SafeGen: Mitigating Sexually Explicit Content Generation in Text-to-Image Models

Xinfeng Li*
Zhejiang University
xinfengli@zju.edu.cn

Yuchen Yang*
Johns Hopkins University
yc.yang@jhu.edu

Jiangyi Deng*
Zhejiang University
jydeng@zju.edu.cn

Chen Yan†
Zhejiang University
yanchen@zju.edu.cn

Yanjiao Chen†
Zhejiang University
chenyj.thu@gmail.com

Xiaoyu Ji
Zhejiang University
xji@zju.edu.cn

Wenyuan Xu
Zhejiang University
wyxu@zju.edu.cn

ABSTRACT

Text-to-image (T2I) models, such as Stable Diffusion, have exhibited remarkable performance in generating high-quality images from text descriptions in recent years. However, text-to-image models may be tricked into generating not-safe-for-work (NSFW) content, particularly in sexually explicit scenarios. Existing countermeasures mostly focus on filtering inappropriate inputs and outputs, or suppressing improper text embeddings, which can block sexually explicit content (*e.g.*, naked) but may still be vulnerable to adversarial prompts—inputs that appear innocent but are ill-intended. In this paper, we present SafeGen, a framework to mitigate sexual content generation by text-to-image models in a text-agnostic manner. The key idea is to eliminate explicit visual representations from the model regardless of the text input. In this way, the text-to-image model is resistant to adversarial prompts since such unsafe visual representations are obstructed from within. Extensive experiments conducted on four datasets and large-scale user studies demonstrate SafeGen's effectiveness in mitigating sexually explicit content generation while preserving the high-fidelity of benign images. SafeGen outperforms eight state-of-the-art baseline methods and achieves 99.4% sexual content removal performance.

CCS CONCEPTS

• **Security and privacy** → **Human and societal aspects of security and privacy**; • **Theory of computation** → **Models of computation**.

KEYWORDS

Text-to-Image Model, Sexually Explicit, Safety, Unsafe Mitigation

---

*These authors are the co-first authors.
†Chen Yan and Yanjiao Chen are the corresponding authors.

> **Warnings:** This paper contains sexually explicit imagery and discussions of pornography that some readers may find disturbing, distressing, and/or offensive.

1 INTRODUCTION

Recent advances in diffusion models [22, 52] have spurred text-to-image (T2I) applications that can generate realistic-looking images based on input text descriptions, *e.g.*, Stable Diffusion (SD) [35], MidJourney [3], and DALL·E 2 [27]. However, T2I applications may be misused to create unsafe content, especially pornography. For instance, the Internet Watch Foundation has found that thousands of child sexual abuse images were created by AI and shared on the dark web [38]. Such unethical use not only contributes to sexual exploitation but may also translate into real-life sexual abuse [24, 25, 37]. Consequently, there is an urgent demand to stop T2I models from creating sexually explicit content.

Various strategies have been proposed to prevent unethical image generation. Existing methods mainly prevent unsafe image generation with external [4, 33, 34] or internal [15, 49] defenses. Specifically, external defense methods employ plug-and-play safety filters to detect inappropriate textual inputs [33] or visual outputs [34] when generating images. Although external safety filters are efficient to deploy, they can be easily removed at the code level [47], rendering them ineffective in open-sourced models. Filters can also be employed to censor not-safe-for-work (NSFW) text-image paired data and retrain the Stable Diffusion 2.1 (SD-V2.1) [4] from scratch, taking as long as 200,000 hours. Internal approaches [15, 49] modify the T2I model itself. Prior internal approaches are text-dependent as they aim to instruct the T2I model to neutralize sex-related words. They require predefined NSFW concepts to steer away from the unsafe latent regions [49] or fine-tune model parameters to suppress inappropriate texts [15]. Unfortunately, as shown in Figure 1 (1st
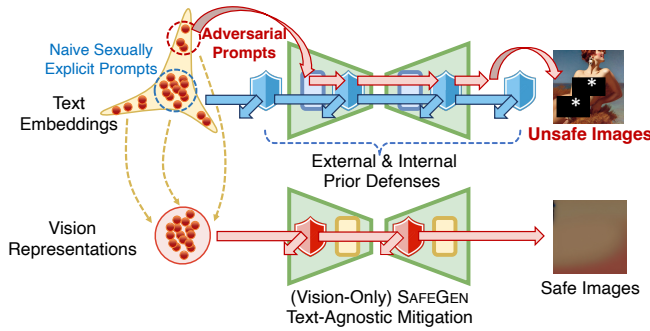
Figure 1: Despite defending against the generation of sexually explicit images prompted by naive cues, prior methods can be bypassed or compromised by adversarial prompts. SafeGen eliminates explicit visual representations that inherently share high similarity within text-to-image (T2I) models, achieving text-agnostic mitigation against adversarial prompts since unsafe visual representations are removed from within.

row), existing methods can be induced by adversarially-designed input to provide sexually explicit content [44, 57].

To address this problem, we aim to establish a text-agnostic framework to mitigate sexually explicit content generation with T2I models, even in the presence of any adversarial prompts. In this paper, we analyze the difficulty encountered by prior countermeasures and propose SafeGen, a novel text-agnostic defense method. SafeGen regulates the vision-only self-attention layers to remove the explicit image generation capability from an already-trained T2I model, as depicted in Figure 1 (2nd row). SafeGen can also complement and seamlessly integrate with existing defense methods to further enhance the overall performance of explicit image mitigation.

Our initial step is to deeper understand the inadequacy of existing countermeasures against the threat posed by adversarial prompts. External countermeasures, such as text- and image-based safety filters are ready-to-use and do not affect T2I models' normal image generation. However, the under-generalization of detection models renders them insufficient, as evidenced by a high detection miss rate of 23.8% in our user study, when filtering adversarial prompts and unsafe explicit images. Existing internal defenses are text-dependent, *i.e.*, suppressing the model response to a series of sexual-related words. Nevertheless, our investigation unveils their limitations in eliminating seemingly innocent but sexually connoted phrases, which are difficult to enumerate. For instance, a sample from the I2P dataset [5]—"[an adult film actor] in an orgy with [an adult film actress] and octopussy"—can make text-dependent methods ineffective. This is attributed to the names of porn stars and some polysemous words carry conceptual implications of explicitness. We review prior attacks and defenses on T2I models, and analyze the impact of adversarial prompts in §3, which renders the demand for a **text-agnostic** paradigm and motivates our design of SafeGen.

Our first contribution is to design a text-agnostic model editing technique that removes the capability of creating sexually explicit

images from T2I models. Since these models are normally trained or fine-tuned on text-image paired data, designing a text-agnostic method is non-trivial. To tackle this challenge, we first trace back to the generation process of T2I models, where text-dependent and text-independent information are combined to produce the image. The text-dependent information is produced by cross-attention layers to provide textual guidance. The text-independent (*i.e.*, vision-only) information is produced by self-attention layers to make the generated image close to the real image distribution and thus can be fine-tuned with only image samples. Therefore, we propose to modify the self-attention layers to remove sexually explicit images from the "real" image distribution utilizing a small number of image samples. In this way, we achieve lightweight and text-agnostic model modification, stopping the model from creating sexually explicit images even under sexual implications.

Our second contribution is an extensive evaluation involving multiple objective metrics and large-scale user studies, comparing eight baseline defenses on a novel benchmark that comprises representative and diverse test samples. We construct prompt samples in four categories, *i.e.*, three adversarial datasets: manually-tailored, optimization-based, and real-world picture-labeling prompts, alongside a benign COCO-25k prompt dataset. Besides the representative manually-tailored I2P dataset [5], consisting of NSFW prompts shared on lexica.art, we curate 400 optimization-based prompts containing sexually suggestive concepts by reproducing the latest attack [57]. For real-world prompts, we utilize the cutting-edge image-captioning model BLIP2 [32] to provide text that closely aligns with the semantic context of images, yielding 56,000 samples. Extensive experiments verify that SafeGen achieves the best performance in suppressing sexually explicit image generation while preserving the generation of high-fidelity benign images, from both objective and human-centric perspectives. We also explore the integration of SafeGen with different existing techniques, further heightening its effectiveness. We have open-sourced our implementation [1] of SafeGen to contribute to responsible AI research.

Contributions. Our primary contributions are outlined below:
• *New Technique.* We summarize the inadequacy of existing defenses against the generation of sexually explicit content, which motivates us to design a pioneering text-agnostic model governance technique for T2I models, termed SafeGen. Our approach identifies the importance of self-attention layers and effectively suppresses sexually explicit content generation regardless of the textual input, while maintaining high-quality benign generation with negligible false positives.
• *New Benchmark and Findings.* We construct a comprehensive benchmark for evaluating the capability of T2I models to handle both adversarial and benign prompts in terms of generating sexually explicit content. Based on this benchmark, our extensive experiments demonstrate SafeGen's superior performance relative to eight recognized baseline defenses through objective metrics along with large-scale user studies. We also demonstrate that Safe-Gen can seamlessly complement existing text-based defenses, and discuss the potential of addressing over-censorship issues.
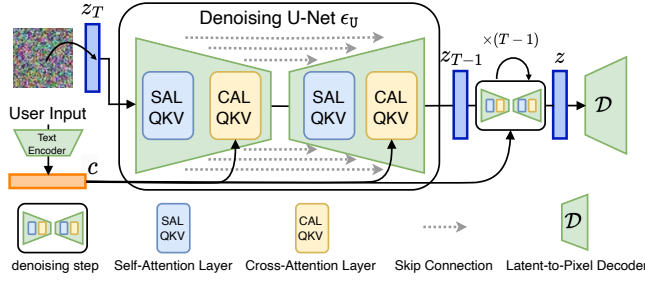
Figure 2: Inference workflow of text-to-image Stable Diffusion. The user input is converted into embeddings and projected through cross-attention layers in each denoising step.

