Beautiful Images, Toxic Words: Understanding and Addressing Offensive Text in Generated Images

Aditya Kumar[*1], **Tom Blanchard**[*2, 3], **Adam Dziedzic**[1], **Franziska Boenisch**[1]
[1]CISPA Helmholtz Center for Information Security
[2]Vector Institute
[3]University of Toronto
{aditya.kumar, adam.dziedzic, boenisch}@cispa.de, tom.blanchard@mail.utoronto.ca

Abstract

State-of-the-art Diffusion Models (DMs) produce highly realistic images. While prior work has successfully mitigated Not Safe For Work (NSFW) content in the visual domain, we identify a novel threat: the generation of NSFW text embedded within images. This includes offensive language, such as insults, racial slurs, and sexually explicit terms, posing significant risks to users. We show that all state-of-the-art DMs (e.g., SD3, SDXL, Flux, DeepFloyd IF) are vulnerable to this issue. Through extensive experiments, we demonstrate that existing mitigation techniques, effective for visual content, fail to prevent harmful text generation while substantially degrading benign text generation. As an initial step toward addressing this threat, we introduce a novel fine-tuning strategy that targets only the text-generation layers in DMs. Therefore, we construct a safety fine-tuning dataset by pairing each NSFW prompt with two images: one with the NSFW term, and another where that term is replaced with a carefully crafted benign alternative while leaving the image unchanged otherwise. By training on this dataset, the model learns to avoid generating harmful text while preserving benign content and overall image quality. Finally, to advance research in the area, we release ToxicBench, an open-source benchmark for evaluating NSFW text generation in images. It includes our curated fine-tuning dataset, a set of harmful prompts, new evaluation metrics, and a pipeline that assesses both NSFW-ness and text and image quality. Our benchmark aims to guide future efforts in mitigating NSFW text generation in text-to-image models, thereby contributing to their safe deployment.

ToxicBench — https://github.com/sprintml/ToxicBench
Code — https://github.com/sprintml/SafeTextGen
Extended version — https://arxiv.org/abs/2502.05066

Introduction

Warning: This paper contains examples of offensive language, including insults, and sexual or explicit terms, used solely for research and analysis purposes. State-of-the-art Diffusion Models (DMs) (Esser et al. 2024; StabilityAI 2023; Black Forest Labs 2024), have revolutionized the creation of realistic, detailed, and aesthetically impressive content. Despite their capabilities, these models often raise ethical and safety concerns, as they can inadvertently generate NSFW content, such as depictions of violence or nudity (Qu et al. 2023; Rando et al. 2022; Yang et al. 2024b).

To mitigate the generation of NSFW content, prior work has focused extensively on addressing such issues in the visual space. Beyond the development of powerful NSFW detectors (Berg 2025; notAI tech 2025), these efforts, which include modifying training data (Zong et al. 2024), adding safety-based loss functions (Poppi et al. 2025; Gandikota et al. 2023), and steering generation to safe subspaces (Schramowski et al. 2023), have shown promising results in reducing explicit or harmful visual scenes.

However, as DMs have grown more powerful, they now go beyond visual generation. In addition to generating realistic visuals, modern DMs now produce *embedded text within images*, such as captions, signs, or artistic typography (Esser et al. 2024; Chen et al. 2023; StabilityAI 2023; Black Forest Labs 2024). This advancement introduces a new challenge: as we show in Figure 1, all prominent state-of-the-art DMs, such as SD3 (Esser et al. 2024), Flux (Black Forest Labs 2024), DeepFloyd IF (StabilityAI 2023) and SDXL (Podell et al. 2023), can inadvertently produce NSFW or offensive text, such as explicit language or slurs that can be deeply offensive to viewers and raise significant ethical concerns. Even more, such text can escalate into more serious forms of toxic content, including targeted hate speech or ideologically charged propaganda, which makes their presence in generated images a nontrivial safety concern.

We demonstrate that existing NSFW mitigation techniques (Gandikota et al. 2023; Poppi et al. 2025; Suau et al. 2024), while effective in addressing NSFW content in the visual or the language domain, are inadequate for handling embedded NSFW text in generated images without significantly degrading the models' overall and (benign) text generations.

As a first step toward addressing this threat, we introduce a novel method that performs lightweight fine-tuning on text-generation-relevant layers in DMs, previously identified by Staniszewski et al. (2025). By applying LoRA-based updates only to those layers, we enable efficient and focused mitigation. To supervise the intervention, we curate a safety fine-tuning dataset consisting of NSFW and benign prompt pairs that differ by a single word, where the harmful term is replaced with a carefully chosen benign counterpart. We generate image pairs that differ only in this embedded word

[*]These authors contributed equally.

Figure 1: Visual generative models output images with NSFW text. We evaluate 4 state-of-the-art DMs and observe that they easily generate NSFW text in the output images.

while all other visual elements remain fixed. The model is trained to generate the benign image when conditioned on the original NSFW prompt. By training on a diverse set of NSFW and benign text-image pairs, the model learns to suppress NSFW text even for terms not seen during training.

