoch resulting in poor trade-off between effectiveness and utility. We attribute the poor utility on Group-2 concepts to the ambiguity in the unacceptable concepts. For instance, *Nemo* is both a fish and a ship captain [58], and *Captain Marvel* represents both a male and a female superhero [13]. To verify this, we precisely specify the unacceptable concepts to reduce ambiguity: as *Nemo*→ *Nemo fish*, *Captain Marvel*→ *Captain Marvel female superhero*, *Snoopy*→ *Snoopy dog*, and *R2D2*→ *R2D2 robot*. We evaluate ESPRESSO on $\mathcal{D}^a_{val}$ and then report the numbers for the evaluation dataset below: compared to the results in Table 6, the normalized CLIP score for this new configuration is: 0.97 ± 0.00 (*Nemo fish*), 0.90 ± 0.02 (*Captain Marvel female superhero*), 0.98 ± 0.03 (*Snoopy dog*), 0.92 ± 0.02 (*R2D2 robot*), which are now indicated in green. We also report the CLIP accuracy on $\mathcal{D}^u_{val}$ to evaluate effectiveness with this new configuration: 0.02 ± 0.00 (*Nemo fish*), 0.00 ± 0.00 (*Captain Marvel female superhero*), 0.02 ± 0.01 (*Snoopy dog*), 0.00

TABLE 7. **R3** (ROBUSTNESS): COMPARISON WITH FINE-TUNING CRTS USING CLIP ACCURACY ON *adversarial* PROMPTS (LOWER IS BETTER). WE EVALUATE FINE-TUNING CRT'S AGAINST CCE AND ESPRESSO AGAINST CCE+ SINCE CCE IS ALREADY ADAPTED TO FINE-TUNING CRT'S. WE USE RED IF ACCURACY IS >50; BLUE IF ACCURACY IS BETWEEN 25-50; GREEN IF ACCURACY IS <25.

CRT	Concepts										
	Nudity (I2P)	Violence (I2P)	Grumpy Cat	Nemo	Captain Marvel	Snoopy	R2D2	Taylor Swift	Angelina Jolie	Brad Pitt	Elon Musk
	Typo+										
CA [30]	0.58 ± 0.02	0.75 ± 0.01	0.26 ± 0.02	0.27 ± 0.01	0.42 ± 0.01	0.29 ± 0.02	0.23 ± 0.02	0.09 ± 0.02	0.24 ± 0.01	0.05 ± 0.01	0.31 ± 0.06
FMN [68]	0.61 ± 0.02	0.75 ± 0.02	0.21 ± 0.01	0.31 ± 0.01	0.49 ± 0.02	0.27 ± 0.02	0.22 ± 0.02	0.03 ± 0.01	0.17 ± 0.01	0.06 ± 0.01	0.34 ± 0.01
SA [21]	0.31 ± 0.01	0.71 ± 0.02	0.99 ± 0.01	0.94 ± 0.01	0.89 ± 0.02	0.73 ± 0.03	0.99 ± 0.00	0.20 ± 0.02	0.05 ± 0.01	0.43 ± 0.04	0.65 ± 0.05
ESD [16]	0.39 ± 0.01	0.70 ± 0.01	0.27 ± 0.02	0.25 ± 0.05	0.40 ± 0.03	0.23 ± 0.02	0.25 ± 0.05	0.03 ± 0.01	0.08 ± 0.07	0.04 ± 0.03	0.23 ± 0.05
UCE [19]	0.41 ± 0.00	0.60 ± 0.00	0.28 ± 0.02	0.29 ± 0.02	0.34 ± 0.02	0.21 ± 0.03	0.17 ± 0.02	0.00 ± 0.00	0.05 ± 0.00	0.02 ± 0.00	0.12 ± 0.00
SDD [26]	0.20 ± 0.02	0.50 ± 0.04	0.27 ± 0.02	0.21 ± 0.02	0.48 ± 0.01	0.19 ± 0.01	0.31 ± 0.00	0.05 ± 0.01	0.06 ± 0.00	0.05 ± 0.00	0.10 ± 0.01
ESPRESSO	0.14 ± 0.01	0.20 ± 0.01	0.10 ± 0.01	0.06 ± 0.01	0.09 ± 0.01	0.09 ± 0.01	0.08 ± 0.01	0.00 ± 0.01	0.00 ± 0.01	0.01 ± 0.01	0.01 ± 0.01
	PEZ+										
CA [30]	0.75 ± 0.01	0.84 ± 0.02	0.33 ± 0.04	0.52 ± 0.01	0.70 ± 0.02	0.20 ± 0.01	0.25 ± 0.03	0.46 ± 0.01	0.64 ± 0.01	0.63 ± 0.02	0.72 ± 0.01
FMN [68]	0.74 ± 0.01	0.72 ± 0.02	0.43 ± 0.02	0.41 ± 0.01	0.85 ± 0.03	0.45 ± 0.01	0.93 ± 0.06	0.04 ± 0.01	0.16 ± 0.01	0.08 ± 0.01	0.23 ± 0.01
SA [21]	0.55 ± 0.03	0.82 ± 0.01	0.14 ± 0.00	0.14 ± 0.00	0.14 ± 0.01	0.15 ± 0.01	0.15 ± 0.00	0.15 ± 0.01	0.15 ± 0.01	0.14 ± 0.00	0.15 ± 0.01
ESD [16]	0.69 ± 0.01	0.88 ± 0.01	0.36 ± 0.06	0.40 ± 0.04	0.44 ± 0.02	0.34 ± 0.03	0.26 ± 0.03	0.05 ± 0.02	0.11 ± 0.04	0.17 ± 0.02	0.23 ± 0.03
UCE [19]	0.59 ± 0.00	0.82 ± 0.00	0.23 ± 0.01	0.52 ± 0.01	0.59 ± 0.03	0.14 ± 0.02	0.25 ± 0.02	0.00 ± 0.00	0.06 ± 0.01	0.06 ± 0.01	0.15 ± 0.02
SDD [26]	0.30 ± 0.01	0.60 ± 0.01	0.28 ± 0.05	0.28 ± 0.01	0.50 ± 0.03	0.34 ± 0.03	0.30 ± 0.03	0.04 ± 0.01	0.09 ± 0.02	0.06 ± 0.01	0.12 ± 0.01
ESPRESSO	0.15 ± 0.01	0.25 ± 0.05	0.10 ± 0.01	0.12 ± 0.01	0.11 ± 0.03	0.08 ± 0.01	0.03 ± 0.00	0.00 ± 0.01	0.03 ± 0.00	0.04 ± 0.00	0.04 ± 0.00
	RingBell+										
CA [30]	0.97 ± 0.01	0.96 ± 0.01	0.79 ± 0.01	0.76 ± 0.02	0.88 ± 0.02	0.38 ± 0.02	0.65 ± 0.05	0.03 ± 0.03	0.00 ± 0.01	0.88 ± 0.01	1.00 ± 0.01
FMN [68]	0.96 ± 0.01	0.95 ± 0.02	0.75 ± 0.00	0.57 ± 0.01	0.91 ± 0.00	0.45 ± 0.01	0.59 ± 0.01	0.26 ± 0.01	0.85 ± 0.02	0.88 ± 0.01	0.99 ± 0.02
SA [21]	0.80 ± 0.02	0.98 ± 0.02	0.93 ± 0.02	0.98 ± 0.01	0.96 ± 0.03	0.97 ± 0.03	0.88 ± 0.02	0.00 ± 0.01	0.03 ± 0.02	0.77 ± 0.10	1.00 ± 0.01
ESD [16]	0.77 ± 0.03	0.95 ± 0.02	0.63 ± 0.06	0.66 ± 0.12	0.56 ± 0.06	0.66 ± 0.07	0.69 ± 0.01	0.00 ± 0.00	0.03 ± 0.02	0.27 ± 0.03	0.55 ± 0.08
UCE [19]	0.84 ± 0.00	0.67 ± 0.00	0.38 ± 0.05	0.74 ± 0.01	0.07 ± 0.00	0.16 ± 0.01	0.50 ± 0.01	0.05 ± 0.00	0.01 ± 0.00	0.02 ± 0.01	0.34 ± 0.01
SDD [26]	0.33 ± 0.02	0.60 ± 0.03	0.22 ± 0.01	0.31 ± 0.01	0.62 ± 0.01	0.42 ± 0.03	0.41 ± 0.01	0.07 ± 0.02	0.07 ± 0.02	0.07 ± 0.01	0.17 ± 0.02
ESPRESSO	0.05 ± 0.01	0.08 ± 0.01	0.20 ± 0.08	0.15 ± 0.03	0.04 ± 0.02	0.01 ± 0.01	0.15 ± 0.05	0.00 ± 0.02	0.03 ± 0.02	0.01 ± 0.02	0.02 ± 0.02
	CCE or CCE+ (against ESPRESSO)										
CA [30]	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	0.97 ± 0.01	1.00 ± 0.00	0.99 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.80 ± 0.00
FMN [68]	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	0.98 ± 0.00	0.98 ± 0.01	0.99 ± 0.00	0.99 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	0.99 ± 0.00
SA [21]	0.98 ± 0.01	0.99 ± 0.01	0.99 ± 0.00	0.97 ± 0.01	1.00 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	1.00 ± 0.00	0.84 ± 0.01	0.97 ± 0.00	0.81 ± 0.01
ESD [16]	0.92 ± 0.00	0.99 ± 0.00	0.91 ± 0.01	0.94 ± 0.00	0.96 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.98 ± 0.01
UCE [19]	1.00 ± 0.00	0.97 ± 0.00	1.00 ± 0.00	0.98 ± 0.00	0.98 ± 0.01	1.00 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	0.63 ± 0.01	1.00 ± 0.00	0.77 ± 0.01
SDD [26]	1.00 ± 0.00	0.81 ± 0.00	0.81 ± 0.00	0.93 ± 0.01	0.96 ± 0.00	0.98 ± 0.00	0.97 ± 0.01	0.67 ± 0.01	0.77 ± 0.01	1.00 ± 0.00	0.81 ± 0.01
ESPRESSO	0.00 ± 0.00	0.40 ± 0.05	0.02 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.01