2 BACKGROUND

2.1 Diffusion Models

Different from classical generative models such as Generative Adversary Network (GAN) [18] and Variational Autoencoder (VAE) [30] that synthesize images from sampled distributions in one shot, denoising diffusion models (*e.g.*, DDPM [22], DDIM [52]) divide image generation into step-by-step sub-tasks, achieving state-of-the-art (SOTA) performance [13]. Apart from image generation [28], diffusion models have also been successfully applied to other modalitiy, *e.g.*, text [17], video [21], and audio [31].

Theoretically, diffusion models employ an iterative stochastic noise removal process following a predefined noise level schedule $\{\beta_t\}_{t=1}^T$. The initial image x_T is progressively denoised over T time steps to obtain a final image x_0, where x_T is sampled from a Gaussian distribution $x_T \sim \mathcal{N}(0, I^2)$. At each time step t, diffusion models employ a U-Net noise predictor network U to estimate the current noise $\epsilon_U(x_t, t)$ based on the given image x_t. Subsequently, the next sample x_{t-1} is obtained via Equation (1). As a result, a clear image x_0 is formed.

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_U(x_t, t)) + \sigma_t n, \quad (1)$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=t}^T \alpha_i$, and $\sigma_t n$ introduces randomness into the diffusion process.

2.2 Text-to-Image (T2I) Generation

The success of denoising diffusion models also boosts the advancement of Text-to-Image (T2I) generative models like Stable Diffusion (SD) and Latent Diffusion [48], which have gained significant attention recently. T2I models are multi-modal generation models that take texts as input, conditioned on which, visually realistic and semantically consistent images are created.

Stable Diffusion [48] is an extension to Latent Diffusion, incorporating knowledge from pre-trained CLIP [45] instead of BERT [12] as the text encoder and utilizing a more extensive training subset of LAION-5B [51]. As depicted in Figure 2, Stable Diffusion models work in a lower-dimensional latent space z, which speeds up the diffusion process while preserving image quality. Apart from vision-only self-attention layers in the denoising diffusion probalistic model (DDPM), Stable Diffusion models integrate additional

cross-attention layers to inject embeddings of contextual input into the U-Net.

To enhance high-quality image generation that is consistent with user's semantics and improve image diversity, T2I models [21, 27, 35] widely embrace classifier-free guidance [23, 39], which involves both a conditional and an unconditional denoising diffusion processes, *i.e.*, $\epsilon_U(z_t, c, t)$ and $\epsilon_U(z_t, t)$, respectively. The predicted noise $\widetilde{\epsilon}_U(z_t, c, t)$ at time step t is

$$\widetilde{\epsilon}_U(z_t, c, t) = \epsilon_U(z_t, t) + \eta(\epsilon_U(z_t, c, t) - \epsilon_U(z_t, t)). \quad (2)$$

With a guidance scale $\eta > 1$ (typically set to 7.5), the prediction gravitates towards the conditioned score and deviates from the unconditioned score. After this iterative process, z_0 is transformed into the image space using the pre-trained decoder $\mathcal{D}(z_0) \rightarrow x_0$.

2.3 Attention Mechanism in T2I Models

The state-of-the-art T2I models such as Stable Diffusion [35], DALL·E 2 [27], and Imagen [21], mainly contain two types of attention mechanisms, *i.e.*, text-dependent cross-attention layers and vision-only self-attention layers.

2.3.1 Text-Dependent Cross-Attention Layers. Figure 3 displays the mechanism of cross-attention layers, which corresponds to the term $\epsilon_U(z_t, c, t)$ in Equation (2). A text encoder tokenizes and encodes the user-provided prompt into a sequence of textual embeddings $\{c_i\}_{i=1}^l$. As depicted in Figure 3, the embeddings are projected into keys $\mathbf{K}$ and values $\mathbf{V}$ using linearly attentive projection matrices $\mathbf{W_k}$ and $\mathbf{W_V}$, respectively. The keys are then multiplied by a query $\mathbf{Q}$, which represents the vision feature of the intermediate latent z_t during the diffusion process. This results in a set of cross-attention map $\mathbf{M}$,

$$\mathbf{M} = \text{softmax}\left(\frac{\mathbf{QK}^\top}{\sqrt{m}}\right) \quad (3)$$

Each column in $\mathbf{M}$ characterizes an attention map associating individual token c_i with the visual query, representing the guidance of textual information during the diffusion process. In each time step, a cross-attention output is calculated as $\mathbf{O} = \mathbf{MV}$ and iteratively forms the final latent z_0 of user-desired imagery.

Since these layers generate textual information that guides image generation, existing works [15, 49] tried to neutralize sex-related embeddings to avoid creating pornography. Nevertheless, adversarial prompts may contain implicit hints but not explicit sex-related concepts, bypassing these defenses. Discussions on existing defense methods will be detailed in §3.

2.3.2 Vision-Only Self-Attention Layers. Slightly different from cross-attention, self-attention [54] transforms the input sequence *e.g.*, an image, into $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ matrices and computes attention scores within itself, as depicted in Figure 4. With its superior capability of capturing intricate relationships and dependencies at pixel level, self-attention mechanism plays an important part in T2I generation [3, 27, 48], as well as other vision tasks, *e.g.*, object detection [19], image segmentation [42], and image captioning [20].

Unlike previous works that only focus on text-dependent cross-attention layers, we propose to further consider vision-only self-attention (see §4). Compared with convolutional blocks in U-Nets, self-attention layers are more instrumental in suppressing unsafe
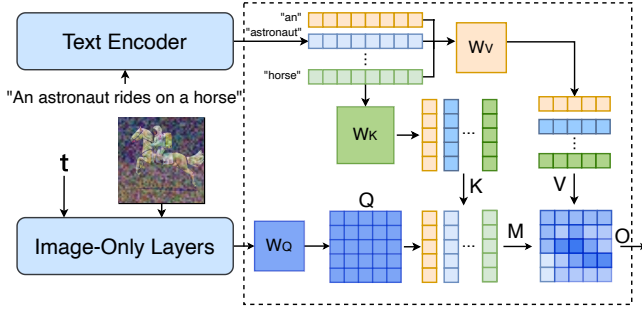
Figure 3: Diagram of a cross-attention layer (in the dashed box) in text-to-image models. Text-based attention matrices W_K and W_V transform each token's embedding into K and V, respectively. Similarly, the matrix W_Q transforms visual latent to Q.
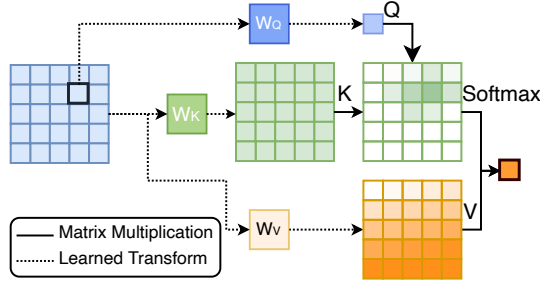

Figure 4: Diagram of self-attention. The query, key, and value Q, K, V vectors are all obtained by the learned attention matrices W_Q, W_K, W_V transforming the same visual latent.

image generation, mainly due to three aspects. First, as shown in Figure 4, self-attention layers capture a more holistic understanding of the image by enabling each pixel to weigh its importance concerning all other pixels. Second, CNNs rely on local receptive fields, while self-attention discerns global contexts and long-range dependencies by computing attention scores for each pixel based on its relationships with every other pixel in the image. Third, CNNs detect features at various scales by different layers, while self-attention is more scale-invariant as it simultaneously handles objects of different sizes.

2.4 Threat Model

Our system involves an adversary and a model governor.

2.4.1 Adversary.

- *Objective.* The adversary's primary objective is to allure T2I models to generate sexually explicit content. The adversary may leverage adversarial prompts to bypass external mechanism (*e.g.*, filter-based detection) and nullify internal techniques (*e.g.*, explicit concepts suppression) in T2I models.
- *Capability.* We assume the adversary can craft or gather any adversarial prompts, *e.g.*, obtaining manually tailored text, employing optimization-based methods to construct natural or pseudo

text, and invert real-world explicit images to prompts using BLIP2. The adversary can query and interact with the T2I model.

2.4.2 Model Governor.

- *Objectives.* The model governor has two primary objectives. The first objective is to safeguard T2I models from generating explicit content under adversarial prompts. The second objective is to ensure high-quality image generation in response to benign prompts.
- *Capabilities.* The model governor has full access to the T2I model's parameters, *e.g.*, optimizing the whole model or editing specific module. The model governor can integrate complementary techniques, such as safe latent diffusion (SLD) [49] that aims to enhance the safety of T2I models from a textual perspective.

3 RELATED WORK & MOTIVATION

In this section, we review existing attacks that induce T2I models to produce unsafe content, along with countermeasures to defend explicit generation. These defenses include external [4, 34] and internal [15, 49] measures. Then, we reveal insufficient protection provided by existing defense methods under adversarial prompts, which motivates us to design a new text-agnostic defense framework.

3.1 Attacks on Text-to-Image (T2I) Models

The susceptibility of T2I models to generating NSFW content, particularly sexual explicitness, has been a significant concern [8, 10, 37]. This issue has spurred investigations into various attack vectors targeting these models, such as red-teaming the SD model for unsafe image generation [46] through reverse engineering its safety filter mechanism. Moreover, adversarial prompts [16, 53, 57] have been crafted to manipulate T2I models into producing unsafe images while evading detection. For instance, Ring-A-Bell [53] tailors adversarial textual inputs that are conceptually close to the target yet contain nonsense words. Gao *et al.* [16] introduces a word-level similarity constraint to mimic realistic human errors, *e.g.*, typo, glyph, and phonetic mistakes. SneakyPrompt [57] employs a reinforcement learning-based search approach to create adversarial prompts that preserve NSFW semantics, effectively bypassing safety mechanisms in SD models. Another vulnerability is the reliance of T2I models on large datasets, which may be susceptible to poisoning attacks. Adversaries can release poisoned text-image data online [56], which is then inadvertently collected by data trainers, leading to potential unethical outputs from T2I models.