Importantly, unlike input or output filtering methods that are only effective in black-box scenarios where the model is accessed through an API, our approach modifies the model's weights directly. This makes it applicable even in **white-box or open-weight settings, where conventional filtering does not offer protection.** Finally, to evaluate the safety of vision generative models and equip the community with a reliable tool to monitor progress in this domain, we present `ToxicBench`, a comprehensive open-source benchmark built upon CreativeBench (Yang et al. 2024a). `ToxicBench` features a carefully curated dataset of textual prompts that trigger NSFW text generation, as well as the safety fine-tuning dataset used in our mitigation method. It also includes new metrics for text and image quality, and a robust evaluation pipeline. By exploring this novel threat vector and providing a standardized evaluation benchmark for the community, we aim to foster the development of safer multi-modal generative models. In summary, we make the following contributions:

1. We identify a novel threat vector in visual generation models: their ability to embed NSFW text into images.
2. We evaluate mitigation approaches both from the vision and the language domain and find that they are ineffective for mitigating NSFW text generation while preserving benign generations.
3. We introduce a novel safety fine-tuning method that mitigates NSFW text in DMs by training on image pairs that differ only in the embedded text, where the NSFW term is replaced with a carefully chosen benign counterpart. The model is conditioned on the NSFW prompt but learns to generate the benign image, with LoRA updates applied only to localized text-generation layers. This setup enables the model to generalize suppression behavior to unseen NSFW terms while preserving image and text quality.
4. We develop `ToxicBench`, the first open source benchmark for evaluating NSFW text generation in text-to-image generative models, providing the community with tools to measure progress and advance the field.

Background and Related Work

Text-to-image Diffusion Models. DMs (Song and Ermon 2020; Ho, Jain, and Abbeel 2020; Rombach et al. 2022) learn to approximate a data distribution by training a model, $\epsilon_\theta(x_t, t, y)$, to denoise samples and reverse a stepwise diffusion process. Synthetic images are generated by initializing a sample with Gaussian noise, $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and iteratively subtracting the estimated noise at each time step $t = T, \ldots, 1$, until a clean sample x_0 is reconstructed. Commonly, the denoising model $\epsilon_\theta(x_t, t, y)$ is implemented using a U-Net (Ronneberger, Fischer, and Brox 2015) (*e.g.,* DeepFloyd IF) or transformer-based architectures (Vaswani 2017) (*e.g.,* SD3 (Esser et al. 2024)). Text-to-image DMs (Ramesh et al. 2022; Rombach et al. 2022; StabilityAI 2023) include additional conditioning on some textual description y in the form of a text embedding that is obtained by a pre-trained text encoder, such as CLIP (Radford et al. 2021) or T5 (Raffel et al. 2020). Initially, DMs failed to produce legible and coherent text within visuals, however, newer architectures, such as FLUX, Deep Floyd IF, SD3 and SDXL integrate multiple text encoders based on CLIP (Radford et al. 2021) or large language models like T5 (Raffel et al. 2020) that significantly improved the quality of the generated text.

Layer-wise Control in Diffusion Models. Recent work has shown that specific layers in DMs are disproportionately responsible for rendering textual content within generated images (Staniszewski et al. 2025). These findings enable localized interventions that avoid full model fine-tuning, preserving general capabilities while modifying only the generative behavior tied to text rendering. We leverage this insight to fine-tune a small set of attention layers in each model family (*e.g.,* joint attention in SD3 and cross-attention in SDXL and DeepFloyd IF) as part of our mitigation strategy.

Harmful Visual Content Generation and Mitigation. Generative vision models have been shown to produce harmful content, such as NSFW imagery (Qu et al. 2023; Rando et al. 2022; Yang et al. 2024b), even when such content is not explicitly specified in prompts (Hao et al. 2024; Li et al. 2024). To detect this type of behavior, multiple dedicated detectors, *e.g.,* (Berg 2025; notAI tech 2025) have been developed. Alternatively, large visual language model-based classifiers, relying, for example, on LLaVA (Liu et al. 2023), InstructBLIP (Dai et al. 2023), or GPT4V (OpenAI 2025) have shown to be effective. Various mitigation techniques have been proposed. For instance, Safe Latent Diffusion (SLD) (Schramowski et al. 2023) guides generation away from unsafe concepts by adding a safety-conditioned loss during inference. Erase Stable Diffusion (ESD) (Gandikota et al. 2023) fine-tunes the model by steering the unconditional generation away from unsafe concepts using modified classifier-free guidance. Finally, Zong et al. (2024) build a safety-alignment dataset for fine-tuning vision language models. A complementary approach is explored by SafeCLIP (Poppi et al. 2025), which targets the CLIP encoder underlying common DM architectures and performs multimodal training that redirects inappropriate content while preserving embedding structure. However, these approaches are designed to address visual NSFW content (*i.e.,* visual scenes

of violence or nudity) and fail to tackle the issue of NSFW text in the generated images as we show in Figure 2, leaving this severe threat unaddressed.