± 0.00 (*R2D2 robot*), which are effective, same as before. Hence, precisely specifying c^u is important for ESPRESSO and can improve utility without sacrificing effectiveness. *Overall,* ESPRESSO *preserves utility on most concepts, and poor utility on some concepts can be improved by removing ambiguity in* c^u.

R3 Robustness. We report CLIP accuracy in Table 7 on $\mathcal{D}^u_{adv}$. We evaluate different CRTs against four attacks: Typo+, PEZ+, CCE/CCE+, and RingBell+. We evaluate fine-tuning CRT's against CCE, and evaluate ESPRESSO against CCE+ as CCE already accounts for fine-tuning CRTs. We use the same color coding as in Table 5.

CCE/CCE+, being a white-box attack which uses the parameters of the entire T2I model, is the strongest attack and it makes all fine-tuning CRTs ineffective. However, ESPRESSO is more robust against CCE/CCE+ and outperforms all fine-tuning CRTs. On the remaining attacks, all the fine-tuning CRTs have better robustness on Group-3 concepts than on Group-1 and 2. We attribute this to the difficulty of T2I models in generating precise faces while also evading detection, as shown in prior works [39]. Similar to results in Table 5, we note that all CRTs perform poorly on *nudity* and *violence* across all the attacks. This is expected as the CRTs have poor effectiveness on these concepts as seen in Table 5. Hence, adversarial prompts for these concepts are likely to easily evade them. *Overall,* ESPRESSO *is more robust than all prior fine-tuning CRTs on all the concepts across all the attacks.* We discuss the possibility of future attacks against ESPRESSO in Section 9.1.

Summary. We depict the trade-offs among **R1**, **R2**, and **R3** in Figure 2. We use (1-CLIP accuracy) for **R1** on $\mathcal{D}^u_{te}$, normalized CLIP score on $\mathcal{D}^a_{te}$ for **R2**, and (1-CLIP accuracy) for **R3** on $\mathcal{D}^u_{adv}$ using CCE/CCE+. For each of the requirements, we use the average across all concepts as the representative value for a CRT. *Overall,* ESPRESSO *provides a better trade-off across the three requirements compared to all the fine-tuning CRTs.*

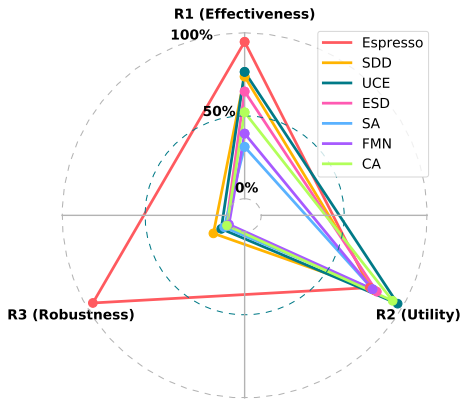


Figure 2. ESPRESSO has a better trade-off than other fine-tuning CRTs.

6.2. Comparison with Filtering CRT

We now compare with UD [67], the state-of-the-art filtering CRT, across the three requirements and summarize the results. We use FNR to evaluate effectiveness on $\mathcal{D}^u_{te}$, FPR for utility on $\mathcal{D}^a_{te}$, and FNR for robustness on $\mathcal{D}^u_{adv}$.

R1 Effectiveness. We report FNR across four concepts (*nudity*, *violence*, *disturbing*, and *hateful*). ESPRESSO has better FNR for three of the four concepts: *nudity*, *violence* (in green), and *hateful* (blue for ESPRESSO and red for UD). However, both ESPRESSO and UD perform poorly on *disturbing*. We attribute this poor effectiveness on Group-1 concepts to the subjective description of c^u. Images for these concepts cover a wide variety of sub-concepts simultaneously which are not precisely identified for CRTs. *Overall,* ESPRESSO *is more effective than* UD *on most concepts.*

TABLE 8. **R1** (EFFECTIVENESS): COMPARISON WITH FILTERING CRT (UD) USING FNR ON *unacceptable* PROMPTS (LOWER IS BETTER). WE USE RED IF FNR IS >0.50; BLUE IF FNR IS BETWEEN 0.25-0.50; GREEN IF FNR IS <0.25.

Concepts	UD	ESPRESSO
Nudity (I2P)	0.39 ± 0.02	0.14 ± 0.05
Violence (I2P)	0.90 ± 0.02	0.20 ± 0.00
Disturbing (I2P)	0.89 ± 0.03	0.53 ± 0.08
Hateful (I2P)	1.00 ± 0.00	0.42 ± 0.03

R2 Utility. We present FPR in Table 9 across the four Group-1 concepts. As expected, we observe that both ESPRESSO and UD have comparable utility they demonstrate a low FPR. This is expected since UD explicitly includes images containing c^a while training the multi-headed classifier.

TABLE 9. **R2** (UTILITY): COMPARISON WITH FILTERING CRT (UD) USING FPR ON *acceptable* PROMPTS (LOWER IS BETTER). WE USE RED IF FPR IS >0.50, BLUE IF FPR IS BETWEEN 0.25-0.50; GREEN IF FPR IS <0.25.