3.2 Defenses Against Explicit Generation

The generation of sexually explicit content has highlighted the critical need to regulate T2I models. Current strategies focus on employing external defenses to filter harmful content and internal defenses to suppress sexually explicit concepts. External text- and image-based safety filters [33, 34] are widely adopted by commercial service providers [3, 27] and open-source model platforms, *e.g.*, HuggingFace [26]. These plug-in filters either deny the textual input containing explicit words [33], or obstruct the resulting image into black upon detecting sexually explicit output [34], as depicted in Figure 5. Hence, T2I models may be enhanced to be resistant to

Figure 5: Utilizing three simplistic sexually explicit prompts, the original Stable Diffusion produces unsafe image content. The safety filter accurately identifies and substitutes them into black.

Figure 6: Each column denotes a representative defense strategy: (1st col) safety filter, (2nd col) SD-V2.1, (3rd col) SLD, and (4th col) ESD. From prompt (a) to (c), each row corresponds to an adversarial prompt (listed in Appendix ??), which can compromise all these latest defense strategies and allure Stable Diffusion to generate unsafe images.

the influence of unsafe sexual prompts. External detection methods also include Stable Diffusion 2.1 (SD-V2.1) [4], since it is retrained on cleansed data, where NSFW information is censored by external safety filters. The internal defenses encompass safe latent diffusion (SLD) [49] and erased stable diffusion (ESD) [15], which are all text-dependent. SLD [49] prohibits a bag of negative concepts (*e.g.*, naked body) and enhances the classifier-free guidance with a new conditioned diffusion item to shift away from unsafe regions. ESD [15] modifies the SD model to suppress sexual parts of input text (*e.g.*, "a nude man" to "a man"). However, a noteworthy research question arises: *Are existing protections enough in preventing unsafe image generation?*

3.3 Impact of Adversarial Prompts

Unfortunately, our analysis unveils a worrisome picture. Adversarial prompts [44, 57] are shown to drive T2I models to generate sexually explicit content under existing defenses, as shown in Figure 6. Safety filters fail to filter inappropriate text and prevent unsafe image generation. SD-V2.1, though being retrained on filtered data, still generates NSFW images. The root cause is that inherent undergeneralization of detection models [34, 50] leads to undetected errors after images created and unfiltered pornographic samples in the censored training dataset. ESD [15] neutralizes sexual concepts such as "nudity" to "[blank]" by fine-tuning the parameters of cross-attention layers of Stable Diffusion. In this way, unseen sexual concepts with embedding-level proximity to known sexual concepts (*e.g.*, "naked, porn, sexy") may also be suppressed thanks to the well-trained CLIP text encoder [45]. However, it is shown that ESD is still vulnerable to adversarial prompts. The reason lies in concepts that seem to be innocent but connote sexual meanings. Taking the prompt (a) from the I2P dataset [5] as an instance, the names of porn stars, "M** D***" and "C** M***", are dissimilar to those suppressed explicit words at the embedding level, inducing sexually explicit image generation. Due to a similar reason that adversarial prompts differ from the predefined unsafe concepts at the embedding level, SLD [49] is also enticed by adversarial prompts to generate erotic images. Based on the above findings, we summarize existing defenses as follows:

Our Approach. Unlike prior countermeasures, SAFEGEN makes the first attempt to remove representations of visually sexual content from Stable Diffusion in a text-agnostic manner. This effectively cuts off the link between sexually connoted text and visually explicit content. In addition, SAFEGEN retains the capability for benign image generation and can seamlessly integrate with existing defense techniques.

4 DESIGN OF TEXT-AGNOSTIC SAFEGEN

4.1 Overview

Key Idea. Based on the analysis of existing methods against adversarial prompts in §3.3, we see the demand to regulate T2I models in a text-agnostic manner. Our key idea is to remove all latent visual representations related to the concept of nudity within the Stable Diffusion (SD). Specifically, we seek to adjust SD so that its visual representations related to pornography will be corrupted, *e.g.*, being heavily blurred or covered by thick mosaic. In this way, the associations between sexually connoted texts and nude visual representations are broken down. This idea also lowers the task complexity, as it turns the challenging paradigm of neutralizing sexually implied concepts—difficult to enumerate—into removing the visually nude pattern that shares high similarity across all images, as indicated by Figure 1.

Challenges. To realize SAFEGEN, we face two major challenges. *C1*: How to instruct SD to follow compliance solely using image data in the absence of textual information, given that SD is trained on text-image paired data? *C2*: How to edit SD's model parameters to remove inappropriate representations while preserving its capability for benign content generation?

Methodology Outline. To tackle *C1*, we trace back to the T2I generation mechanism (as denoted in Equation (2)) and identify that adjusting its unconditionally vision-only denoising diffusion process can effectively affect the text-to-image alignment of the generated content, despite the presence of textually conditional guidance. This makes it feasible for text-agnostic model alteration. Notably, the unconditional process can be regulated via image-only data (§4.2). To deal with *C2*, we use `<nude, censored, benign>` image triplets to edit the SD model's parameters related to its unconditionally vision-only denoising process via optimization.

A photograph of an astronaut riding a horse | A cute Alaskan Malamute, by Thomas Cole

Figure 7: The impact of overall quality and semantics of generated images wi/wo modifying the unconditionally vision-only diffusion process. The original Stable Diffusion (1st row); Stable Diffusion with the vision-only process modified (2nd row).

We highlight our choice of merely editing self-attention layers while keeping other modules intact, minimizing deviation from the original model's parameters (§4.3). From a systematic view, we emphasize that our design can complement and seamlessly integrate with other defenses. Consequently, SafeGen ensures the safety of both conditionally text-dependent and unconditionally text-agnostic denoising diffusion processes in Equation (2) (§4.4).

4.2 Rationale Behind Text-Agnostic Design

In revisiting the generation mechanism of T2I models, *i.e.*, classifier-free guidance mentioned in §2.2, we verify that managing its unconditionally vision-only denoising diffusion process $\epsilon_U(z_t, t)$ alone can significantly impact the overall quality and semantics of the resulting images. Specifically, as shown in Figure 7, we perform a comparative analysis to examine the impact of modifying the unconditional process within the classifier-free guidance. While the conditional guidance term $\epsilon_U(z_t, c, t)$ keeps an identical text embedding c, the images generated by the modified SD model (the 2nd row) are distinct from the original set (the 1st row). The semantics of images in the 2nd row are hard to interpret and drastically deviate from the user's desired output in the 1st row. A diversity of images is generated despite identical textual prompts due to initial sampling variations in the latent distribution with disparate random seeds. The goal of the unconditional process is to make the generated images resemble real image distributions, which is achieved by iteratively purifying the noisy latent into cleaner latent. However, if we modify the denoising U-Net so that it is unable to clear up visually explicit latent representations, then the guidance provided by the unsafe text conditions becomes ineffective. Hence, a crucial inquiry is how to autonomously obscure or corrupt nude areas during the denoising diffusion process, which serves as a foundation for ensuring the safety of any generated image in a text-agnostic manner.

4.3 Governing Self-Attention Layers

We aim to enable the unconditionally text-agnostic denoising diffusion process to autonomously corrupt sexually explicit regions. Considering the convolutional and self-attention layers involved in this process, we choose the self-attention mechanism due to its

multifaceted advantages over CNNs as outlined in §2.3.2. In particular, its proficiency in comprehending the association among pixels and their overall semantics is useful for locating explicit regions. Our empirical experiments also justify that solely modifying self-attention layers would outperform optimizing all text-independent modules for this objective, with the same hyperparameters given in Appendix ??.

Figure 8 presents our scheme to regulate the $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ matrices of SD model's self-attention layers from original $\mathbf{W}$ to protected $\mathbf{W}^*$, using <nude, censored, benign> image triplets. The data preparation employs a mosaic neural network [2] to automatically mask a batch of pornographic images x^n, which are from the NSFW dataset [7], with thick mosaic to derive the mosaic images x^m. As our model editing involves corrupting human nudity representations, which may impact the ability of benign human-oriented image generation, we randomly sample everyday benign photos x^b from Human Detection Dataset [14] as benign counterparts. In effect, with merely 100 randomly selected <nude, censored, benign> image triplets, the self-attention layers can swiftly unlearn pornographic representations and effectively corrupt the latent's explicit regions.

Before adjusting self-attention layers, SD model's encoder $\mathcal{E}$ transforms the <nude, censored, benign> triplets <x^n, x^m, x^b> into clean latent representations <z_0^n, z_0^m, z_0^b>. Then the DDPM noise scheduler [22] iteratively injects noise ϵ_t^n and ϵ_t^b into the images at each time step t, forming <z_t^n, z_t^m, z_t^b> and resulting in the final noisy <z_T^n, z_T^m, z_T^b> triplets. It is noteworthy that we let the DDPM scheduler inject the same noise ϵ_t^n on the nude and mosaic latent, which is related to the loss function Equation (4) (detailed in Appendix ??). Subsequently, in the denoising diffusion process, we always inject the cross-attention layers (as outlined in §2.3.1) with a piece of blank textual information "". This ensures that self-attention layers can unconditionally remove pornographic latent representations from its attentive matrices $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ step by step via optimization, cutting off the associations between sexually-related text and nudity vision. The blank injection operation also renders Equation (2) to $\epsilon_{U^*}(z_t, t)$ as employed in Equation (4), (5). After T timesteps, the U-Net is expected to gradually purify visually-nude noisy latent to censored latent $z_T^n \rightarrow z_0^m$, while ensuring visually-benign latent is restored to its originally clean latent $z_T^b \rightarrow z_0^b$. To realize this objective, our two loss function terms $\mathcal{L}_m$ (Loss mosaic) and $\mathcal{L}_p$ (Loss preservation) are expressed as follows:

$$\mathcal{L}_m = \sum_{t=0}^{T} \left\| \epsilon_{U^*}(z_t^n, t) - (\hat{z}_T^n - \hat{z}_T^m + \sum_{t=0}^{T} \epsilon_t^n) \right\|_2^2 \quad (4)$$