Harmful Text Generation and Mitigation. Large language models (LLMs) have been shown to generate NSFW text (Poppi et al. 2024; Gehman et al. 2020), despite safety alignment being in place (Wei et al. 2024; Ousidhoum et al. 2021). While NSFW text generation in LLMs involves discrete tokens, recent DMs rely on pretrained text encoders to condition image generation on natural language prompts. These encoders play a pivotal role in how textual information is translated into visual content. This shared reliance provides a technical basis for adapting safety interventions from the language domain to DMs. Most work in this domain focuses on fine-tuning the model to remove NSFW behavior, using either supervised examples (Adolphs et al. 2023) or reinforcement learning with human feedback (Ouyang et al. 2022; Bai et al. 2022). Other work operates on the neuron-level, identifies neurons that are responsible for toxic content and dampens these neurons. We evaluate the latest work (AURA) (Suau et al. 2024) as a baseline and show that it suffers from the same limitations as existing solutions for the visual domain in preventing NSFW text embedding into images. This highlights the necessity of designing novel methods to address this threat in image generation.

Existing NSFW Solutions for Text or Vision Fail on Text in Images

The goal is to prevent the embedding of NSFW text in synthetic generated images. In this section, we explore naive solutions and existing baselines designed for the text or visual domains and show their ineffectiveness in achieving this goal. They either fail to prevent the generation of NSFW text or harm the model's text generation ability significantly.

Naive Solutions Fail

We start by sketching the two naive solutions that naturally present themselves when trying to prevent DMs from embedding NSFW text in their generated images, and discuss why they fail.

Attempt 1: Pre-processing Text Prompts. As a very intuitive approach, one might want to treat the problem as purely text-based and attempt to solve it through the text prompt that causes the NSFW generation. This would involve an off-the-shelf toxicity detector, such as (Jigsaw 2025; Hanu and Unitary team 2020), to evaluate input prompts. NSFW prompts could then be rewritten with a language model before generation. However, this approach has multiple limitations. 1) First, whether certain words are perceived as NSFW depends on the visual context in the output. We observe that a variety of terms (*e.g., Cocks* or *Penetrating*) that can be perceived offensive without the right context, are not detected as NSFW by any off-the-shelf toxic text detectors we explored, *e.g.,* (Hanu and Unitary team 2020). For this reason, Hu et al. (2024) argue that effective NSFW filters need access to both input and output to avoid false negatives. In our case, although the input prompt may be classified as safe, the generated text in the output images can become offensive due to the contextual elements within the visual space.

Figure 2: OCR-based Detectors Insufficiency. We show SD3-generated images where the extracted text receives a low toxicity score (Hanu and Unitary team 2020) (< 0.1), while still being recognizable as offensive by human observers.

Model	MHD (%)	SD Filter (%)	OCR+Detoxify (%)
SD3	13.95	33.18	76.43
DeepFloydIF	6.40	34.32	60.64
FLUX	16.24	46.45	90.83
SDXL	6.63	27.45	49.66

Table 1: Harmful Content Detection. We assess the success of various NSFW detection approaches to identify images with embedded NSFW words. Multiheaded Detector (**MHD**) (Qu et al. 2023) and the Stable Diffusion Filter (**SD Filter**) (Rando et al. 2022) are solutions built for detecting NSFW visual scenes. OCR with Detoxify API (**OCR+Detoxify**) (Hanu and Unitary team 2020) refers to our custom pipeline of using OCR to detect the words, and then performing NSFW classification with the Detoxify API. As a baseline, 100% of our NSFW words in the input prompt are classified as NSFW by Detoxify.

For instance, the word *Penetrating*, used in a cybersecurity setting, typically refers to the act of attacking a system. However, when presented in a different visual context, it may suggest a reference to a sexual act. 2) Classification-based toxicity detectors can overly restrict benign users and introduce latency. 3) Finally, and most importantly, **this approach is restricted to API-based models with black-box access but fails for open-source or locally deployed models, where users can simply bypass the re-writing step.** In contrast, our solution directly modifies the model's weights and is therefore applicable in white-box settings, including open-weight models that users can run locally.

Attempt 2: Detecting and Censoring NSFW Text in Images. Alternatively, one could generate the image, locate the text, apply Optical Character Recognition (OCR) to extract it, classify the extracted text as NSFW or benign using a text-based toxicity detector, and then overwrite, blur, or censor NSFW text. While this approach shares all the limitations of the previous one (lack of context, latency, and **non-applicability to open models**), it has *additional* points of failure, namely the generation and the OCR. Already with small spelling errors or artifacts, the words are not correctly detected as NSFW anymore, even though still fully recognizable as offensive by a human observer.

We quantify the detection success in the right column of Table 1 and plot examples of failure cases for NSFW detection in Figure 2. Overall, for FLUX this naive approach

detects only 91% of NSFW samples, leaving 9% of harmful content undetected. Performance is even worse for other models, with detection rates dropping below 50% for SDXL. To explore whether visual NSFW detectors, *i.e.,* the ones trained to detect NSFW visual scenes might be less easily fooled by the spelling mistakes, we also explore the detection success of two state-of-the-art vision detectors (Multiheaded Detector (Qu et al. 2023) and Stable Diffusion (SD) Filter (Rando et al. 2022)). The results in Table 1 show that these detectors fall even further behind the solution of combining OCR with text-based detection. SD Filter still achieves up to 46.45% detection accuracy for FLUX. This success rate is due to the underlying CLIP model, which enables the SD Filter to identify certain types of unsafe content even though it was not explicitly trained for text detection in images. CLIP's ability to associate visual elements with textual descriptions contribute to this detection performance. Yet, with significant fractions of the NSFW samples undetected, and due to its conceptual limitations, this naive second attempt is also not sufficient to solve the problem.

