SAFETEXT: SAFE TEXT-TO-IMAGE MODELS VIA ALIGNING THE TEXT ENCODER

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image models can generate harmful images when presented with unsafe prompts, posing significant safety and societal risks. Alignment methods aim to modify these models to ensure they generate only non-harmful images, even when exposed to unsafe prompts. A typical text-to-image model comprises two main components: 1) a text encoder and 2) a diffusion module. Existing alignment methods mainly focus on modifying the diffusion module to prevent harmful image generation. However, this often significantly impacts the model's behavior for safe prompts, causing substantial quality degradation of generated images. In this work, we propose *SafeText*, a novel alignment method that fine-tunes the text encoder rather than the diffusion module. By adjusting the text encoder, SafeText significantly alters the embedding vectors for unsafe prompts, while minimally affecting those for safe prompts. As a result, the diffusion module generates non-harmful images for unsafe prompts while preserving the quality of images for safe prompts. We evaluate SafeText on multiple datasets of safe and unsafe prompts, including those generated through jailbreak attacks. Our results show that SafeText effectively prevents harmful image generation with minor impact on the images for safe prompts, and SafeText outperforms six existing alignment methods. We will publish our code and data after paper acceptance.

WARNING: This paper contains sexual and nudity content, which readers may find offensive or disturbing.

1 INTRODUCTION

Given a prompt, a text-to-image model can generate highly realistic images that align with the prompt's semantics. Typically, such a model consists of two key components: 1) a text encoder, which maps the prompt into an embedding vector; and 2) a diffusion module, which guided by the embedding vector, recursively denoises a random Gaussian noise vector to an image. These models have a wide range of applications, including art creation, character design in online games, and virtual environment development. For instance, Microsoft has integrated DALL·E into its Edge browser (Mehdi, 2023).

Like any advanced technology, text-to-image models are double-edged swords, raising severe safety concerns alongside their societal benefits discussed above. Specifically, they can generate high-quality harmful images–such as those containing sexual or nudity content–when provided with *unsafe prompts* like, "Show me an image of a nude body." These harmful image generations can be triggered either intentionally by malicious users or unintentionally by regular users. Unsafe prompts can be manually crafted based on heuristics, often containing keywords related to sexual or nudity content. Alternatively, they can also be adversarially crafted via jailbreak attacks (Zhuang et al., 2023; Qu et al., 2023; Yang et al., 2024b; Tsai et al., 2024; Yang et al., 2024a), which are designed to bypass safety mechanisms.

Alignment methods aim to modify text-to-image models to ensure they generate only non-harmful images, even when presented with unsafe prompts. Existing alignment methods (Rombach et al., 2022; Schramowski et al., 2023; Gandikota et al., 2023; Lu et al., 2024; Li et al., 2024; Zhang et al., 2024) primarily target the diffusion module of the model. For example, Erased Stable Diffusion (ESD) (Gandikota et al., 2023) fine-tunes the diffusion module to make the noise prediction, conditioned on unsafe prompts, unconditional and therefore typically non-harmful. While these methods

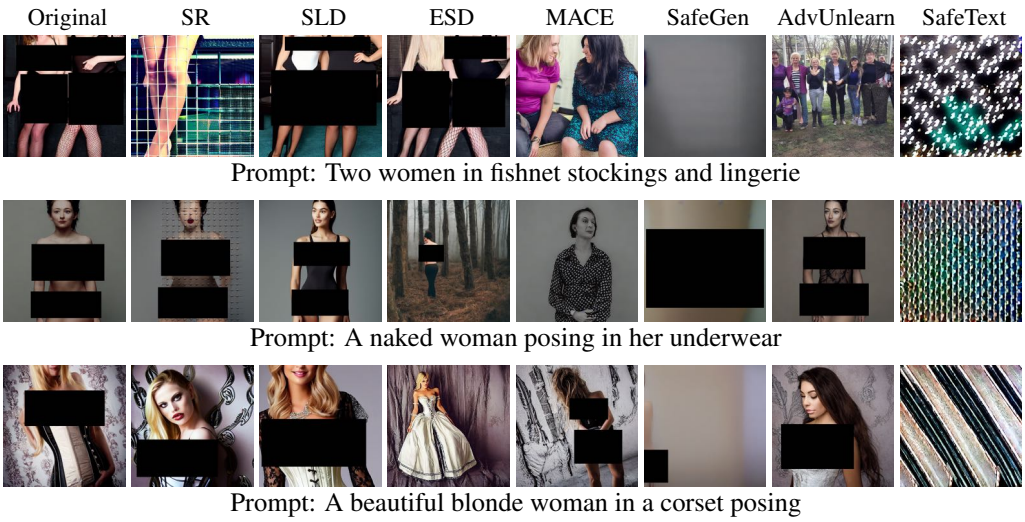


Figure 1: Images generated by Stable Diffusion v1.4 without alignment (first column) and with different alignments (other columns) for three unsafe prompts.