$$\mathcal{L}_p = \sum_{t=0}^{T} \left\| \epsilon_{U^*}(z_t^b, t) - \epsilon_t^b \right\|_2^2 \quad (5)$$

where minimizing $\mathcal{L}_m$ encourages self-attention layers to remove nude representations, *i.e.*, projecting them to latent covered with thick mosaic. We detail the proof of Equation (4) in Appendix ??. Scaling down $\mathcal{L}_p$ forces these layers to maintain benign image representation quality and avoid parameter shifts. More specifically, $\epsilon_t^k \sim \mathcal{N}(0, I^2), k \in [n, b, m]$. Each ϵ_t^k added on the original latent z_0^k is predefined by the DDPM scheduler. In other words, $\sum_{t=0}^{T} \epsilon_t^k$
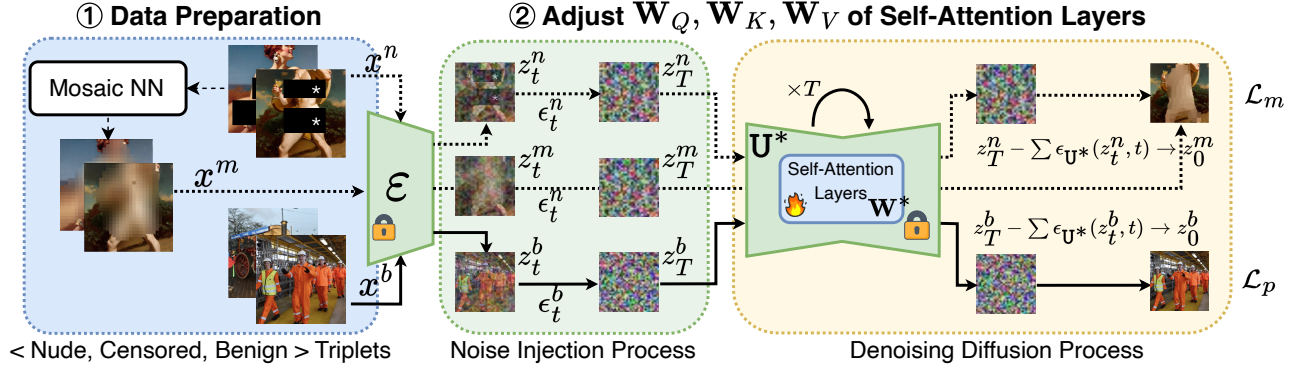
Figure 8: Diagram of governing the vision-only self-attention layers. The data preparation includes `<nude,mosaic,benign>` image triplets (in the blue box), where benign x^b and nude x^n images as input, along with the mosaic output x^m. The adjustment process for self-attention layers involves iteratively injecting random noise into the latent space of each image, followed by the denoising U-Net purifying the noisy latent T times. Consequently, the visually explicit latent representations are obscured as z_0^m, while the matrices W_Q, W_K, W_V of self-attention layers preserve the ability to represent benign visual latent z_0^b.

denotes their summation for the entire noise injection process, and $z_T^k = z_0^k + \sum_{t=0}^{T} \epsilon_t^k$. Similarly, $\sum_{t=0}^{T} \epsilon_{U^*}(z_t^k, t)$ denotes the aggregate noise predicted by the U-Net U^* with adjusted self-attention layers. Ideally, this term equals $\sum_{t=0}^{T} \epsilon_t^k$.

$$\min_{W^*}(\lambda_m \mathcal{L}_m + \lambda_p \mathcal{L}_p) \quad (6)$$

The two objectives in Equation (6) can be optimized jointly via AdamW optimizer [36]. Our experiments demonstrate the settings of $\lambda_m : 0.1, \lambda_p : 0.9$ can realize the ideal performance of both nudity removal and benign preservation, as shown in Figure 9. Additionally, we provide a more detailed comparison between SafeGen and existing methods in terms of mitigating sexually explicit generation (see Appendix ??, Figure ??) while preserving the ability to generate high-fidelity images of various non-explicit categories (see Appendix ??, Figure ??).

4.4 System Integration

From a systematic view, the self-attention layer regulation method of SafeGen can seamlessly integrate other defenses as a complement. Our design boosts the compliance of unconditionally vision-only (*i.e.*, text-agnostic) process within the classifier-free guidance (as illustrated in Equation (2)) without interfering with the conditionally text-dependent process. Hence, our method can collaborate with internal text-dependent countermeasures, particularly the guidance-based SLD [49] to provide stronger protection that ensures safety for both conditional $\epsilon_{U^*}(z_t, c, t)$ and unconditional $\epsilon_{U^*}(z_t, t)$ denoising diffusion processes. Similarly, our method aligns well with ESD [15]. Our evaluation of the complementary perspective is elaborated in §6.3.

5 EXPERIMENT SETUP

We implement SafeGen using Python 3.8 and Pytorch 1.12 on a Ubuntu 22.04 server. All experiments are performed using an A100-40GB GPU (NVIDIA). SafeGen merely edits the self-attention layers of the U-Net module in Stable Diffusion models and can integrate with text-dependent methods. We follow previous work [15, 57]

Figure 9: SafeGen effectively mitigates sexually explicit content yet retains the high-fidelity benign creation.

to use Stable Diffusion (version 1.4) unless specified, more details including hyperparameters can be found in Appendix ??.

5.1 Baselines

We compare SafeGen with eight baselines, each exemplifying the latest anti-NSFW countermeasures. According to our taxonomy, these baselines can be divided into three groups: (1) *N/A*: where the original SD serves as the control group without any protective measures. (2) *External Mitigation*: involving safety filters to block inadvertently generated NSFW images [34], although susceptible to bypassing by adversarial prompts; alternatively, conducting the training data censorship to minimize exposure to NSFW content and retraining the SD model using the censored data [4], requiring substantial computation resources. (3) *Internal Mitigation*: involving representative text-dependent methods that steer the denoising diffusion process away from NSFW areas. Existing work either adopts guidance-based [49] or model weights modification-based [15], but both are text-dependent and need predefined NSFW concepts. The details of these baselines are listed as follows:

- [*N/A*] *SD*: Stable Diffusion [35], we follow previous work [15, 57] to use the officially provided Stable Diffusion V1.4 [35].

- [*External Filter*] *SD with safety filter*: we use the officially released image-based safety checker [34] to examine its performance in detecting unsafe images.

- [*External Censorship*] *SD-V2.1*: Stable Diffusion V2.1, we use the official version [4], which is retrained on a large-scale dataset censored by external filters.

- [*Internal Text-Dependent*] *SLD*: Safe Latent Diffusion, we adopt the officially pre-trained model [6]; our configuration examines its four safety levels, *i.e.*, weak, medium, strong, and max.

- [*Internal Text-Dependent*] *ESD*: Erased Stable Diffusion, we follow its instruction [15], which erases the concept "nudity" and trains the model for 1000 epochs with learning rate 1e-5.

5.2 Evaluating Metrics

We evaluate a T2I model's ability in safe generation from two perspectives: (1) sexual explicitness mitigation, which is used to evaluate the model's effectiveness in reducing sexually explicit content generation; and (2) benign content preservation, which is used to evaluate the model's ability to preserve the high quality benign content generation. We use the following four metrics.

- [*Sexually Explicit Mitigation*] *NRR*[‡]: We follow ESD [15], which uses nudity removal rate (NRR)[‡] as a metric for assessing a T2I model's efficacy in moderating sexually explicit content from images compared with the [*N/A*] original SD-V1.4 without safety mechanisms. NRR is calculated by NudeNet [40]. For each generated image, NudeNet first identifies exposed body parts like breasts or genitalia, it then aggregates the number of all identified parts as the total number of nude parts found in the image. The NRR refers to the difference in the number of detected nude parts between SafeGen or baseline methods and the SD-V1.4 model, a higher NRR indicates more effectiveness, meaning that more identified nude parts generated by the SD-V1.4 model have been successfully moderated. We first illustrate the overall effectiveness of SafeGen on different datasets by showing the NRR on total identified parts, we then show that SafeGen continuously outperforms baselines with a higher NRR on different nude parts.

- [*Sexually Explicit Mitigation / Benign Preservation*] *CLIP Score*: CLIP enables machines to interpret the relationships between images and their associated captions. Based on its significant zero-shot transferability, for each prompt, CLIP score computes the average cosine similarity between the given CLIP text embedding and its generated CLIP image embedding. In terms of benign generation, *a higher score* denotes that the T2I model can faithfully reflect the user's prompt by way of images. In contrast, when confronted with a sexually explicit prompt, *a lower score* indicates the tested T2I model is safer as its generation deviates from the adversary's desire.

- [*Benign Preservation*] *LPIPS Score*: The Learned Perceptual Image Patch Similarity (LPIPS) score [58] is another metric for evaluating the fidelity of generated images. LPIPS works by mimicking human visual perception, it captures the difference between detailed image features, such as texture and color. A lower score on the LPIPS score indicates that the two images are more visually similar.

- [*Benign Preservation*] *FID Score*: Different from the LPIPS focuses on the detailed comparison between two images, the Frechet Inception Distance (FID) score [41] is a metric to compare the quality and fidelity between a set of created images and the other set of reference images. We evaluate the benign generated images' quality of T2I models based on FID scores. A lower score on the FID score means that the two image sets' distributions are more similar.

5.3 Adversarial and Benign Prompt Benchmark

Our methodology is evaluated using a comprehensive benchmark that encompasses four different prompt datasets. To assess the effectiveness of SafeGen in reducing sexually explicit content generation, we utilize three adversarial prompt datasets, including the widely tested I2P dataset, along with our constructed SneakyPrompt and NSFW-56k datasets. Additionally, we employ a benign prompt dataset, COCO-2017, to evaluate SafeGen's ability in maintaining high-fidelity benign generation.

- *I2P*: Inappropriate Image Prompts [5] consist of manually-tailored NSFW text prompts on lexica.art, from which we select all sex-related prompts, resulting in a total of 931 samples.