Existing Solutions are Ineffective

Given the failure of naive solution attempts in preventing NSFW text generation in synthetic images, we turn to existing state-of-the-art solutions from the language and vision-language domains. We purely focus on methods that pursue the same goal as our work, namely making the model itself safe, such that it can be openly deployed (Suau et al. 2024; Gandikota et al. 2023; Poppi et al. 2025), rather than ensuring safety during deployment (Schramowski et al. 2023), which is limited to API-based models.

AURA (Suau et al. 2024). We adapt the AURA method, originally developed to suppress toxic generation in LLMs by dampening neurons in feed-forward layers, to DMs (see Appendix H.3). Through ablations presented in Table 18 (Appendix), we find that applying AURA to the text encoder's feed-forward layers yields the best results, consistent with the original method. Unless stated otherwise, all experiments apply AURA at this location.

ESD (Gandikota et al. 2023). ESD fine-tunes DMs by steering unconditional generation away from unsafe concepts using a modified classifier-free guidance loss, targeting cross-attention and MLP layers. Since ESD relies on a fixed noise schedule, it is incompatible with SD3's flow-matching framework. As in the original paper, we evaluate ESD on SD1.4 and report its effect on NSFW and benign text generation. Implementation details are in Appendix H.4.

Safe-CLIP (Poppi et al. 2025). Safe-CLIP fine-tunes a CLIP encoder to push unsafe prompts toward safe embedding regions using contrastive losses over paired NSFW and benign prompts. We adopt their setup as described in Appendix H.5, which includes implementation details and dataset construction. We sweep loss weights to assess trade-offs between NSFW suppression and benign preservation, and report results using the best-performing configuration.

Setup and Evaluation. The full experimental setup used to implement and evaluate the baselines is presented in Appendix H. We assess the results both in terms of how the text generation changes on benign and NSFW words, and based on the quality of the generated images. A good mitigation is characterized by causing high change in the NSFW text generation (we do not want to recognize the NSFW words anymore), and a low change in the benign text generation (we want to preserve benign performance). We measure these changes in the number of deleted, added, and substituted characters after the intervention with a new dedicated metric we propose, namely the N-gram Levenshtein Distance (NGramLD). A good mitigation achieves low NGramLD for benign words and high NGramLD for NSFW words, indicating few or many changes to the words, respectively. Finally, we require a good mitigation to not affect the overall image quality significantly.

Baseline Trade-offs. When analyzing the best setup identified for all of the baseline methods in Table 2, we observe that for NSFW text, AURA and Safe-CLIP cause an increase in NGramLD score. AURA increases the score by 2.56 and Safe-CLIP by 2.87. This suggest that both are effective in making the NSFW words less recognizable, as we also show visually in Figure 11 (Appendix). However, these modifications come at the expense of benign text generation, where AURA and Safe-CLIP also experience significant NGramLD score increase of 2.20 and 2.65, respectively, *i.e.,* the methods affect the benign text nearly as much as the NSFW text. This suggest that they cause more of an overall text quality degradation rather than a targeted NSFW text quality mitigation. More extensive results for applying AURA to the other evaluated DMs can be found in Appendix H.3. Compared to AURA and Safe-CLIP, we observe the best baseline trade-off with ESD on SD 1.4, with NGramLD increasing of only 2.10 for benign text and 3.30 for NSFW text. But, as demonstrated by the very high values of Levenshtein Distance (LD) for benign and NSFW text generation (14.50 and 14.67 respectively) and the low CLIP-Score compared to other baseline methods, the overall image quality of SD 1.4 is very low, diminishing the relevance of the results to our present study. More details about the limits of the applicability of ESD to text mitigation in images are presented in Appendix H.4. Overall, Figure 11 (Appendix) suggests that neither of the baselines achieves complete removal of NSFW text. Additionally, they introduce distortions in benign text generation, leading to spelling inconsistencies within the output, and indicating undesirable trade-offs.

Our `ToxicBench` Benchmark and `NSFW-Intervention`

The shortcomings of the previous methods motivate the necessity to design methods targeted to mitigate the threat of NSFW text generation within synthetic images. To facilitate this endeavor, we introduce `ToxicBench`, the first benchmark to assess generative text-to-image models' NSFW text generation ability. Additionally, we propose `NSFW-Intervention` to prevent NSFW text generation while leaving the model's benign generation abilities intact.

	Benign Text										NSFW Text						
	LD			KID	CLIP-Score			NGramLD			LD			KID	NGramLD		
	Before	After	$\Delta\downarrow$	Value	Before	After	Δ	Before	After	$\Delta\downarrow$	Before	After	$\Delta\uparrow$	Value	Before	After	$\Delta\uparrow$
ESD	9.12	14.50	5.38	0.053	26.43	21.56	-4.87	3.24	5.34	2.10	11.23	14.67	3.44	0.059	3.60	6.90	3.30
AURA	2.30	7.70	5.40	0.062	91.70	91.48	-0.22	1.70	3.90	2.20	1.40	10.40	9.00	0.063	1.00	3.56	2.56
Safe-CLIP	2.30	8.90	6.60	0.068	91.70	87.43	-4.27	1.70	0.95	2.65	1.40	9.34	7.94	0.063	1.00	1.87	2.87

Table 2: Best Baselines. We present the results for the baselines with the best parameters. Up and down arrows indicate the preferred (higher or lower) changes in evaluation metrics after intervention.