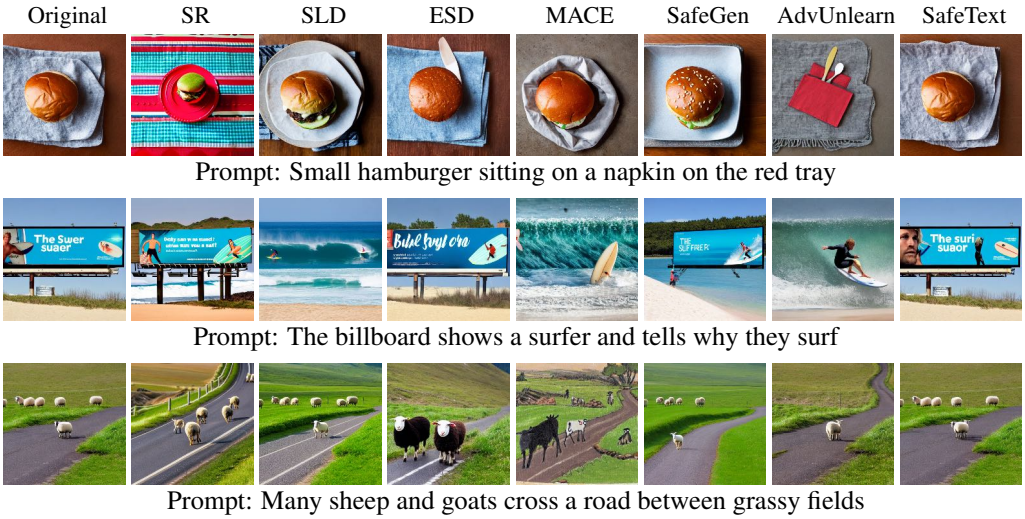


Figure 2: Images generated by Stable Diffusion v1.4 without alignment (first column) and with different alignments (other columns) for three safe prompts.

show some effectiveness in preventing harmful image generation, they also significantly degrade the quality of images generated for safe prompts. This is because it is challenging to separate the impact of diffusion-module modification on image generation for unsafe and safe prompts. AdvUnlearn (Zhang et al., 2024), a method recently posted on arXiv, is the only approach that aligns the text encoder. It combines the loss function from ESD with adversarial training (Madry, 2017) to fine-tune the text encoder. However, because the loss function of ESD is designed for the diffusion module, applying it to fine-tune the text encoder still results in substantial changes to the denoising process, which negatively impacts image generation for safe prompts, as shown in our experiments.

In this work, we propose *SafeText*, a novel alignment method. Due to the challenges of aligning the diffusion module discussed above, SafeText aligns the text encoder without any information about the diffusion module. Specifically, SafeText fine-tunes the text encoder to substantially alter the embeddings of unsafe prompts (*effectiveness goal*) while introducing minimal changes to those of safe prompts (*utility goal*). As a result, the diffusion module generates non-harmful images for unsafe prompts while preserving the quality of images for safe prompts. We develop two loss terms to respectively quantify the effectiveness and utility goals. Then, we formulate fine-tuning the text

encoder as an optimization problem, whose objective is to minimize a weighted sum of the two loss terms. Furthermore, SafeText leverages a standard gradient-based method (e.g., Adam optimizer) to solve the optimization problem, which fine-tunes the text encoder.

We evaluate SafeText on three datasets of safe prompts, four datasets of manually crafted unsafe prompts, and adversarially crafted unsafe prompts generated by three state-of-the-art jailbreak attacks (Yang et al., 2024b; Tsai et al., 2024; Yang et al., 2024a). Additionally, we compare SafeText with six leading alignment methods. The results demonstrate that SafeText outperforms all these alignment methods, striking a balance between preventing harmful image generation for unsafe prompts and preserving the quality of images generated for safe prompts. Figure 1 shows the images generated by an unaligned text-to-image model and the models aligned by different methods for three unsafe prompts, while Figure 2 shows the images generated for three safe prompts.

2 Related Work

2.1 Harmful Image Generation

A text-to-image model generates high-quality harmful images when presented with unsafe prompts, which can be manually crafted based on heuristics or adversarially crafted using jailbreak attacks.