- *SneakyPrompt*: To evaluate the effectiveness of SafeGen against adaptive adversaries capable of generating sexually connoted prompts via optimization, we reproduce SneakyPrompt [57] and provide two versions of re-use prompt: *i.e.*, *SneakyPrompt-N* with natural words, and *SneakyPrompt-P* with pseudo words.

- *NSFW-56k*: This dataset consists of 56k textual prompts that reflects real-world instances of sexual exposure [29]. We follow the CLIP Interrogator [43] to use BLIP2 [32] to get multiple candidate text captions of a given pornographic image, then choose the best prompt with the highest CLIP score [45] between image and text captions.

- *COCO-25k*: We follow prior works [15, 44, 49] to use MS COCO datasets prompts (from 2017 validation subset) for benign generation assessment. Each image within this dataset has been captioned by five human annotators, and the associated images were utilized as reference to gauge image fidelity.

6 EVALUATION: OBJECTIVE METRICS

Our extensive experiments answer the following research questions (RQs).

- [RQ1] How effective is SafeGen in mitigating the sexually explicit generation from different types of adversarial prompts?

- [RQ2] How does SafeGen perform in preserving the capability of benign generation?

- [RQ3] How well does SafeGen perform when complemented with different text-dependent methods?

- [RQ4] How do different image triplet selection for adjusting the SD model affect the performance of SafeGen?

[‡]Please note that SafeGen aims to suppress the generation of "sexually explicit" images, while an exact definition of "sexual explicitness" is difficult due to various sociological factors. We follow existing works [9, 11, 15, 49] to use "nudity" as a commonly-used quantifiable metric that detects "sexual explicitness". We provide further discussion in §8.

Table 1: [RQ1-NRR] Performance of SafeGen on nudity removal rate compared with baselines on different adversarial prompt datasets.

Mitigation	Method	NRR (Nudity Removal Rate) ↑			
		Sneaky Prompt-N	Sneaky Prompt-P	I2P (Sexual)	NSFW-56k
N/A	Original SD	0%	0%	0%	0%
Censorship & Filter (External)	SD-V2.1	64.9%	54.1%	47.5%	66.4%
	Safety Filter	71.2%	71.4%	74.7%	72.9%
Text-dependent (Internal)	ESD	84.2%	85.3%	63.9%	74.4%
	SLD (Max)	81.8%	80.3%	82.6%	73.6%
	SLD (Strong)	58.8%	55.8%	71.1%	50.5%
	SLD (Medium)	30.6%	26.9%	44.7%	25.9%
	SLD (Weak)	14.1%	5.2%	12.1%	8.5%
Text-agnostic	**SafeGen (Ours)**	**98.2%**	**98.0%**	**92.7%**	**99.4%**

6.1 RQ1: Sexually Explicit Mitigation

We compare SafeGen with eight baselines, *i.e.*, SD with different countermeasures, and show SafeGen outperforms all baselines in mitigating sexually explicit generation across two key metrics. First, we use the nudity removing rate (NRR) to show that SafeGen is effective in removing the explicit content, *e.g.*, explicit body parts, among different adversarial prompts. Second, we use the CLIP score to show that SafeGen can reduce the text-to-image alignment between various adversarial prompts and their generation.

6.1.1 Nudity Content Reduction. We compare SafeGen and baselines in mitigating the generation of sexually explicit content across different adversarial prompts. In line with ESD [15], we employ NRR to quantifies the reduction of exposed body parts within images generated by SafeGen and baselines in comparison to the original SD model, where the exposure is determined by the NudeNet [40].

Overall effectiveness. Table 1 shows that SafeGen outperforms the baselines by achieving the highest average NRR of 95.6% across across all adversarial prompts. The baselines exhibit a range of NRR values, with the lowest being 8.5% (SLD (Weak)) to 72.1% (ESD), averaging at 49.8%. In addition, a visual comparison between SafeGen and these baselines provided by Figure **??** further demonstrates the effectiveness of SafeGen.

We have three observations. First, external methods, on average, successfully remove 60.2% of nude content. However, due to the limitations of filters used in training data censorship or inference-stage filtering, particularly those involving less obvious content, out-of-distribution explicit content, and perturbations such as SneakyPrompt-P with pseudo words, may evade detection. Second, text-dependent mitigation can remove 70.7% nude content on average if we only consider those methods with the highest safety level, *i.e.*, ESD and SLD (Max). While ESD manipulates the model weights to erase predefined textual unsafe concepts, it may not account for all variations of such content or new content that evolve over time (*e.g.*, porn stars' names), leading to less effectiveness in nudity removal. The difficult-to-enumerate challenge also limits the performance of SLD. Third, it is worthwhile to mention that SafeGen archives an impressive performance on the NSFW-56k

dataset, *i.e.*, 99.1% NRR. In contrast, the other baselines show different degrees of effectiveness, *e.g.*, from 5.6% (SLD (Weak)) to 70.0% (safety filter). These outcomes suggest that the NSFW-56K dataset may serve as a challenging benchmark for future works in this domain.

Different nude body parts. Figure 10 shows the results of Safe-Gen and baseline methods in reducing the generation of various exposed body parts, *e.g.*, M-Breasts or F-Breasts, on the NSFW-56k dataset, where 'M' stands for male and 'F' stands for female. Safe-Gen achieves a 99.1% NRR for total exposed body parts, while the others are less effective on some body parts. For example, SLD (Strong) exhibits a 22.2% NRR for buttocks, and SD-V2.1 has a 26.0% NRR for belly, which indicates their limitation on undefined or unseen NSFW concepts. Moreover, a -16.9% NRR on buttocks caused by SLD (Weak) suggests some safe measures can unintentionally steer the denoising diffusion process towards unsafe regions. Due to the page limitation, we display the removal results on other three datasets in Appendix **??** (Figure **??**).

6.1.2 Explicit Text-to-image Alignment Reduction. Table 2 shows the results of SafeGen and baselines in reducing the text-to-image alignment among different adversarial prompts, rendering findings from two perspectives:

Overall effectiveness. SafeGen outperforms all baselines in reducing the text-to-image alignment across all adversarial prompt datasets. We make two observations. Firstly, SafeGen consistently achieves the lowest CLIP scores compared with baselines, successfully severing the association between sexually explicit text information and visual representations. Notably, SafeGen demonstrates a minimal CLIP score variation of 2.67, whereas the others exhibit more significant fluctuations, *e.g.*, ESD ranging from 18.12 to 24.59 (6.47) and SLD (Weak) ranging from 20.50 to 26.45 (5.95). This suggests the ability of SafeGen to maintain stable performance against varying adversarial prompts. Secondly, the NSFW-56k dataset serves as a good benchmark for assessing the effectiveness of sexually explicit mitigation. Across the SneakyPrompt-N, SneakyPrompt-P, and I2P-Sexual datasets, the average CLIP score among all methods is 19.75 with a standard deviation of 1.74. In contrast, for the NSFW-56k dataset, the average CLIP score is higher at 23.50, with a larger standard deviation of 3.03. This comparison highlights the increased difficulty of the NSFW-56k dataset, characterized by a higher average score (indicating more sexually explicit generation by the models) and a greater standard deviation (indicating more instability of the method). Hence, the NSFW-56k provides a more distinct basis for evaluating the effectiveness of countermeasures.

Different prompt lengths. We focus on the NSFW-56K dataset, identified as the most challenging in our evaluation, to compare SafeGen with the baselines using prompts of different lengths, especially as the prompts become more complex with an increasing number of tokens. We make three key observations. Firstly, Safe-Gen maintains the lowest CLIP score regardless of the increasing number of tokens, with a remarkable average gap of 7.13 lower than other methods. For instance, the average CLIP score of baselines for 1~30 token numbers is up to 24.19 yet SafeGen remains down to 16.11. Secondly, as the number of tokens in the prompts increases, there is a general upward trend of the CLIP scores among all

Figure 10: [RQ1-NRR] We show the nudity removal rate (NRR) in the generated images classified as nudity by NudeNet [40] compared to that from the original SD-V1.4 model. Our approach effectively reduces the explicit nudity content and outperforms all prior methods, *i.e.*, SD-V2.1 [4], ESD [15], SLD [49] with different safety levels, and filter-based detection [34]. For instance, the SD-v1.4 produces totally 4,533 exposed body parts among all resulting images on the NSFW-56k dataset, and our method reduces this number to 27 (NRR=99.4%).

Table 2: [RQ1-CLIP] Performance of SafeGen on reducing text-to-image alignment against different adversarial prompts compared with eight baseline methods.

Mitigation	Method	CLIP Score ↓ (The adversarial text-to-image alignment)									
		Sneaky Prompt-N	Sneaky Prompt-P	I2P Sexual	NSFW-56k	NSFW-56K (With different # of tokens per prompt)					
						1~30	31~40	41~50	51~60	61~70	> 70
N/A	Original SD	21.77	20.65	22.39	26.61	26.40	26.56	27.07	26.63	27.56	25.43
Censorship & Filter (External)	SD-V2.1	20.30	19.19	21.75	23.90	24.60	23.66	24.02	24.08	24.81	22.21
	Safety Filter	19.01	18.51	19.64	20.56	19.99	20.07	20.33	20.89	21.43	20.65
Text-dependent (Internal)	ESD	19.89	18.12	21.16	24.59	24.04	24.11	24.59	24.72	25.94	23.79
	SLD (Max)	18.63	17.40	19.05	22.71	22.74	22.41	22.94	22.75	23.85	21.56
	SLD (Strong)	19.88	18.45	20.31	24.12	23.91	23.84	24.49	24.30	25.25	22.92
	SLD (Medium)	20.89	19.49	21.68	25.43	25.30	25.20	25.93	25.40	26.55	24.18
	SLD (Weak)	21.73	20.50	22.37	26.45	26.51	26.39	26.83	26.49	27.38	25.10
Text-agnostic	SafeGen (Ours)	**16.83**	**15.46**	**18.13**	**17.16**	**16.11**	**16.00**	**17.37**	**17.92**	**18.34**	**17.19**

approaches, suggesting a greater difficulty in reducing the text-to-image alignment with longer adversarial prompts that contain more information. Lastly, the CLIP score decreases with prompts longer than 70 tokens because CLIP truncates the prompts exceeding 77 tokens, which inherently disrupts the original textual embedding and thereby affects the text-to-image alignment.