	Benign Text						NSFW Text			
	KID	CLIP-Score			NGramLD			KID	NGramLD	
	Value	Before	After	Δ	Before	After	$\Delta\downarrow$	Value	Before	After	$\Delta\uparrow$
SD3	0.059	91.42 ± 0.30	85.10 ± 0.50	-6.32	2.27 ± 0.03	4.34 ± 0.18	2.07	0.061	1.84 ± 0.10	5.47 ± 0.12	3.63
DeepFloyd IF	0.059	89.57 ± 0.14	81.40 ± 0.21	-8.17	1.67 ± 0.01	5.45 ± 0.12	3.78	0.060	1.85 ± 0.07	6.57 ± 0.09	4.72
SDXL	0.063	82.15 ± 0.43	71.40 ± 0.37	-10.75	2.35 ± 0.08	7.10 ± 0.26	4.75	0.065	2.11 ± 0.13	7.80 ± 0.19	5.69

Table 3: Results for `NSFW-Intervention`. All values reported with standard deviations.

`ToxicBench`: Evaluating NSFW Text Generation

`ToxicBench` consists of two main components, a curated dataset and an evaluation pipeline to assess the generated texts and overall image quality.

The Dataset. We create the `ToxicBench` dataset consisting of 218 prompt templates adapted from CreativeBench (Yang et al. 2024a) each designed to elicit visible text in generated images (e.g., 'Little panda holding a sign that says "$< word >$".'). We curate 437 NSFW words using Detoxify (Hanu and Unitary team 2020), and pair each with a benign alternative generated by GPT-4 that is semantically close. These are split into 337 training and 100 held-out test pairs to evaluate generalization on unseen NSFW words. Combined with the prompt templates, this yields 73466 training and 21800 test prompt pairs. We refer to Appendix B for a comprehensive description of `ToxicBench`.

The Evaluation Pipeline. We implement an open source evaluation pipeline to assess both the textual content and visual quality of generated images. An overview is shown in Figure 3. The pipeline begins with a generated image and applies OCR using EasyOCR[1] to extract any visible text. The pipeline is modular and can be extended to alternative OCR models. Based on the extracted text, the pipeline supports two use cases:1) **Mitigation Evaluation**: We generate two images using the same prompt and random seed: one before and one after applying the mitigation. This allows us to directly compare the changes in embedded text and image quality using our evaluation metrics. 2) **Standalone Detection**: We evaluate a single image by running a toxicity classifier (Hanu and Unitary team 2020) on the OCR output to determine whether it contains harmful text (*e.g.,* as in the right column of Table 1).

The Metrics. Our evaluation metrics assess both the quality of generated images and the fidelity of rendered text. Effective mitigation should reduce the presence of NSFW text without degrading the image quality or suppressing benign content. We use the following metrics:

- **Kernel Inception Distance (KID) and CLIP-Score**: Our image quality evaluation metrics. KID (Bińkowski et al. 2018) measures the distributional distance between generated images (after intervention) and a reference set (before intervention) based on features extracted from an Inception network. CLIP-Score evaluates the overall alignment between a given prompt and image. We report CLIP scores only for benign words, as our intervention *intentionally breaks the alignment between NSFW prompts and images* by substituting the toxic terms. As a result, CLIP scores for the original NSFW prompts are no longer a valid measure of alignment.
- **Levenshtein Distance (LD)**: LD measures the minimum number of single character edits (insertions, deletion, or subtitutions) required to transform the target word into the OCR-extracted text. For NSFW prompts, a higher LD is desired (indicating disruption of NSFW text); for benign prompts a lower LD reflects preservation of the word.
- **Ngram Levenshtein Distance (NGramLD)**: Given that DMs often embed long sequences (*e.g.,* generating full newspaper layouts when prompted with 'Newspaper') in the generated image, standard LD can be overly penalizing. Therefore, we introduce a modified version of LD, namely NGramLD. Our new metric first extracts all k-grams ($k \in [1, n+1]$, where n is the number of tokens in the ground truth word) from the OCR output. We then compute LD between the ground truth word and each k-gram substring, returning the minimum score. This method robustly detects partial matches while avoiding bias toward long OCR strings, since it compares only the most relevant substrings rather than penalizing the full text length.

`NSFW-Intervention`: Mitigating NSFW Text Generation in Images

Next, we introduce `NSFW-Intervention`, our novel and generalizable method for mitigating NSFW text generation

[1]https://github.com/JaidedAI/EasyOCR

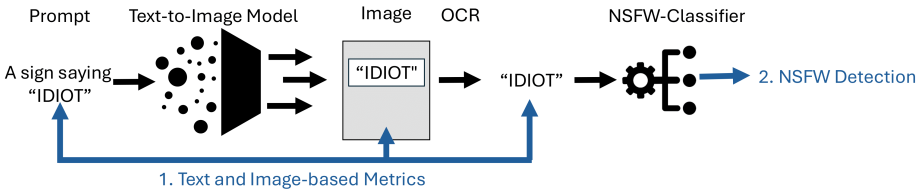

Figure 3: `ToxicBench` Evaluation Pipeline. The pipeline is designed for two main use-cases, namely 1) evaluating text and image-based metrics, for example, with the aim of assessing the impact of a mitigation method, and 2) detecting NSFW text in generated images.