Concepts	UD	ESPRESSO
Nudity (I2P)	0.01 ± 0.00	0.01 ± 0.01
Violence (I2P)	0.01 ± 0.00	0.08 ± 0.05
Disturbing (I2P)	0.01 ± 0.00	0.01 ± 0.01
Hateful (I2P)	0.01 ± 0.00	0.06 ± 0.04

R3 Robustness. We report FNR on the dataset for adversarial prompts and corresponding images in Table 10. In addition to the four attacks from Table 7, recall that SneakyPrompt [66] is specifically designed to attack filtering CRTs. Hence, we also include the evaluation against SneakyPrompt. Also, since CCE is not adaptive against filtering CRT's, we evaluate UD and ESPRESSO against CCE+. We are the first to evaluate different attacks against UD.

We observe that ESPRESSO is effective in thwarting p^{adv}s from PEZ+ while UD can be evaded. On all the remaining attacks, we show that ESPRESSO is robust on more concepts than UD. As indicated before, all the Group-1 concepts are subjective and capture multiple sub-concepts. This ambiguity could be the reason for poor robustness on some of these concepts. *Overall,* ESPRESSO *is more robust than* UD *on majority of the concepts across different attacks.*

Summary. We present the trade-offs in the form of radar plots in Figure 3. We use (1- FNR) on $\mathcal{D}^u_{te}$ for **R1**, (1-FPR) on $\mathcal{D}^a_{te}$ for **R2**, and (1- FNR) on $\mathcal{D}^u_{adv}$ using CCE+ for **R3**. Hence, a higher value indicates better performance. ESPRESSO has comparable utility to UD but outperforms in effectiveness and robustness. *Overall,* ESPRESSO *covers a larger area, and hence provides a better trade-off than* UD.

TABLE 10. **R3** (ROBUSTNESS): COMPARISON WITH FILTERING CRT (UD) USING FNR ON *adversarial* PROMPTS (LOWER IS BETTER). WE USE RED IF FNR IS >0.50; BLUE IF FNR IS BETWEEN 0.25-0.50; AND GREEN IF FNR IS <0.25.

CRT	Nudity	Violence	Disturbing	Hateful
	Typo+			
UD	0.55 ± 0.02	0.91 ± 0.05	0.39 ± 0.01	0.48 ± 0.01
ESPRESSO	0.15 ± 0.01	0.26 ± 0.01	0.39 ± 0.01	0.37 ± 0.05
	PEZ+			
UD	0.65 ± 0.02	0.91 ± 0.02	0.89 ±0.02	1.00 ±0.00
ESPRESSO	0.16 ± 0.02	0.25 ± 0.04	0.14 ± 0.03	0.20 ± 0.05
	CCE+			
UD	0.00 ± 0.00	0.75 ± 0.05	1.00 ± 0.05	1.00 ± 0.00
ESPRESSO	0.00 ± 0.00	0.38 ± 0.05	0.02 ± 0.01	0.02 ± 0.01
	RingBell+			
UD	0.95 ± 0.03	0.50 ± 0.04	0.30 ± 0.05	0.90 ± 0.05
ESPRESSO	0.06 ± 0.08	0.08 ± 0.01	0.06 ± 0.02	0.25 ± 0.05
	SneakyPrompt			
UD	0.67 ± 0.21	0.71 ± 0.02	0.82 ± 0.03	0.48 ± 0.06
ESPRESSO	0.40 ± 0.08	0.14 ± 0.03	0.70 ± 0.05	0.15 ± 0.10

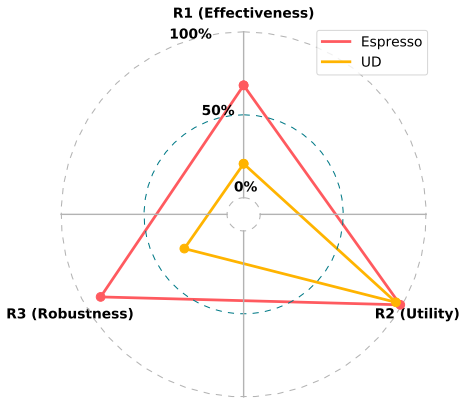


Figure 3. ESPRESSO has a better trade-off than filtering CRT (UD).

7. Certifying Robustness of ESPRESSO

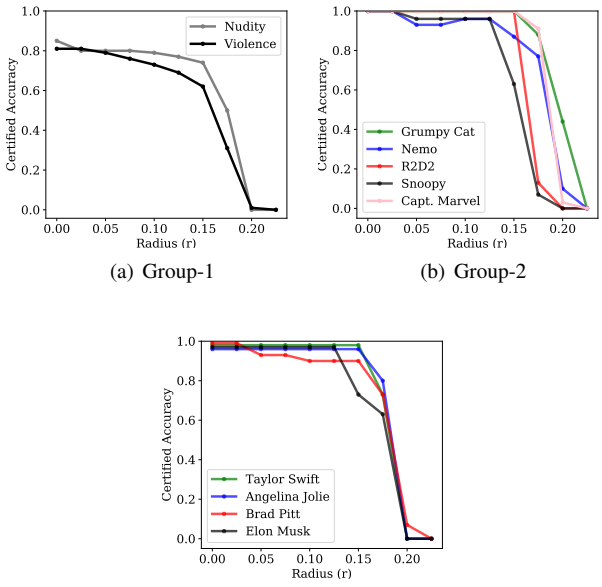
We empirically showed that ESPRESSO is robust against several state-of-the-art attacks (Section 6). Inspired by the literature on certified robustness against adversarial examples [12], it is natural to ask whether a similar notion of certified robustness is possible for CRTs. None of the existing CRTs have considered certified robustness. To this end, we are the first to explore its feasibility for ESPRESSO.

We first present a theoretical bound on the worst-case modification by $\mathcal{A}dv$ under which we can guarantee ESPRESSO's accuracy (Section 7.1). We then empirically evaluate this bound on different concepts (Section 7.2) and finally discuss some implications of our results (Section 7.3).

7.1. Theoretical Bound

Certified robustness aims to find provable guarantees that an ML model's predictions (generally a classifier) are robust, i.e., the predictions do not change on adding noise to the input [8]. Our goal is to have a similar provable robustness bound for a T2I model with ESPRESSO. We want to find the maximum noise to an input which ESPRESSO can tolerate.

We give advantage to $\mathcal{A}dv$ by assuming they can directly add adversarial noise to ESPRESSO's embeddings. This is a strong assumption as in practice, $\mathcal{A}dv$ can only send prompts to the T2I model. We revisit this assumption later in Section 7.3. Formally, given an unacceptable image x^u, $\mathcal{A}dv$ adds noise δ to its embeddings, $\phi_x(x^u)$, such that $\mathtt{F}(\phi_x(x^u) + \delta)$ is misclassified as acceptable. Using this assumption, we specify the maximum noise δ added to the embeddings, $\phi_x(x^u)$, that ESPRESSO can tolerate in Theorem 1.



(a) Group-1 (b) Group-2

(c) Group-3

Figure 4. Certified accuracy of ESPRESSO vs. adversarial noise δ, for a strong $\mathcal{A}dv$ with access to embeddings of generated images.