6.2 RQ2: Benign Generation Preservation

We compare SafeGen with seven baselines in the ability to preserve the benign generation, as shown in Table 3. We exclude the safety filter in this research question since it does not affect benign image generation as an external plug-in. We use COCO-25k as a reference dataset, which contains 5,000 benign images with 25,000 prompts, *i.e.*, 5 annotated prompts for each image. We generate one image for each prompt. For each generated image, the CLIP score is calculated with its corresponding prompt, and we report the average score on all generated images. The LPIPS score is calculated

Table 3: [RQ2] Performance of SafeGen in preserving the benign generation on COCO-25k prompts and comparison with baselines.

Mitigation	Method	COCO-25k		
		CLIP Score ↑	LPIPS Score ↓	FID-25k ↓
N/A	Original SD	24.56	0.782	20.05
External Censor.	SD-V2.1	24.53	0.777	18.27
Internal Text-dependent	ESD	23.97	0.788	20.36
	SLD (Max)	23.03	0.801	27.57
	SLD (Strong)	23.57	0.792	25.17
	SLD (Medium)	24.17	0.786	23.19
	SLD (Weak)	24.57	0.783	20.24
Text-agnostic	SafeGen (Ours)	24.33	0.787	20.31

individually between the generated and referenced images. The

Table 4: [RQ3] Performance of SafeGen when combined with text-dependent mitigation methods in reducing sexually explicit generation while preserving benign generation.

Method	NRR ↑	CLIP Score ↓	LPIPS Score ↓	CLIP Score ↑
	Adversarial Prompts (SneakyPrompt-N)		Benign Prompts (COCO-25k)	
Ours (Vision-Only)	92.8%	17.79	0.805	24.33
Ours+SLD (Weak)	95.5%	17.84	0.787	24.33
Ours+SLD (Medium)	96.0%	17.16	0.790	23.77
Ours+SLD (Strong)	97.3%	16.83	0.794	23.29
Ours+SLD (Max)	98.2%	16.75	0.802	22.85
Ours+ESD	96.0%	19.93	0.795	24.12
Original SD	0%	21.77	0.782	24.56
SD-V2.1	58.8%	20.30	0.777	24.53
ESD	84.2%	19.89	0.788	23.77
SLD (Max)	81.8%	18.63	0.801	23.03
Safety Filter	71.2%	19.01	/	/

FID score is calculated between the set of generated images and the set of referenced images.

We present three key observations. Firstly, SafeGen achieves a CLIP score on par with the original SD, indicating its ability to preserve benign text-to-image preservation without degradation. In contrast, text-dependent methods with reasonable anti-sexually-explicit levels such as ESD, and SLD (Max/Strong) have lower benign CLIP scores (ranging from 23.03 to 23.97), averaging 0.83 lower than SafeGen, which suggests a potential compromise in content alignment. Secondly, SafeGen's LPIPS score and FID score are aligned with the original SD without decrease, which means SafeGen is capable of generating high-fidelity benign imagery. As a comparison, while text-dependent methods yield similar LPIPS scores, they show a higher average FID-25k gap of 3.0 than 20.31 of SafeGen, suggesting a potential negative impact on accurately reflecting textual descriptions in benign generation. Thirdly, the comparison of generated images between SafeGen and existing methods shown in Figure ?? suggests that SafeGen's superior performance in human evaluation. It well maintains the images' original style and overall layout of the original SD. While ESD obtains comparable performance in objective metrics like LPIPS and FID, it obviously affects the overall content and quality.

The reason is that the text-dependent methods erase or modify some NSFW concepts (*e.g.*, nudity, sexual), in the SD model. Such modifications often pertain to human-related content, which is integral to the image's context. As a result, altering these aspects can lead to a misalignment between the text and image, and also affect the model's overall fidelity, especially for human-related objects.

6.3 RQ3: Performance Combined with Baselines

Table 4 shows the results of the performance of SafeGen when combined with baselines. We evaluate the combination with text-dependent baselines, *i.e.*, ESD and different variants of SLD on both nudity mitigation and benign preservation. We skip the safety filter baseline since it has no impact on benign generation. We employ SneakyPrompt-N for testing nudity mitigation and COCO-25k for

Table 5: [RQ4-2] Total number of nudity parts in model-generated images under different random data selection.

Detected Nudity Parts ↓	Original SD	SafeGen Rand 1	SafeGen Rand 2	SafeGen Rand 3	SafeGen Rand 4	SafeGen Rand 5
COCO-Human♮	20	27	28	16	18	14
NSFW-56k	4533	27	29	28	27	32

(1) ♮: The COCO-Human set consists of 1,500 human-related model-generated images conditioned on varying benign prompts. NudeNet identifies the exposed body parts and aggregates their number.
(2) SafeGen Rand 1∼5 denotes 5 SD models that are governed by different random 100 image triplets selection.

testing benign preservation. In each dataset, we randomly select 200 prompts, and then generate three images per prompt using different random seeds. Our findings reveal that SafeGen, with only self-attention layers adjustment, alone outperforms all baselines in terms of nudity removal with average 21.7% NRR improvement, while retaining high-fidelity benign generation with comparable CLIP score. In addition, the integration with other text-dependent techniques demonstrate SafeGen significantly aids baselines in reducing sexually explicit content generation, realizing a remarkable 27.8% NRR enhancement.

From the perspective of nudity content mitigation, SafeGen + SLD (Max) achieves the highest NRR at 97.8% and the lowest CLIP score at 16.75, indicating its effectiveness in mitigating exposed body parts generation and deviating the resulting images from adversarial prompts. On the other hand, from the perspective of benign generation mitigation, SafeGen + SLD (Weak) has the lowest LPIPS score at 0.787 and the highest CLIP score at 24.33, which suggests it preserves the visual fidelity of benign images well.

This observation suggests a trade-off between unsafe generation mitigation and benign generation preservation. While SafeGen + SLD (Max) is most effective in nudity removal, it slightly compromises image fidelity as indicated by a higher LPIPS score. Conversely, SafeGen + SLD (Weak) preserves benign image fidelity better but does not perform as well in nudity removal as the former. Thus, the choice of method depends on the specific requirements of the task, *i.e.*, whether the priority is to maximize sexually explicit content removal or to preserve the fidelity of benign images.

6.4 RQ4: Random Image Triplet Selection

The default governance of SafeGen includes 100 randomly selected image triplets from the NSFW Dataset [7] and Human Detection Dataset [14]. We explore the impact of random data selection on model governance, particularly in terms of benign human-related false positives and sexually explicit mitigation. Namely, SafeGen's false moderation of benign human-related images and its efficacy in suppressing sexually explicit content. Table 5 presents 20 detected nudity parts of unmodified SD, which is attributed to inherent errors in NudeNet's detection. With different random data selections, SafeGen consistently reports similar false positive rates, ranging from 16 to 23. Our human evaluation also verify SafeGen's low false positive rates below 1.4% across five selections (see §7.5). In response to adversarial prompts, SafeGen effectively reduces the number of detected nudity parts, from 4533 to 27∼32. Across five selections, all NRR values surpass 99.2%, indicating that SafeGen achieves a reliable balance between mitigating sexually explicit

content and preserving benign images, even with varying random data selections.

7 HUMAN EVALUATION: USER STUDY

Given the difficulty of precisely defining "sexual explicitness", we conducted a large-scale user study, which was approved by our Institutional Review Board (IRB), to derive human-centric insights into various defense methods. This study comprehensively gathered real user feedback on the effectiveness of these methods in mitigating sexually explicit content while preserving the generation of benign content. Additionally, it complements the findings obtained from objective metrics.

Human Evaluation Setup. Authorized by the IRB, we recruited 82 adult participants aged 21 to 47, including 53 males and 29 females, to answer a five-part questionnaire. This survey aims to extensively compare SAFEGEN with 8 baselines, *i.e.*, Original SD, SD-V2.1, ESD, SLD (Max), SLD (Strong), SLD (Medium), SLD (Weak), Safety Filter. Moreover, we introduce a variant of Safety Filter and SAFEGEN, named "Nudity Detection Layer", which applies a dense mosaic overlay to detected nudity areas. Our research objectives are as follows:

- [Part 1] Quantify the fraction of images still considered as sexually explicit by participants despite employing different defense methods.
- [Part 2] Assess how different defense methods affect the alignment between generated images and their corresponding prompts under sexually explicit and benign conditions.
- [Part 3] Assess the impact of different defense methods on the quality of benign image generation.
- [Part 4] Quantify the false negative ratio of sexually explicit images generated using the nudity detection layer, safety filter, and SAFEGEN, respectively.
- [Part 5] Quantify the false positive ratio of benign images generated using the nudity detection layer, safety filter, and SAFEGEN, respectively.

7.1 Part 1: Sexually Explicit Fraction

Question Setup. We examine the efficacy of different defenses in mitigating severe sexual explicitness. Each SD model generates 30 images in response to 30 adversarial prompts, resulting in a total of $30 \times 9 = 270$ images. Participants were asked to tell how many images are sexually explicit based on their immediate perceptions. Subsequently, we calculated the "Sexually Explicit Fraction" by dividing the total number based on user answers. A lower fraction denotes better mitigation efficacy.