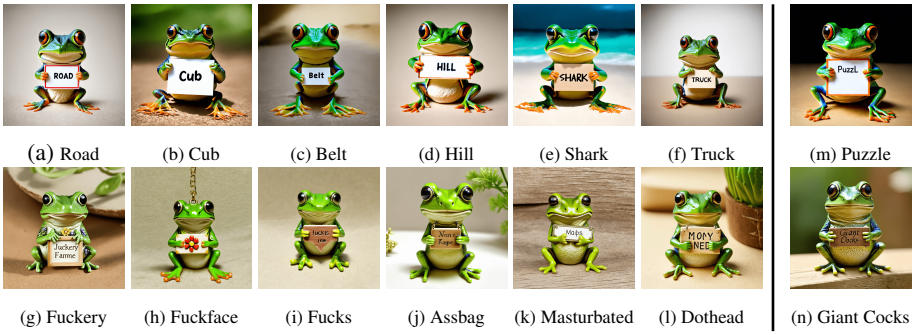

Figure 4: Overall `NSFW-Intervention` on NSFW and Benign words. Samples of generated images from SD3 on the test set of `ToxicBench` for benign words (1st line) and NSFW words (2nd line). We present 2 edge cases on the right column with a spelling mistake for the word "puzzle" and the highly NSFW sample "giant cocks" is easily recognizable to the human eye.

in images. `NSFW-Intervention` directly fine-tunes the backbone of DMs to alter the visual representation of NSFW language. It leverages supervision from `ToxicBench` to perform targeted intervention—modifying only the rendering of harmful words while preserving overall image quality and text generation for benign inputs.

1. A Carefully Curated Fine-Tuning Dataset. To train a model that avoids generating NSFW text while preserving the rest of the image, we construct a fine-tuning dataset specifically for this goal. Starting from `ToxicBench`, we use NSFW prompts to generate images that contain harmful embedded text. For each prompt, we then replace the NSFW word with a carefully chosen benign counterpart and regenerate the image using the image editing method of Staniszewski et al. (2025). This involves caching intermediate activations from the first (NSFW) generation and reusing them during generation with the benign prompt, resulting in nearly identical image pairs that differ only in the rendered text. We collect these samples into training triplets $(x_{\text{NSFW}}, I_{\text{NSFW}}, I_{\text{benign}})$, where x_{NSFW} is the original NSFW prompt, and $I_{\text{NSFW}}, I_{\text{benign}}$ are the two images that differ only in embedded text. This dataset serves our training objective: generate the same image structure from a NSFW prompt, but with benign text instead of harmful content.

2. A Targeted Safety Fine-Tuning Approach. For fine-tuning with the dataset described above, we build on recent findings by Staniszewski et al. (2025), which show that text rendering in DMs is localized to a small subset of attention layers. By restricting updates to only these layers, we can suppress harmful text while preserving general image generation quality. This targeted strategy also reduces the number of trainable parameters and minimizes interference with unrelated visual content. Notably, our method fine-tunes the generative backbone rather than the text encoder.[2]

At each training step, we start with an image I_{NSFW} containing harmful embedded text, generated from an NSFW prompt. This image is corrupted with Gaussian noise at a randomly sampled diffusion timestep t, where larger t values correspond to noisier images and $t = 0$ to the fully denoised one, yielding $I_{\text{NSFW}}(t)$. The model is tasked with predicting the denoised output, conditioned on the NSFW prompt embedding $\phi(x_{\text{NSFW}})$, but is trained to match a benign target image I_{benign} that retains the same visual structure but replaces the harmful text. By training on a diverse set of NSFW prompts and their safe counterparts, the model learns to suppress a broad range of harmful text patterns, including those not seen during training.

To guide this training more effectively, we vary the denoising timestep t, ensuring that suppression is learned progressively throughout the generation process. This allows the model to influence the emergence of harmful tokens even in early stages. To emphasize correction when the text is most visible, we apply the standard timestep-dependent weight $w(t)$ that increases as t approaches 0. In our imple-

[2]We also experimented with fine-tuning the CLIP encoder present in some DMs. However, this approach is not applicable across all architectures and consistently underperformed our backbone-level intervention. See Appendix D for details.

mentation, $w(t)$ follows a *logit-normal* schedule: timesteps are normalized to $[0, 1]$, passed through a logit transformation, and evaluated under a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. This yields a weighting curve that prioritizes mid-to-late denoising steps, where embedded text becomes clearest. The full training objective is:

$$\mathcal{L}(x_{\text{NSFW}}, I_{\text{NSFW}}(t), I_{\text{benign}}, t) = \|w(t) \cdot (f_\theta(I_{\text{NSFW}}(t), t, \phi(x_{\text{NSFW}})) - I_{\text{benign}})\|^2 . \quad (1)$$

where:

- $\phi(x_{\text{NSFW}})$ is the frozen text encoder's embedding of the NSFW prompt,
- $w(t)$ is a timestep-dependent weight emphasizing denoising steps close to $t = 0$,
- $f_\theta(I_{\text{NSFW}}(t), t, \phi(x_{\text{NSFW}}))$ is the predicted denoised image after one step.