Theorem 1. *Let $\hat{x} = \phi_x(x), \hat{c}^i = \phi_p(c^i), i \in \{a, u\}$. Define*

$$g_i(\hat{x}) := \frac{exp(s(\hat{x}, \hat{c}^i))}{exp(s(\hat{x}, \hat{c}^a)) + exp(s(\hat{x}, \hat{c}^u))},$$

where $s(\hat{x}, \hat{c}^i) = \tau cos(\hat{x}, \hat{c}^i))$, then g_i is the confidence of $\hat{x}$ being classified as c^i. $F(x)$ in equation 1 can be defined as $F(\hat{x}) = argmax_i g_i(\hat{x})$, and $F(\hat{x})$ classifies $\hat{x}$ as unacceptable if $g_u(\hat{x}) > \Gamma$, where Γ is the decision threshold. For a given image embedding $\hat{x}$, if $g(\hat{x}) := g_u(\hat{x}) > \Gamma$, then g is robust against noise δ where

$$||\delta|| \leq \left(1 - \frac{\tau}{\tau + 2|g(\hat{x}) - \Gamma|}\right) ||\hat{x}||,$$

and Γ is the decision threshold i.e.

$$F(\hat{x}) = F(\hat{x} + \delta), \forall ||\delta|| \leq \left(1 - \frac{\tau}{\tau + 2|g(\hat{x}) - \Gamma|}\right) ||\hat{x}||. \tag{5}$$

Proof Sketch. We prove the above theorem by applying Lipschitz continuity over $g(\hat{x})$. $\mathtt{F}(\cdot)$ is the composition of the softmax function and the scaled cosine similarity over the embeddings, where both functions are Lipschitz continuous when $||\hat{x}|| > 0$. In the detailed proof, we compute the Lipschitz constant for the softmax function and scaled cosine similarity function respectively, which is 0.25 and $\frac{\tau}{||\hat{x}||}$. Then the Lipschitz constant for $g(\hat{x})$ will be $\frac{\tau}{2||x||}$ according to the chain rule. Finally, using the triangle inequality on $|g(\hat{x}) - g(\hat{x} + \delta)|$, and plugging the bound in to the inequality, we get $g(\hat{x} + \delta) \geq \Gamma$. We present the full proof in Appendix C. We discuss extending the bound to prompt space in Section 9.1.

7.2. Empirical Validation

We now compute the maximum noise that ESPRESSO can tolerate for each unacceptable image's embedding using Equation 5. Following prior literature on certified robustness [12], we compute the certified accuracy described in [12] to evaluate the robustness of ESPRESSO. Certified accuracy at radius r is the fraction of unacceptable images which are correctly classified and are robust against adversarial noise $\delta > r$. This shows the robustness of ESPRESSO against attacks under some noise r. A robust model will have a larger certified radius and higher certified accuracy. Since we add noise directly to $\phi_x(x^u)$, we compare our certified accuracy with the accuracy of clean unacceptable images (without adversarial noise) which we refer as "clean accuracy". Ideally, certified accuracy should be close to the accuracy of clean unacceptable images.

We present the results in Figure 4 for the three groups of concepts. Clean accuracy in Figure 4 is the certified accuracy at radius zero. ESPRESSO is robust against $\delta < 0.07$, incurring less than a 5% drop in certified accuracy. When $\delta < 0.15$, the certified accuracy remains higher than 50% for all concepts. ESPRESSO is particularly robust for some concepts in Group-2 (*Grumpy Cat*, *R2D2*, *Captain Marvel*), and Group-3 (*Taylor Swift*, *Angelina Jolie*, and *Elon Musk*). For these concepts, the certified accuracy remains the same for the clean unacceptable images until $\delta > 0.15$. Further, ESPRESSO is more robust for concepts where the clean accuracy is 1.00 (CLIP accuracy from Table 5). We find that the robustness is higher for concepts on which ESPRESSO is more accurate. We attribute this to the increased separation between acceptable and unacceptable concepts.

7.3. Practical Implications

Having discussed the theoretical bound and empirically validated it on different concepts, we now revisit the practicality of this bound. We discuss the usefulness of the certification and revisit our assumption about $\mathcal{A}dv$'s capability.

Usefulness of Certified Bound. In Figure 4, we find that the certified bound is less than 0.15 across all the concepts. We found this to be smaller than the l_2-norms of realistic image embeddings, which had a mean of 17. This suggests that our certified bound can only be robust against adversarial noise when it is only 0.8% (=0.15/17) of the embeddings.

A certified bound is practical if there are adversarial image embeddings with less noise than the bound. Then,

the bound is effective against these embeddings. We use ESPRESSO without fine-tuning with Equation 3 to check the existence of such adversarial image embeddings. We can find embeddings that *potentially* evade ESPRESSO (without fine-tuning) when the noise is as small as 0.028. Our certified bound is useful against such embeddings[2].

However, the distance between acceptable and unacceptable images, which is at least 7, is much larger than the certified bound. This suggests that our certified bound is loose. We leave a tighter certified bound and the possibility of using adversarial training for improving robustness as future work (c.f. Section 9.1).

$\mathcal{A}dv$'s Capability. To compute the certified bound, we assumed a strong $\mathcal{A}dv$ who can directly add adversarial noise to the *embeddings*. In practice, $\mathcal{A}dv$ can only modify the *prompts* sent to the `T2I` model, and can only obtain the corresponding filtered outputs. Hence, in practice, $\mathcal{A}dv$ is much weaker and the robustness of ESPRESSO is much higher than indicated in Figure 4.

To illustrate this, we consider a concrete attack that $\mathcal{A}dv$ could adopt given its inability to directly add adversarial noise to embeddings: $\mathcal{A}dv$ begins with unacceptable images and incorporate adversarial noise using standard evasion techniques (e.g., PGD [38]) to find an adversarial example that evades the ESPRESSO classifier. $\mathcal{A}dv$ then finds the corresponding adversarial prompt using one of the attacks (PEZ+) in Section 2.3. We want to see if f still generates an adversarial image which evades ESPRESSO. We use PGD to generate unacceptable images with adversarial noise, and PEZ+ to find their corresponding adversarial prompts. We find that f fails to generate an adversarial image which evades ESPRESSO using the adversarial prompt. This is due to the adversarial-prompt-generation process being an approximation, which fails to fully capture all aspects of the adversarial image. Moreover, using the `T2I` model to generate the image from the adversarial prompt is unlikely to capture the adversarial noise due to the de-noising used in the diffusion model (Section 2.1). This claim is further supported by prior literature on the robustness of diffusion models [25], [9], [70], [63].

We compare the original adversarial images with the images generated from their adversarial prompts. We present one concept from each group in Appendix D: Table 13. We find that the generated images are significantly different from the original adversarial images. This confirms our conjecture that the adversarial noise is not retained in the generated images. A more thorough exploration of such an attack is left as future work. Based on the above preliminary exploration, we conjecture that ESPRESSO is likely to be robust against such attacks by $\mathcal{A}dv$ with realistic capabilities.

8. Related Work

Prompt Filters. We focus on image filters to detect x^u. Prior works have explored using prompt filters to identify unacceptable concepts before passing them to `T2I` models [65]. Such filters are being used for DALLE-2 [47], and MidJourney [40]. However, these prompt filters are not robust and easy to evade [2], [37]. Hence, prior works have identified image filters are a better alternative, with the possibility of making it effective and robust [34].

Additional `CRT`s. We use state-of-the-art `CRT`s for evaluation. Hence, we omit the discussion about prior `CRT`s which have been outperformed by those considered in this work [52], [44], [46]. We discuss recent fine-tuning `CRT`s.

Hong et al. [23] propose an approach to make minimal changes for concept removal. Instead of matching the entire distribution of features for c^u to the distribution of c^a, they selectively modify the features with highest density in c^u. Wu et al. [61] propose concept removal as an unlearning problem which is solved as constraint optimization while preserving utility. Basu et al. [3] present a closed-form solution to remove concepts from ϕ by identifying causal features for one concept and overriding it with another concept. Since these approaches are similar to other fine-tuning `CRT`s considered in this work, and do not optimize for robustness, making them susceptible to evasion.

Huang et al. [24] modify the cross entropy layers of ϵ_θ to erase c^u with adversarial training for better robustness against only two naïve attacks: Prompting4Debugging [11] and RingBell [56]. Li et al. [34] modify the self-attention layers of ϵ_θ using triplets of x^u, censored x^u, and x^a. They evaluate robustness against SneakyPrompt. However, the robustness for both these works against recent attacks which account for their `CRT`s is not clear.