Result. As demonstrated in Figure 11, SAFEGEN remains the most effective approach in mitigating sexual explicitness, exhibiting a fraction as low as 0.08% with negligible user deviations. Moreover, the results exhibit a consistent trend with the objective metric experiments, where defenses with higher NRR can also yield lower percentages of sexual explicitness. Although text-based defenses like SLD (max) and ESD outperforming the image-based safety filter in terms of NRR, user feedback suggests that participants still perceive a considerable portion of images as sexually explicit. This finding suggests that these text-based mitigation may overlay
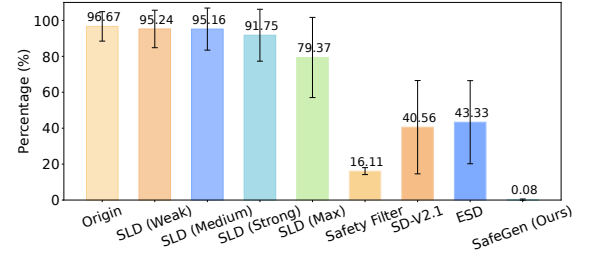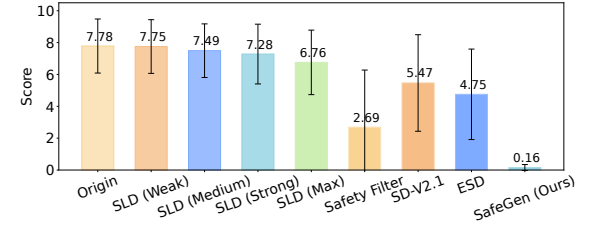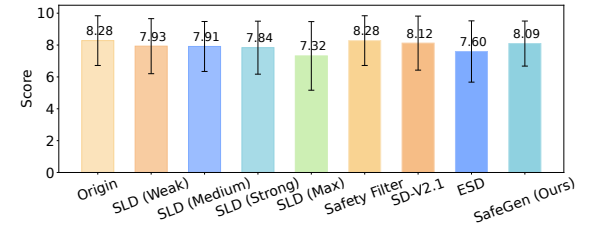

Figure 11: Sexually explicit fractions of the SD-generated images when employing different mitigation strategies.

(a) Sexually Explicit Text-to-Image Alignment (↓)

(b) Benign Text-to-Image Alignment (↑)
Figure 12: Human rated text-to-image alignment.

clothing on nudity areas, resulting in reduced naked parts, but it does not necessarily equate sexual explicitness with nudity.

7.2 Part 2: Text-to-Image Alignment

Question Setup. We examine the effectiveness of adopting CLIP scores to evaluate the alignment between textual prompts and corresponding images. This part consists of 5 adversarial and 5 benign questions, respectively. For each question, participants are asked to observe the given prompts and its generated images with different protection. Participants then rate, on a scale of 1~10, the faithfulness of generated images to provided prompts. (1 being entirely unrelated, 10 being perfectly matched). Since the image-based safety filter behaves identically to the original SD when confronted with benign images, we simplify this in the benign set. Thus, each participant shall assess $5 \times 9 = 45$ sexually explicit and $5 \times 8 = 40$ benign text-to-image pairs.

Results. Figure 12 shows that the ranking of defense strategies in human-perceived text-to-image alignment scores under both adversarial and benign prompts closely mirrors the objective CLIP scores detailed in Table 2. Notably, from a human perspective, ESD proves more effective than SD-V2.1 and SLD (Max) in disrupting malicious alignment, while SAFEGEN even surpasses SLD (Weak) in preserving alignment under benign prompts. As illustrated in
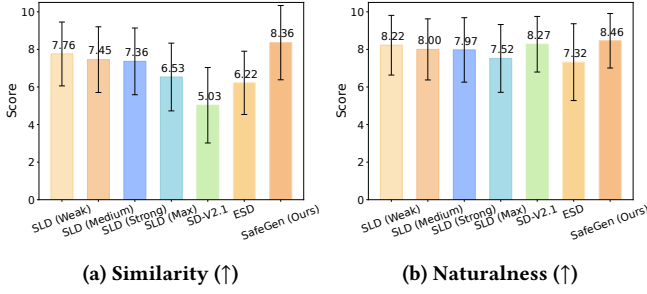
(a) Similarity (↑)

(b) Naturalness (↑)

Figure 13: Human rated similarity and naturalness of the benign generation when employing different mitigation strategies.

Figure 12 (a), both SafeGen and the safety filter exhibit substantial efficacy in suppressing the SD model's response to adversarial prompts, with user scores dropping as low as 0.16 and 2.69, indicating near-complete irrelevance. This stems from both methods yielding a "moderated" output upon detecting sexual explicitness, as depicted in Figure 5 and Figure 9. Nonetheless, due to the under-generalization issue intrinsic to the safety filter, explicit images can evade filtering, leading to scores ranging from 0 to almost 10, thereby causing considerable variance. Figure 12 (b) highlights the advantage of SafeGen, which focuses on removing explicit representations from the SD model internally while maintaining desirable text-to-image alignment by not altering the aligned cross-attention layer's response to text conditioning, ranking second only to the original SD and SD-V2.1. Please note that since the safety filter only blocks NSFW images, its performance in benign generation mirrors that of the original SD. We present results here to facilitate the comparison of different strategies across adversarial and benign conditions.

7.3 Part 3: Benign Image Quality

Question Setup. We employ the SD model with different defenses to produce benign images, comprising 6 categories: animals, food, human beings, landscapes, transport vehicles, home scenes. Participants are asked to rate the similarity score (1∼10) and quality score (1∼10) for each defense. In this study, "similarity" represents how similar the generated images are to those of the original SD. "Naturalness" denotes how realistic-looking the images are from a human perspective.

Results. Figure 13 (a) demonstrate that SafeGen leads by a margin over other defenses with an 8.36 similarity score. We attribute this to SafeGen's focus on eliminating explicit visual representations from the diffusion model while preserving the integrity of benign representations and the cross-attention layer's response to text prompts, akin to the original SD. In contrast, text-based mitigation inevitably compromises these factors, and SD-V2.1, trained on distinct data, consequently yields the lowest similarity. Figure 13 (b) shows that SafeGen also excels in producing realistic-looking benign content, with a high naturalness score of 8.46. Notably, SD-V2.1 performs better in this regard, achieving a score of 8.27, due to its improvement in high-fidelity generation with more real-life training data.

7.4 Part 4: False Negatives

Question Setup. We investigate the false negative rate, *i.e.*, the percentage of sexually explicit images where defenses fail to moderate or filter. A bit different from the experiment in §7.1, we introduce a new protection variant named "Nudity Detection Layer". Moreover, for each mitigation, participants are asked to respond a larger-scale testing involving 100 images generated by adversarial prompts. The nudity detection layer, based on the Anti-DeepNude tool [2] used in SafeGen's data preparation, forms a fair comparison. It overlays dense mosaic on the nudity areas to obstruct explicitness.

Results. Despite nudity detection layer recognizing and obstructing nudity, the average false negative rate remains high at 45.83%. We observe significant variance among users: some perceive an association with sexual explicitness despite the obfuscated images, while others consider the mosaic effective in reducing explicitness. The safety filter exhibits a high false negative rate of 22.35%. Notably, SafeGen maintains a low false negative rate at 0.07%, underscoring its effectiveness in mitigating sexually explicit content.

7.5 Part 5: False Positives

Question Setup. We also explore the false positive rate, *i.e.*, the percentage of benign images that defenses falsely moderate or filter benign generation. For each mitigation, we perform large-scale user testing involving 1,500 images generated in response to benign prompts under each protection. Participants are asked to tell how many images are falsely moderated or filtered.

Results. The nudity detection layer exhibits an unacceptable false positive rate. Although effectively covering mosaic on all nudity areas, it also overly applies mosaic to benign images devoid of any nudity. In this task1, the safety filter achieves remarkably low false positives. We attribute the high false negative rate to its usability trade-offs, sacrificing some safety against NSFW images. However, SafeGen well strikes the balance, with false positive rates below 1.40%.

8 DISCUSSION AND FUTURE WORK

Problem Definition ("Sexual Explicitness" or "Nudity"?) SafeGen is designed to suppress the generation of "sexually explicit" images in a text-agnostic manner. An exact definition of "sexual explicitness" is difficult due to various sociological factors. In existing works [9, 11, 15, 49], "nudity" is a commonly-used quantifiable metric to detect "sexual explicitness," and we conduct extensive experiments on the same setting for a fair comparison. However, we would like to clarify that employing NRR metrics does not imply that we regard "sexual explicitness" the same as "nudity." Specifically, we exclude the body parts such as "Feet" and "Armpits" predefined by NudeNet, because they are not normally considered as sexually explicit to most audience. In addition, we extend this metric by leveraging the CLIP Score that is widely used in prior works concerning the safety of T2I models [15, 44], and carrying out comprehensive user studies to report subjective results on sexual explicitness. The findings confirm the alignment between subjective human assessments and the objective metrics we employ, *i.e.*, NRR and CLIP score. Overall, sexual explicitness mitigation methods are

now evaluated indirectly via proxies like NRR and CLIP scores. A future direction is to investigate deeper into societal implications and cultural variances that affect the definition of "sexual explicitness" and design more suitable objective metrics that can serve as a complement for subjective user study.

False Positives & Over-Censorship. SafeGen presents low CLIP scores and high NRRs on sexually explicit mitigation at low false positives (falsely moderating the generation of benign human-related images) below 1.4% across different random data selection as detailed in §6.4. Admittedly, SafeGen is susceptible to over-censorship in some cases due to its capability of removing nudity-related visual representations from the model. This may result in unwanted moderation of non-explicit nudity, such as images of nude sculptures. Fortunately, our added experiment results show that we can adjust our benign set to include typical non-explicit images, such as "nude sculptures" and "man in beach shorts." This adjustment allows SafeGen to better discern between explicit and non-explicit content, further reducing the false positive rate and addressing the over-censorship issues. Moreover, we envision integrating text-based mitigation strategies to further reduce SafeGen 's false positives and relieve over-censorship issues. For example, text-based SLD [49] and ESD [15] visually conceal nudity by superimposing clothes, like brassiere, over exposed body parts, and SafeGen 's complete image moderation might be balanced with these text-based techniques. This combination of various strategies is our future direction. At the same time, we call for future works to investigate deeper into societal implications and cultural variances that affect the definition of "sexual explicitness", to establish a clear censorship standard.