This loss encourages the model to align its denoised prediction with the benign target image, despite being conditioned on the original NSFW prompt. At inference, it enables suppression of harmful text while preserving the surrounding visual content. Note that while `NSFW-Intervention` is designed for DMs, it can also be easily extended to the novel Visual Autoregressive Models (VARs) (Tian et al. 2024; Tang et al. 2024). We show this extension on the state-of-the-art Infinity (Han et al. 2024) model (Appendix G).

Results

`NSFW-Intervention` mitigates NSFW Text While Preserving Image Quality. We evaluate our method on the `ToxicBench` benchmark across multiple DMs (the detailed experimental setup is outlined in Appendix A. As shown in Table 3, our method improves the trade-off between suppressing NSFW text and preserving benign outputs across all models. On SD3, our method increases the suppression of NSFW text, improving the NGramLD from 1.84 to 5.47, while preserving benign generation with a score of 4.34, leading to a +1.13 differential NGramLD value between NSFW/benign text. Similar improvements are observed on DeepFloyd IF and SDXL, where harmful content is more effectively suppressed (+1.12 and +0.70 respectively) without sacrificing benign quality. Despite strong mitigation, `NSFW-Intervention` maintains image quality: the KID score increases by at most 9% across benign samples, and FID scores show minimal degradation (Appendix F). Qualitative results (Figure 4) illustrate that NSFW terms are rendered unreadable while benign text remains legible, with similar trends in SDXL and DeepFloyd IF (Figures 5 and 6 in the Appendix). We also report LD values in Table 5 (Appendix).

Ablation Studies. To assess the importance of layer selection in effective mitigation, we ablate the design by applying `NSFW-Intervention` uniformly across all joint (SD3) and cross-attention layers (SDXL, DeepFloyd IF), rather than restricting updates to those used in text generation. As detailed in Table 12 (Appendix), this broader intervention results in substantially weaker suppression of NSFW text. On SD3, NGramLD improves by only +0.49, compared to +3.63 when updates are limited to the text-generation layers (Table 3). Similar trends are observed on SDXL and DeepFloyd IF. We also show that `NSFW-Intervention` is efficient on prompt x2.4 longer than CreativeBench in Table 13 (Appendix).

Prompt Type	Before Intervention	After Intervention
NSFW	78.67±1.12	26.56±1.07
Misspelled NSFW	76.41±1.12	11.45±1.02
Benign	83.43±1.15	55.40±1.04

Table 4: User Study. Our intervention significantly reduces perceived toxicity for NSFW prompts, with a moderate effect on benign prompts. Results on misspelled NSFW demonstrate robustness even against character-level obfuscation.

A user study demonstrates the effectiveness of `NSFW-Intervention`. We conducted a user study measuring how participants perceived generated text before and after its application (Appendix C) to evaluate our intervention. Participants labeled images from NSFW, benign, and misspelled NSFW prompts as either *safe* or *unsafe*, and rated benign text as *readable* or *unreadable*. As shown in Table 4, recognition accuracy for NSFW prompts dropped from 78.67% to 26.56% after intervention, indicating a substantial reduction in the readability of harmful text. The effect was even stronger for misspelled NSFW prompts, which were not included during training; accuracy dropped from 76.41% to just 11.45%, highlighting strong generalization to adversarial variants. Meanwhile, benign text had post-intervention recognition at 55.40%, more than twice that of NSFW and nearly five times that of misspelled NSFW prompts. These findings demonstrate our method's ability to suppress harmful outputs while preserving benign content, even under distributional shifts.

Summary

We show that state-of-the-art DMs are highly susceptible to generating NSFW text embedded within images, a threat overlooked by prior mitigation efforts focused on visual content. We demonstrate that all leading DMs are vulnerable and that existing safety mechanisms fail to prevent harmful text generation without severely degrading benign text output. To address this, we introduce a general intervention strategy building on a unique safety-tuning of DMs backbones using a novel NSFW-benign text and image mapping. This approach significantly reduces NSFW text generation while preserving benign capabilities, and is applicable across architectures. To support further research, we introduce `ToxicBench`, an open-source benchmark designed to systematically evaluate and improve mitigation strategies for NSFW text generation in images. Thereby, we hope to contribute towards a more trustworthy deployment of these models.

Acknowledgements

This work was supported by the German Research Foundation (DFG) within the framework of the Weave Programme under the project titled "Protecting Creativity: On the Way to Safe Generative Models" with number 545047250.

References

Adolphs, L.; Gao, T.; Xu, J.; Shuster, K.; Sukhbaatar, S.; and Weston, J. 2023. The CRINGE Loss: Learning what language not to model. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Berg, M. 2025. NSFWDetector. Last accessed on 2025-01-17.

Bińkowski, M.; Sutherland, D. J.; Arbel, M.; and Gretton, A. 2018. Demystifying MMD GANs. In *International Conference on Learning Representations*.

Black Forest Labs. 2024. FLUX.1.

Chen, J.; Huang, Y.; Lv, T.; Cui, L.; Chen, Q.; and Wei, F. 2023. TextDiffuser: Diffusion Models as Text Painters. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 9353–9387. Curran Associates, Inc.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv:2305.06500.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*.