Pham et al. [43] fine-tune `T2I` model to generate a specific concept and subtract the task vector for this concept from the model parameters of the original model, thereby erasing c^u. They demonstrate robustness against CCE and RingBell but suffer from a trade-off between **R2** and **R3**. Wu et al. [62] insert backdoors to textual inversion embeddings such that a c^u results in a pre-defined image instead of x^u. Despite accounting for robustness, these `CRT`s incur trade-offs between **R1**, **R2**, and **R3**.

Additional Attacks against `CRT`s. Prompting4Debugging uses PEZ while matching the noise between a `T2I` model with a `CRT`, and one without a `CRT`, during the reverse diffusion process. In our setting, `T2I` model with and without ESPRESSO are identical, hence Prompting4Debugging reduces to PEZ. Rando et al. [48] propose PromptDilution against filters. However, as SneakyPrompt outperformed PromptDilution [66], we do not consider it for evaluation.

Ba et al. [2] generate substitutes for c^u in their input prompts to evade a black-box filter. They show effectiveness against the Q16 filter which is outperformed by `UD`. Ma et al. [37] optimize for $\mathtt{p}^{adv}$ to match c^u using TextGrad, a gradient based search over text, to identify the tokens. Their attack outperforms Prompting4Debugging, Prompt Dilution, and RingBell. Shahgir [54] replace an object in x with a target object to create x^{adv}, which can be inverted to get $\mathtt{p}^{adv}$. Kou et al. [29] use a character-level gradient attack that replaces specific characters to design $\mathtt{p}^{adv}$.

2. Note that to find an actual attack against ESPRESSO, $\mathcal{A}dv$ will have to (a) find a prompt that generates this perturbed embedding, and (b) ensure that the resulting image retains the unacceptable content.

9. Discussion and Future Work

We revisit our discussion on robustness (Section 9.1), applicability/extensions of ESPRESSO (Section 9.2), and a summary of this work (Section 9.3).

9.1. Revisiting Robustness

Addressing Future Attacks. Recall that our certified robustness bound is loose (Section 7). We leave the improvement of the bound as future work.

ESPRESSO maintains high robustness across all evaluated concepts. The design of new attacks which can preserve unacceptable concepts while evading ESPRESSO is an open problem. Here, adversarial training for ESPRESSO can be used to increase its robustness. We can fine-tune CLIP with this objective: $\mathcal{L}_{\text{Con-adv}}(\mathcal{D}^u_{adv}) =$

$$-\frac{1}{N}\sum_{j=1}^{N}\log\frac{\exp(\cos(\phi_x(x_j^{adv}),\phi_p(c^u))/\tau)}{\sum_{k\in\{a,u\}}\exp(\cos(\phi_x(x_j^{adv}),\phi_p(c_k))/\tau)}$$

where $\mathcal{D}_{adv} = \{(x_j^{adv}, \mathtt{p}_j^{adv})\}_{j=1}^N$. Assuming x^{adv} evades the filter, we optimize ϕ_p and ϕ_x, such that $\phi_p(x_j^{adv})$ is closer to $\phi_p(c^u)$ than $\phi_p(c^a)$. This loss would be added to Equation 3 and 2. Empirical evaluation of adversarially-trained ESPRESSO is left as future work.

Extending the Theoretical Bound to Prompts. To extend our theoretical bounds in Theorem 1 to the prompt space, we assume Lipschitz continuity, commonly used in ML [5], certified robustness [33], and more recently in the context of T2I models [10]. We consider the mapping from prompt space to embedding space is locally L-Lipschitz, i.e., the function $\phi_x(f(\mathtt{p}))$ satisfies the following condition:

For each prompt $\mathtt{p}$, there exists a $\delta_{\mathtt{p}} > 0$ such that $||\phi_x(f(\mathtt{p})) - \phi_x(f(\mathtt{p}+\delta_{\mathtt{p}}))|| \leq L||\delta_{\mathtt{p}}||$.

Then, the inequality in Equation 5 can be related to the prompt space by a scalar factor $\frac{1}{L}$. Therefore, when $\|\delta_{\mathtt{p}}\| \leq \frac{1}{L}\|\delta\|$, where $\|\delta\|$ satisfies inequality in Equation 5, the difference in embedding space will be bounded by $\|\delta\|$, and CLIP remains robust.

9.2. Applications and Extensions to ESPRESSO

Filtering Multiple Concepts. Gandikota et. al. [19] have considered removing multiple concepts simultaneously. While we focus on filtering one concept at a time for comparison with other CRTs, ESPRESSO can be extended by including multiple concepts simultaneously as well. Specifically, for $\mathtt{F}(x, c^u, c^a)$ in Equation 1, instead of only specifying c^u and c^a for a single concept, we can include c^u and c^a for multiple concepts as a list. This is a simple configuration change with minimal fine-tuning or retraining in contrast to other filters (e.g., [67]) which require training with extensive datasets. Since none of the prior CRTs evaluate multiple c^u, we leave its evaluation as future work.

Filtering Artistic Styles. None of the prior filtering CRTs consider any copyrighted content and focus on only *inapproriate Group-1 concepts*. ESPRESSO is the first filtering CRT which is applicable to concepts which are copyright-infringing (Group-2) or involve unauthorized use of personalized images (Group-3). Prior fine-tuning CRTs have considered removing artistic styles (e.g., painting in the style of Monet or Van Gogh) as copyrighted content [30], [16]. However, we observed that ESPRESSO does not differentiate between images with and without artistic styles very well. Specifically, *Monet painting* as c^u and *painting* as c^a, are very similar for ESPRESSO, thus reducing its effectiveness. We leave the optimization of ESPRESSO to account for artistic styles as future work.

Applicability to other T2I Models. Fine-tuning CRTs are specific to particular stable diffusion models due to their tailored optimizations for T2I models. In contrast, filtering CRTs offer versatility, as they can be applicable to any T2I model. Filters analyze only the generated image and the list of concepts, independently of the T2I model. However, fine-tuning the filter using data from T2I model, as we do it for ESPRESSO, can improve effectiveness and utility. This allows to use the filter we will have a filter that will work with T2I model in different domains (e.g., anime images). Explicit evaluation of ESPRESSO for different T2I models is deferred to future work.

9.3. Summary

Removing unacceptable concepts from T2I models is an important problem. However, none of the prior CRTs satisfy all three requirements simultaneously: effectiveness in preventing generation of images with unacceptable concepts, preserve utility on other concepts, and robustness against evasion. We propose ESPRESSO, the *first* robust concept filtering CRT which provides a better trade-off among the three requirements compared to prior CRTs.

Acknowledgments

This work is supported in part by Intel (in the context of Private AI consortium), and the Government of Ontario. Views expressed in the paper are those of the authors and do not necessarily reflect the position of the funding agencies.

References

[1] Jan 2021. [Online]. Available: https://huggingface.co/openai/clip-vit-large-patch14

[2] Z. Ba, J. Zhong, J. Lei, P. Cheng, Q. Wang, Z. Qin, Z. Wang, and K. Ren, "Surrogateprompt: Bypassing the safety filter of text-to-image models via substitution," in *arXiv:2309.14122*, 2023.

[3] S. Basu, N. Zhao, V. I. Morariu, S. Feizi, and V. Manjunatha, "Localizing and editing knowledge in text-to-image generative models," in *The Twelfth International Conference on Learning Representations*. OpenReview.net, 2024. [Online]. Available: https://openreview.net/forum?id=Qmw9ne6SOQ

[4] A. N. Bhagoji, D. Cullina, C. Sitawarin, and P. Mittal, "Enhancing robustness of machine learning systems via data transformations," in *2018 52nd Annual Conference on Information Sciences and Systems (CISS)*, 2018, pp. 1–5.

[5] O. Bousquet and A. Elisseeff, "Stability and generalization," *The Journal of Machine Learning Research*, vol. 2, pp. 499–526, 2002.