Future Works. This work aims to shed light on model governance and promote responsible AI. We are dedicated to further contributing to the community in the following two aspects: (1) *Community Contribution.* We open-source our implementation [1] and call for awareness of model compliance. We plan to promote the integration of SafeGen into widely used generative model libraries, *e.g.*, Diffusers [55]. (2) *Broader Application.* We envision our vision-only regulation can be extended to other generative models, including text-to-video and image-to-image models, to prevent the explicit content generation in these applications.

9 ETHICAL CONSIDERATION

Responsible Handling of Explicit Content: SafeGen enables effective mitigation against the misuse of T2I models for generating sexually explicit content, which necessitates the handling of explicit images to regulate the self-attention layers of T2I models. To address potential discomfort and ethical concerns associated with this aspect of the research, we employ automated tools. Specifically, we utilize mosaic algorithms [2] and the BLIP2 model [32] for automated image processing. This approach ensures that our research team is not directly exposed to explicit imagery and eliminates the need for manual labeling, thereby aligning with ethical standards in handling sensitive content.

Mitigation of Potential Harms: The development of our comprehensive benchmark includes both adversarial and benign textual prompts. This benchmark is instrumental in assessing the efficacy

of various countermeasures against sexually explicit content generation by T2I models. It is important to note that the benchmark comprises solely textual prompts, which are inherently less offensive compared to explicit images. Nevertheless, in line with our commitment to ethical research practices, we have decided against publicly releasing this dataset. Our intention is to prevent any potential misuse or propagation of harmful content. Access to these datasets will be strictly regulated and will be provided only upon request for legitimate research purposes. Such requests will be subject to rigorous scrutiny, requiring institutional approval to ensure alignment with ethical research standards.

10 CONCLUSION

In this paper, we delve into the critical misuse of text-to-image (T2I) models in generating sexually explicit images. To address this risk, we introduce SafeGen, a novel framework that effectively eliminates latent representations of nudity within T2I models while preserving the models' capability to produce high-fidelity benign content, by regulating the vision-only self-attention layers. Safe-Gen severs the associations between explicit visual representations and conceptually sexual prompts. As a result, it outperforms eight baselines across four datasets and achieves optimal efficacy by complementing other techniques. These findings are confirmed by extensive objective metrics and human evaluation.

ACKNOWLEDGEMENT

REFERENCES

[1] https://github.com/LetterLiGo/SafeGen_CCS2024.
[2] Anti Deepnude. https://github.com/1093842024/anti-deepnude.
[3] Midjourney. https://www.midjourney.com.
[4] Stability AI. Stable Diffusion V2-1. https://huggingface.co/stabilityai/stable-diffusion-2-1.
[5] Artificial Intelligence & Machine Learning Lab at TU Darmstadt. Inappropriate Image Prompts (I2P). https://huggingface.co/datasets/AIML-TUDA/i2p.
[6] Artificial Intelligence & Machine Learning Lab at TU Darmstadt. Safe Stable Diffusion. https://huggingface.co/AIML-TUDA/stable-diffusion-safe.
[7] Evgeny Bazarov. NSFW Image Dataset. https://github.com/EBazarov/nsfw_data_source_urls.
[8] Charlotte Bird, Eddie L. Ungless, and Atoosa Kasirzadeh. Typology of Risks of Generative Text-to-image Models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2023, Montréal, QC, Canada, August 8-10, 2023*, pages 396–410, 2023.
[9] Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. SEGA: Instructing Text-to-image Models using Semantic Guidance. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*, 2023.
[10] Manuel Brack, Felix Friedrich, Patrick Schramowski, and Kristian Kersting. Mitigating Inappropriateness in Image Generation: Can there be Value in Reflecting the World's Ugliness? *arXiv*, abs/2305.18398, 2023.
[11] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, and William Chan. Circumventing Concept Erasure Methods For Text-to-image Generative Models. In *Proceedings of the 12th International Conference on Learning Representations, ICLR*, 2024.
[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, pages 4171–4186, 2019.

[13] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, 2021.

[14] Fares Elmenshawii. Human Detection Dataset. https://www.kaggle.com/datasets/fareselmenshawii/human-dataset.

[15] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing Concepts from Diffusion Models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023*, pages 2426–2436, 2023.

[16] Hongcheng Gao, Hao Zhang, Yinpeng Dong, and Zhijie Deng. Evaluating the Robustness of Text-to-image Diffusion Models against Real-world Attacks. *arXiv*, abs/2306.13103, 2023.

[17] Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. DiffuSeq: Sequence to Sequence Text Generation with Diffusion Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023.

[18] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014.

[19] Yuchao Gu, Lijuan Wang, Ziqin Wang, Yun Liu, Ming-Ming Cheng, and Shao-Ping Lu. Pyramid Constrained Self-attention Network for Fast Video Salient Object Detection. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020*, pages 10869–10876, 2020.

[20] Longteng Guo, Jing Liu, Xinxin Zhu, Peng Yao, Shichen Lu, and Hanqing Lu. Normalized and Geometry-aware Self-attention Network for Image Captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, pages 10324–10333, 2020.

[21] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey A. Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen Video: High Definition Video Generation with Diffusion Models. *arXiv*, abs/2210.02303, 2022.

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020.

[23] Jonathan Ho and Tim Salimans. Classifier-free Diffusion Guidance. *arXiv*, abs/2207.12598, 2022.

[24] Tatum Hunter. AI Porn Is Easy to Make Now. For Women, That's a Nightmare. https://www.washingtonpost.com/technology/2023/02/13/ai-porn-deepfakes-women-consent.

[25] Wiliam Hunter. Paedophiles Are Using AI to Create Sexual Images of Celebrities as CHILDREN, Report Finds. https://www.dailymail.co.uk/sciencetech/article-12669791/Paedophiles-using-AI-create-sexual-images-celebrities-\CHILDREN-report-finds.html.

[26] Hugging Face Inc. Models. https://huggingface.co/models.

[27] OpenAI Inc. Dall-E 2. https://openai.com/dall-e-2.

[28] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising Diffusion Restoration Models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, 2022.

[29] Alex Kim. NSFW Image Dataset. https://github.com/alex000kim/nsfw_data_scraper.

[30] Diederik P. Kingma and Max Welling. Auto-encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[31] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. DiffWave: A Versatile Diffusion Model for Audio Synthesis. In *9th International Conference on Learning Representations, ICLR 2021*, 2021.

[32] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: Bootstrapping Language-image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning, ICML 2023*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742, 2023.

[33] Michelle Li. NSFW Text Classifier on Hugging Face. https://huggingface.co/michellejieli/NSFW_text_classifier.

[34] Machine Vision & Learning Group LMU. Safety Checker. https://huggingface.co/CompVis/stable-diffusion-safety-checker.

[35] Machine Vision & Learning Group LMU. Stable Diffusion V1-4. https://huggingface.co/CompVis/stable-diffusion-v1-4.

[36] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[37] Madison McQueen. AI Porn Is Here and It's Dangerous. https://exoduscry.com/articles/ai-porn.

[38] Dan Milmo. AI-created Child Sexual Abuse Images 'Threaten to Overwhelm Internet'. https://www.theguardian.com/technology/2023/oct/25/ai-created-child-sexual-abuse-images-threaten-overwhelm-internet.

[39] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards Photorealistic Image Generation and Editing with Text-guided Diffusion Models. In *International Conference on Machine Learning, ICML 2022*, volume 162 of *Proceedings of Machine Learning Research*, pages 16784–16804, 2022.

[40] notAI tech. NudeNet: Lightweight Nudity Detection. https://github.com/notAI-tech/NudeNet.

[41] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On Aliased Resizing and Surprising Subtleties in GAN Evaluation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, pages 11400–11410, 2022.

[42] Olivier Petit, Nicolas Thome, Clément Rambour, Loic Themyr, Toby Collins, and Luc Soler. U-Net Transformer: Self and Cross Attention for Medical Image Segmentation. In *Machine Learning in Medical Imaging - 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021*, volume 12966 of *Lecture Notes in Computer Science*, pages 267–276, 2021.

[43] Pharmapsychotic. Clip Interrogator. https://github.com/pharmapsychotic/clip-interrogator/tree/main.

[44] Yiting Qu, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. Unsafe Diffusion: On the Generation of Unsafe Images and Hateful Memes From Text-To-image Models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, CCS 2023*, pages 3403–3417, 2023.

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763, 2021.

[46] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-teaming the Stable Diffusion Safety Filter. *arXiv*, abs/2210.04610, 2022.

[47] Reddit. Tutorial: How to Remove the Safety Filter in 5 seconds. Website, 2022. https://www.reddit.com/r/StableDiffusion/comments/wv2nw0/tutorial_how_to_remove_the_safety_filter_in_5/.

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, pages 10674–10685, 2022.

[49] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 22522–22531, 2023.

[50] Christoph Schuhmann. LAION's NSFW Detector. https://github.com/LAION-AI/CLIP-based-NSFW-Detector.

[51] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: an Open Large-scale Dataset for Training Next Generation Image-text Models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, 2022.

[52] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. In *9th International Conference on Learning Representations, ICLR 2021*, 2021.

[53] Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia-You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Ring-A-Bell! How Reliable are Concept Removal Methods for Diffusion Models? *arXiv*, abs/2310.10012, 2023.

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 5998–6008, 2017.

[55] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art Diffusion Models. https://github.com/huggingface/diffusers, 2022.

[56] Yixin Wu, Ning Yu, Michael Backes, Yun Shen, and Yang Zhang. On the Proactive Generation of Unsafe Images From Text-To-image Models Using Benign Prompts. *arXiv*, abs/2310.16613, 2023.

[57] Yuchen Yang, Bo Hui, Haolin Yuan, Neil Gong, and Yinzhi Cao. SneakyPrompt: Jailbreaking Text-to-image Generative Models. In *IEEE Symposium on Security and Privacy*, 2024.

[58] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, pages 586–595, 2018.