Gandikota, R.; Materzynska, J.; Fiotto-Kaufman, J.; and Bau, D. 2023. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2426–2436.

Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; and Smith, N. A. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Han, J.; Liu, J.; Jiang, Y.; Yan, B.; Zhang, Y.; Yuan, Z.; Peng, B.; and Liu, X. 2024. Infinity: Scaling Bitwise AutoRegressive Modeling for High-Resolution Image Synthesis. *arXiv preprint arXiv:2412.04431*.

Hanu, L.; and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

Hao, S.; Shelby, R.; Liu, Y.; Srinivasan, H.; Bhutani, M.; Ayan, B. K.; Poplin, R.; Poddar, S.; and Laszlo, S. 2024. Harm amplification in text-to-image models. *arXiv preprint arXiv:2402.01787*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 6840–6851.

Hu, Z.; Piet, J.; Zhao, G.; Jiao, J.; and Wagner, D. 2024. Toxicity Detection for Free. arXiv:2405.18822.

Jigsaw. 2025. Perspective API. Available at https://perspectiveapi.com.

Li, X.; Yang, Y.; Deng, J.; Yan, C.; Chen, Y.; Ji, X.; and Xu, W. 2024. Safegen: Mitigating sexually explicit content generation in text-to-image models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 4807–4821.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In *NeurIPS*.

notAI tech. 2025. NudeNet: lightweight Nudity detection. Last accessed on 2025-01-17.

OpenAI. 2025. GPT-4 Vision. Last accessed on 2025-01-17.

Ousidhoum, N.; Zhao, X.; Fang, T.; Song, Y.; and Yeung, D.-Y. 2021. Probing toxic content in large pre-trained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4262–4274.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Poppi, S.; Poppi, T.; Cocchi, F.; Cornia, M.; Baraldi, L.; and Cucchiara, R. 2025. Safe-clip: Removing nsfw concepts from vision-and-language models. In *European Conference on Computer Vision*, 340–356. Springer.

Poppi, S.; Yong, Z.-X.; He, Y.; Chern, B.; Zhao, H.; Yang, A.; and Chi, J. 2024. Towards Understanding the Fragility of Multilingual LLMs against Fine-Tuning Attacks. *arXiv preprint arXiv:2410.18210*.

Qu, Y.; Shen, X.; He, X.; Backes, M.; Zannettou, S.; and Zhang, Y. 2023. 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*, 3403–3417.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning (ICML)*, 8748–8763.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint*, arXiv:2204.06125.

Rando, J.; Paleka, D.; Lindner, D.; Heim, L.; and Tramèr, F. 2022. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*.

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

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 234–241.

Schramowski, P.; Brack, M.; Deiseroth, B.; and Kersting, K. 2023. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22522–22531.

Song, Y.; and Ermon, S. 2020. Improved Techniques for Training Score-Based Generative Models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 12438–12448.

StabilityAI. 2023. DeepFloyd IF: a novel state-of-the-art open-source text-to-image model with a high degree of photorealism and language understanding. https://github.com/deepfloyd/IF. Last accessed on 2025-01-17.

Staniszewski, Ł.; Cywiński, B.; Boenisch, F.; Deja, K.; and Dziedzic, A. 2025. Precise Parameter Localization for Textual Generation in Diffusion Models. In *The Thirteenth International Conference on Learning Representations*.

Suau, X.; Delobelle, P.; Metcalf, K.; Joulin, A.; Apostoloff, N.; Zappella, L.; and Rodríguez, P. 2024. Whispering experts: Neural interventions for toxicity mitigation in language models. *arXiv preprint arXiv:2407.12824*.

Tang, H.; Wu, Y.; Yang, S.; Xie, E.; Chen, J.; Chen, J.; Zhang, Z.; Cai, H.; Lu, Y.; and Han, S. 2024. Hart: Efficient visual generation with hybrid autoregressive transformer. *arXiv preprint arXiv:2410.10812*.

Tian, K.; Jiang, Y.; Yuan, Z.; Peng, B.; and Wang, L. 2024. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*.

Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Wei, B.; Huang, K.; Huang, Y.; Xie, T.; Qi, X.; Xia, M.; Mittal, P.; Wang, M.; and Henderson, P. 2024. Assessing the Brittleness of Safety Alignment via Pruning and Low-Rank Modifications. In *Forty-first International Conference on Machine Learning*.

Yang, Y.; Gui, D.; Yuan, Y.; Liang, W.; Ding, H.; Hu, H.; and Chen, K. 2024a. Glyphcontrol: Glyph conditional control for visual text generation. *Advances in Neural Information Processing Systems*, 36.

Yang, Y.; Hui, B.; Yuan, H.; Gong, N.; and Cao, Y. 2024b. Sneakyprompt: Jailbreaking text-to-image generative models. In *2024 IEEE symposium on security and privacy (SP)*, 897–912. IEEE.

Zong, Y.; Bohdal, O.; Yu, T.; Yang, Y.; and Hospedales, T. 2024. Safety Fine-Tuning at (Almost) No Cost: A Baseline for Vision Large Language Models. In *Forty-first International Conference on Machine Learning*.