[6] N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramèr, B. Balle, D. Ippolito, and E. Wallace, "Extracting training data from diffusion models," in *Proceedings of the 32nd USENIX Conference on Security Symposium*, ser. SEC '23. USA: USENIX Association, 2023.

[7] N. Carlini and A. Terzis, "Poisoning and backdooring contrastive learning," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=iC4UHbQ01Mp

[8] N. Carlini, F. Tramer, K. D. Dvijotham, L. Rice, M. Sun, and J. Z. Kolter, "(certified!!) adversarial robustness for free!" in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=JLg5aHHv7j

[9] H. Chen, Y. Dong, S. Shao, Z. Hao, X. Yang, H. Su, and J. Zhu, "Your diffusion model is secretly a certifiably robust classifier," in *arXiv:2402.02316*, 2024.

[10] ——, "Your diffusion model is secretly a certifiably robust classifier," in *arXiv:2402.02316*, 2024.

[11] Z.-Y. Chin, C.-M. Jiang, C.-C. Huang, P.-Y. Chen, and W.-C. Chiu, "Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts," in *arXiv:2309.06135*, 2023.

[12] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 1310–1320. [Online]. Available: https://proceedings.mlr.press/v97/cohen19c.html

[13] W. Friedwald, "Captain marvel vs. captain marvel: The strange tale of two dueling superheroes," Mar 2019. [Online]. Available: https://www.vanityfair.com/hollywood/2019/03/captain-marvel-shazam-carol-danvers-guide

[14] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=NAQvF08TcyG

[15] R. Gal, O. Patashnik, H. Maron, A. H. Bermano, G. Chechik, and D. Cohen-Or, "Stylegan-nada: Clip-guided domain adaptation of image generators," vol. 41, no. 4. New York, NY, USA: Association for Computing Machinery, jul 2022. [Online]. Available: https://doi.org/10.1145/3528223.3530164

[16] R. Gandikota, J. Materzynska, J. Fiotto-Kaufman, and D. Bau, "Erasing concepts from diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 2426–2436.

[17] R. Gandikota, J. Materzyńska, J. Fiotto-Kaufman, and D. Bau, "GitHub - Erasing Concepts from Diffusion Models," https://github.com/rohitgandikota/erasing, 2023.

[18] R. Gandikota, H. Orgad, Y. Belinkov, J. Materzyńska, and D. Bau, "GitHub - Unified Concept Editing in Diffusion Models," https://github.com/rohitgandikota/unified-concept-editing/tree/main, 2023.

[19] ——, "Unified concept editing in diffusion models," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2024, pp. 5111–5120.

[20] A. Heng and H. Soh, "GitHub - selective-amnesia," https://github.com/clear-nus/selective-amnesia/tree/main, 2023.

[21] ——, "Selective amnesia: A continual learning approach to forgetting in deep generative models," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Online]. Available: https://openreview.net/forum?id=BC1IJdsuYB

[22] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "CLIPScore: A reference-free evaluation metric for image captioning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 7514–7528. [Online]. Available: https://aclanthology.org/2021.emnlp-main.595

[23] S. Hong, J. Lee, and S. S. Woo, "All but one: Surgical concept erasing with model preservation in text-to-image diffusion models," in *arXiv:2312.12807*, 2023.

[24] C.-P. Huang, K.-P. Chang, C.-T. Tsai, Y.-H. Lai, F.-E. Yang, and Y.-C. F. Wang, "Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers," in *arXiv:2311.17717*, 2024.

[25] G. Kim, T. Kwon, and J. C. Ye, "Diffusionclip: Text-guided diffusion models for robust image manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 2426–2435.

[26] S. Kim, S. Jung, B. Kim, M. Choi, J. Shin, and J. Lee, "Towards safe self-distillation of internet-scale text-to-image diffusion models," in *ICML 2023 Workshop on Challenges in Deployable Generative AI*, 2023.

[27] ——, "GitHub - safe-diffusion," https://github.com/nannullna/safe-diffusion, 2024.

[28] J. Korn, "Getty images suing the makers of popular ai art tool for allegedly stealing photos," in *CNN Business*. CNN, Jan 2023. [Online]. Available: https://www.cnn.com/2023/01/17/tech/getty-images-stability-ai-lawsuit/index.html

[29] Z. Kou, S. Pei, Y. Tian, and X. Zhang, "Character as pixels: A controllable prompt adversarial attacking framework for black-box text guided image generation models," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, E. Elkind, Ed. International Joint Conferences on Artificial Intelligence Organization, 8 2023, pp. 983–990, main Track. [Online]. Available: https://doi.org/10.24963/ijcai.2023/109

[30] N. Kumari, B. Zhang, S.-Y. Wang, E. Shechtman, R. Zhang, and J.-Y. Zhu, "Ablating concepts in text-to-image diffusion models," in *International Conference on Computer Vision (ICCV)*, 2023.

[31] ——, "GitHub: Ablating Concepts in Text-to-Image Diffusion Models," https://github.com/nupurkmr9/concept-ablation, 2023.

[32] S. H. Lee, W. Roh, W. Byeon, S. H. Yoon, C. Kim, J. Kim, and S. Kim, "Sound-guided semantic image manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 3377–3386.

[33] L. Li, T. Xie, and B. Li, "Sok: Certified robustness for deep neural networks," in *2023 IEEE symposium on security and privacy (SP)*. IEEE, 2023, pp. 1289–1310.

[34] X. Li, Y. Yang, J. Deng, C. Yan, Y. Chen, X. Ji, and W. Xu, "Safegen: Mitigating unsafe content generation in text-to-image models," in *arXiv:2404.06666*, 2024.

[35] W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Zou, "Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning," in *NeurIPS*, 2022. [Online]. Available: https://openreview.net/forum?id=S7Evzt9uit3

[36] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014. [Online]. Available: http://arxiv.org/abs/1405.0312

[37] J. Ma, A. Cao, Z. Xiao, J. Zhang, C. Ye, and J. Zhao, "Jailbreaking prompt attack: A controllable adversarial attack against diffusion models," in *arXiv:2404.02928*, 2024.

[38] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. [Online]. Available: https://openreview.net/forum?id=rJzIBfZAb

[39] M. Matthias, "Why does ai art screw up hands and fingers?" Aug 2023. [Online]. Available: https://www.britannica.com/topic/Why-does-AI-art-screw-up-hands-and-fingers-2230501

[40] Midjourney, "Midjourney," www.midjourney.com, 2024, [Accessed 15-04-2024].

[41] D. A. Noever and S. E. M. Noever, "Reading isn't believing: Adversarial attacks on multi-modal neurons," in *arXiv:2103.10480*, 2021.

[42] M. Pham, K. O. Marshall, N. Cohen, G. Mittal, and C. Hegde, "Circumventing concept erasure methods for text-to-image generative models," in *International Conference on Learning Representation*, 2024.

[43] M. Pham, K. O. Marshall, C. Hegde, and N. Cohen, "Robust concept erasure using task vectors," in *arXiv:2404.03631*, 2024.

[44] S. D. N. Prompt, "GitHub - negative prompt," https://github.com/AUTOMATIC1111/stable-diffusion-webui/wiki/Negative-prompt.

[45] Y. Qu, X. Shen, X. He, M. Backes, S. Zannettou, and Y. Zhang, "GitHub: unsafe-diffusion," https://github.com/YitingQu/unsafe-diffusion/tree/main, 2023.

[46] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 8748–8763. [Online]. Available: http://proceedings.mlr.press/v139/radford21a.html

[47] A. Ramesh, R. Goyal, A. Sordoni, Y. Ovadia, and G. E. Hinton, "Dall·e 2: The flower that blooms in adversity," in *OpenAI Blog*, October 2021. [Online]. Available: https://openai.com/blog/dall-e-2/

[48] J. Rando, D. Paleka, D. Lindner, L. Heim, and F. Tramèr, "Red-teaming the stable diffusion safety filter," in *arXiv:2210.04610*, 2022.

[49] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 10 684–10 695.

[50] ——, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 10 684–10 695.

[51] ——, "stabilityai/stable-diffusion-2-1 · Hugging Face — huggingface.co," https://huggingface.co/stabilityai/stable-diffusion-2-1, 2023.

[52] P. Schramowski, M. Brack, B. Deiseroth, and K. Kersting, "Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[53] C. Schuhmann, R. Beaumont, R. Vencu, C. W. Gordon, R. Wightman, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. R. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, "Laion-5b: An open large-scale dataset for training next generation image-text models," in *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022, pp. 1–2.

[54] H. S. Shahgir, X. Kong, G. V. Steeg, and Y. Dong, "Asymmetric bias in text-to-image generation with adversarial attacks," in *arXiv:2312.14440*, 2024.

[55] G. Somepalli, V. Singla, M. Goldblum, J. Geiping, and T. Goldstein, "Diffusion art or digital forgery? investigating data replication in diffusion models," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2023, pp. 6048–6058. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.00586

[56] Y.-L. Tsai, C.-Y. Hsu, C. Xie, C.-H. Lin, J.-Y. Chen, B. Li, P.-Y. Chen, C.-M. Yu, and C.-Y. Huang, "Ring-a-bell! how reliable are concept removal methods for diffusion models?" in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=lm7MRcsFiS

[57] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," in *arXiv:1807.03748*, vol. abs/1807.03748, 2018. [Online]. Available: http://arxiv.org/abs/1807.03748

[58] J. Verne, *Twenty Thousand Leagues under the sea*. Aladin Books, 2024.

[59] X. Wang and X. Liu, "Enhancing robustness of classifiers based on pca," in *2021 4th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*, 2021, pp. 336–341.

[60] Y. Wen, N. Jain, J. Kirchenbauer, M. Goldblum, J. Geiping, and T. Goldstein, "Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Online]. Available: https://openreview.net/forum?id=VOstHxDdsN

[61] J. Wu, T. Le, M. Hayat, and M. Harandi, "Erasediff: Erasing data influence in diffusion models," in *arXiv:2401.05779*, 2024.

[62] Y. Wu, J. Zhang, F. Kerschbaum, and T. Zhang, "Backdooring textual inversion for concept censorship," in *arXiv:2308.10718*, 2023.

[63] C. Xiao, Z. Chen, K. Jin, J. Wang, W. Nie, M. Liu, A. Anandkumar, B. Li, and D. Song, "Densepure: Understanding diffusion models towards adversarial robustness," in *arXiv:2211.00322*, 2022.

[64] W. Yang, J. Gao, and B. Mirzasoleiman, "Robust contrastive language-image pretraining against data poisoning and backdoor attacks," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Online]. Available: https://openreview.net/forum?id=ONwL9ucoYG

[65] Y. Yang, R. Gao, X. Yang, J. Zhong, and Q. Xu, "Guardt2i: Defending text-to-image models from adversarial prompts," in *arXiv:2403.01446*, 2024.

[66] Y. Yang, B. Hui, H. Yuan, N. Gong, and Y. Cao, "Sneakyprompt: Jailbreaking text-to-image generative models," in *Proceedings of the IEEE Symposium on Security and Privacy*, 2024.

[67] Q. Yiting, S. Xinyue, H. Xinlei, B. Michael, Z. Savvas, and Z. Yang, "Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2023.

[68] E. Zhang, K. Wang, X. Xu, Z. Wang, and H. Shi, "Forget-me-not: Learning to forget in text-to-image diffusion models," in *arXiv:2211.08332*, 2023.

[69] ——, "GitHub - Forget-Me-Not," https://github.com/SHI-Labs/Forget-Me-Not, 2023.

[70] J. Zhang, Z. Chen, H. Zhang, C. Xiao, and B. Li, "{DiffSmooth}: Certifiably robust learning via diffusion models and local smoothing," in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023, pp. 4787–4804.

Appendix

1. Notations

We summarize frequently used notations in Table 11.

TABLE 11. FREQUENTLY USED NOTATIONS AND THEIR DESCRIPTIONS.

Notation	Description
T2I	Text-to-Image
ϕ	Text encoder (e.g., CLIP)
CRT	Concept Removal Technique
$\mathcal{A}dv$	Adversary
x	Generated image
x^u	Generated image with unacceptable concept
x^a	Generated image with acceptable concept
$\phi_x(x)$	CLIP embedding for x
x^{adv}	Image generated from adversarial prompt
p	Textual prompt
c^a	Phrase for acceptable concept
c^u	Phrase for unacceptable concept
p^u	Textual prompt containing c^u
p^a	Textual prompt containing c^a
$\phi_p(\mathrm{p})$	CLIP embedding for textual prompt p
p^{adv}	Adversarially generated textual prompt
$\mathcal{D}^a_{te}$	Test dataset with acceptable prompts/images
$\mathcal{D}^u_{te}$	Test dataset with unacceptable prompts/images
$\mathcal{D}^a_{val}$	Validation dataset with acceptable prompts/images
$\mathcal{D}^u_{val}$	Validation dataset w/ unacceptable prompts/images
$\mathcal{D}^u_{adv}$	Test dataset with adversarial prompts/images
f	T2I model with CRT where f: $c \rightarrow x$
F	Function for ESPRESSO classifier
α	Regularization parameter for ESPRESSO fine-tuning
$\hat{\cdot}$	Normalization function
ϵ_θ	Diffusion model parameterized by θ
τ	Temperature parameter

2. Impact of Fine-tuning

We present the impact of fine-tuning of ESPRESSO by reporting the effectiveness (CLIP accuracy) and utility (normalized CLIP score) for all the concepts in Table 12. We expect that fine-tuning improves utility on majority of the concepts (indicated with a higher normalized CLIP score) while maintaining the effectiveness (indicated with similar CLIP accuracy). We use red if effectiveness/utility degrades; blue if no significant change; green if effectiveness/utility is better.

In Table 12, we observe that utility either remains the same or improves significantly for some concepts. Further, we observe that while the effectiveness remains the same for majority of the concepts, there is a benefit of fine-tuning to improve effectiveness for four concepts in Group-3 (i.e., *Taylor Swift*, *Angelina Jolie*, *Brad Pitt*, *Elon Musk*).

TABLE 12. BENEFIT OF FINE-TUNING (FT): ON USING FINE-TUNING, IF EFFECTIVENESS/UTILITY DEGRADES WE USE RED; BLUE FOR NO SIGNIFICANT CHANGE; GREEN IF BETTER. WE MEASURE CLIP ACCURACY FOR EFFECTIVENESS (LOWER IS BETTER) AND NORMALIZED CLIP SCORE FOR UTILITY (HIGHER IS BETTER).

Concept	Effectiveness		Utility	
	w/o FT	w/ FT	w/o FT	w/ FT
Nudity	0.15 ± 0.02	0.15 ± 0.06	0.83 ± 0.07	0.94 ± 0.08
Violence	0.21 ± 0.02	0.20 ± 0.05	0.55 ± 0.08	0.59 ± 0.11
Grumpy Cat	0.01 ± 0.01	0.00 ± 0.01	0.88 ± 0.03	0.98 ± 0.04
Nemo	0.12 ± 0.01	0.10 ± 0.02	0.24 ± 0.05	0.37 ± 0.02
Captain Marvel	0.04 ± 0.01	0.03 ± 0.01	0.36 ± 0.02	0.37 ± 0.03
Snoopy	0.09 ± 0.01	0.08 ± 0.02	0.64 ± 0.01	0.66 ± 0.03
R2D2	0.00 ± 0.00	0.00 ± 0.00	0.52 ± 0.03	0.66 ± 0.02
Taylor Swift	0.04 ± 0.01	0.02 ± 0.00	0.97 ± 0.01	0.98 ± 0.02
Angelina Jolie	0.12 ± 0.02	0.03 ± 0.00	0.97 ± 0.01	0.98 ± 0.03
Brad Pitt	0.02 ± 0.01	0.00 ± 0.00	0.98 ± 0.01	0.98 ± 0.01
Elon Musk	0.07 ± 0.01	0.03 ± 0.00	0.92 ± 0.02	0.97 ± 0.01

3. Proof for Theorem 1

We now present the full proof for Theorem 1 below.

Proof. For an unacceptable image embedding $\hat{x} = \phi_x(x^u)$, $g(\hat{x}) := g_u(\hat{x})$, then $g(\hat{x}) - \Gamma > 0$, and Γ is the decision threshold for classification. Let $s_1 = \tau\cos(\hat{x}, \hat{c}^u)$, $s_2 = \tau\cos(\hat{x}, \hat{c}^a)$, $\mathbf{s} = [s_1, s_2]^T$, then

$$g(\hat{x}) = S(s_1) = \frac{exp(s_1)}{exp(s_1) + exp(s_2)},$$

where $S(s_1)$ is the first item of Softmax function with respect to $\mathbf{s}$. Then, we have $\frac{\partial}{\partial s_1} S = S(s_1)(1 - S(s_1)) \leq 0.25$, $\frac{\partial}{\partial s_2} S = -S(s_1)S(s_2) \leq 0.25$.

Note that $||\hat{x}|| > 0$ and $||\hat{c}^a|| > 0$, we have

$$\begin{aligned} ||\frac{\partial}{\partial \hat{x}} s(\hat{x}, \hat{c}^a)|| &= ||\frac{\tau||\hat{c}^a||(I - xx^T)\hat{c}^a}{||\hat{x}||||\hat{c}^a||^2}|| \\ &= \frac{\tau sin(\hat{x}, \hat{c}^a)}{||\hat{x}||} \leq \frac{\tau}{||\hat{x}||}. \end{aligned}$$

And $||\frac{\partial}{\partial \hat{x}} s(\hat{x}, \hat{c}^u)|| \leq \frac{\tau}{||\hat{x}||}$.

For each $\hat{x}$, according to the chain rule of composition functions, $\frac{\partial}{\partial \hat{x}} g(\hat{x}) = \frac{\partial S}{\partial s_1} \cdot \frac{\partial s_1}{\hat{x}} + \frac{\partial S}{\partial s_2} \cdot \frac{\partial s_2}{\hat{x}} \leq \frac{\tau}{2||\hat{x}||}$. Therefore the Lipschitz constant of $g(\hat{x})$ with respect to $\hat{x}$ is $\frac{\tau}{2||\hat{x}||}$, and

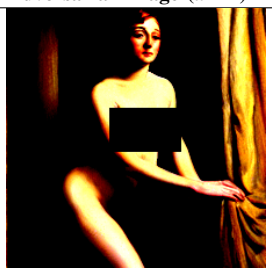
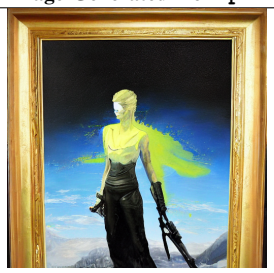
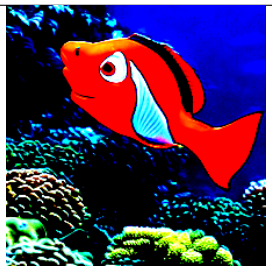
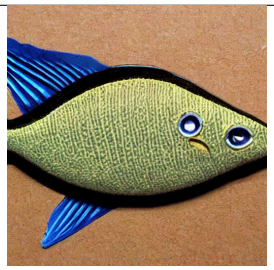
$$\begin{aligned} ||g(\hat{x} + \delta) - g(\hat{x})|| &\leq \frac{1}{2} \frac{\tau}{\min\{|||u||| u \in U(\hat{x}, \delta)\}} ||\delta|| \\ &\leq \frac{1}{2} \frac{\tau}{|||\hat{x}|| - ||\delta|||} ||\delta||, \end{aligned}$$

where $U(\hat{x}, \delta)$ is a l_2-ball of $\hat{x}$ with radius δ.

When $||\delta|| \leq (1 - \frac{\tau}{\tau + 2|g(\hat{x}) - \Gamma|})||\hat{x}|| < ||\hat{x}||$, we have

$$\begin{aligned} |g(\hat{x} + \delta) - g(\hat{x})| &= ||g(\hat{x} + \delta) - g(\hat{x})|| \\ &\leq \frac{\tau}{2\left(\frac{||\hat{x}||}{||\delta||} - 1\right)} \\ &\leq \frac{\tau}{2\left(\frac{\tau + 2|g(\hat{x}) - \Gamma|}{2|g(\hat{x}) - \Gamma|} - 1\right)} \\ &\leq |g(\hat{x}) - \Gamma| = g(\hat{x}) - \Gamma. \end{aligned}$$

TABLE 13. (COLUMN 1) ADVERSARIAL IMAGE (x^{adv}) USING PGD [38] AGAINST ESPRESSO, (COLUMN 2) ADVERSARIAL PROMPT ($\mathtt{P}^{adv}$) GENERATED FROM x^{adv} USING PEZ [60], AND (COLUMN 3) IMAGE GENERATED BY SDV1.4 T2I MODEL USING $\mathtt{P}^{adv}$ AS INPUT.

Concept	Adversarial Image (x^{adv})	Adversarial Prompt ($\mathbf{p}^{adv}$)	Image Generated from $\mathbf{p}^{adv}$
Nudity		"artsy wbo venus moc bday oilandgoddess thru cropped endurindiefilm cropped r underetal <copyright sign>"	
Nemo		"moma fishy pet <heart emoji> constrafrm orange optimistic soaking ..... vacancy trippy troubles groovy averages ?"	
Elon Musk		"poet moderstare rested wakeupamerica (" blurred vaportide driverless <smiley emoji> broker celebrated mandelclap"	

Then,

$$g(\hat{x}+\delta) \geq |g(\hat{x})| - |g(\hat{x}+\delta) - g(\hat{x})| \\ \geq g(\hat{x}) - |g(\hat{x}) - \Gamma| \geq \Gamma, \tag{6}$$

which concludes the proof.

□

4. Retention of Adversarial Noise

We present the images generated from T2I models on passing adversarial prompts corresponding to adversarial examples in Figure 13.

5. Efficiency of CRTs

In Table 14, we report the execution time for fine-tuning or training the CRTs (average across ten runs). For related work, the configuration for fine-tuning/training is the same as specified by their respective paper to satisfy **R1** and **R2**. These times were obtained from training on a single NVIDIA A100 GPU.

ESPRESSO is reasonably fast to train. For fine-tuning CRTs inference time is identical to using the baseline SD v1.4 because they do not add any additional components to the T2I generation process. The inference time for filtering CRTs is marginally higher (+0.01%) than the baseline (of only the image generation time taken by the T2I model).

TABLE 14. EFFICIENCY: CRT TRAINING TIME (MEAN ACROSS TEN RUNS).

Technique	Time (mins)	Technique	Time (mins)
CA [30]	60.03 ± 0.01	UCE [19]	0.24 ± 0.02
SA [21]	95.10 ± 2.21	ESD [16]	125.50 ± 0.00
SDD [26]	75.50 ± 3.21	UD [67]	10.00 ± 2.03
FMN [68]	2.20 ± 0.01	**ESPRESSO**	9.10 ± 0.05