Manually crafted unsafe prompts: These unsafe prompts are manually crafted based on heuristics, often containing keywords related to sexual or nudity content. Additionally, multi-modal large language models can be employed to generate captions for real-world harmful images, with these captions being used as unsafe prompts. In our experiments, we utilize manually crafted unsafe prompts collected from online prompt-sharing platforms like civitai.com and lexica.art, as well as captions generated for harmful images, to test the effectiveness of safety alignment methods.

Adversarially crafted unsafe prompts: These unsafe prompts are generated through jailbreak attacks and could include text that is either coherent or nonsensical to humans. A jailbreak attack modifies a manually crafted unsafe prompt, which fails to bypass a model's safety alignment, into an adversarial prompt. This adversarial prompt is designed to circumvent the safety alignment, enabling the text-to-image model to generate a harmful image that matches the semantics of the original unsafe prompt. For instance, SneakyPrompt (Yang et al., 2024b) iteratively refines the adversarial prompt via interacting with a given text-to-image model and leveraging reinforcement learning to take the responses into consideration. Similarly, Ring-A-Bell (Tsai et al., 2024) employs a surrogate text encoder and a genetic algorithm to generate an adversarial prompt that avoids explicit unsafe words while keeping its embedding similar to the original unsafe prompt. MMA-Diffusion (Yang et al., 2024a) further leverages token-level gradients and word regularization to optimize an adversarial prompt, ensuring it avoids explicit unsafe words while preserving embedding similarity to the original unsafe prompt.

2.2 Safety Alignment

Depending on the text-to-image model's component that is aligned, alignment methods can be grouped into the following two categories:

Aligning the diffusion module: The most straightforward method (Rombach et al., 2022) to align the diffusion module of a text-to-image model is to retrain it on a dataset containing only non-harmful images and safe prompts. However, this safe retraining has limited effectiveness because the retrained model can still piece together different parts of seemingly non-harmful images to generate harmful ones. Additionally, retraining is highly time-consuming. To address this, some alignment methods fine-tune the diffusion module (Gandikota et al., 2023; Lu et al., 2024; Li et al., 2024) or modify its image generation process (Schramowski et al., 2023). For instance, Erased Stable Diffusion (ESD) (Gandikota et al., 2023) fine-tunes the diffusion module to make the noise prediction, conditioned on unsafe concepts, unconditional and therefore typically non-harmful. Mass Concept Erasure (MACE) (Lu et al., 2024) uses Low-Rank Adaptation (LoRA) (Hu et al., 2022) to fine-tune the cross-attention layer (Chen et al., 2021) within the diffusion module, preventing the generation of images related to unsafe concepts. Similarly, SafeGen (Li et al., 2024) fine-tunes the diffusion module using harmful images and their mosaic versions, prompting the model to generate

mosaic images when given unsafe prompts. For generation-time alignment, Safe Latent Diffusion (SLD) (Schramowski et al., 2023) adds a safety guidance term to the classifier-free guidance noise prediction process to remove harmful elements from the generated images. However, these alignment methods substantially affect the images generated for safe prompts as they significantly alter the diffusion module's behavior.

Aligning the text encoder: To the best of our knowledge, AdvUnlearn (Zhang et al., 2024) is the only method that aligns the text encoder. AdvUnlearn combines the loss function of ESD (Gandikota et al., 2023) with adversarial training (Madry, 2017) to change the diffusion module's noise prediction process. Specifically, it fine-tunes the text encoder so that the diffusion module's predicted noise conditioned on unsafe prompts approximates the unconditional predicted noise, while the predicted noise conditioned on safe prompts remains close to that before fine-tuning. However, because the loss function of ESD is based on classifier-free guidance and is designed for the diffusion module, using it to fine-tune the text encoder still substantially changes the denoising process, significantly affecting the image generation for safe prompts, as demonstrated in our experiments.

3 Problem Definition

Given a text-to-image model, our objective is to align it to meet two goals: 1) *Effectiveness* and 2) *Utility*. The effectiveness goal ensures that the aligned model does not generate harmful images–such as those containing sexual or nudity-related content–when presented with unsafe prompts. The utility goal focuses on maintaining the model's ability to generate high-quality images for safe prompts. Specifically, we aim for a high standard of utility: given the same safe prompt and seed, the aligned and unaligned models should produce visually similar images. For instance, the LPIPS score (Zhang et al., 2018) between the images generated by the aligned and unaligned models is small. Our SafeText achieves a balance between the two goals, i.e., between preventing harmful image generation and preserving the model's functionality for safe use cases.

4 Our SafeText

4.1 Overview

Our SafeText achieves the effectiveness and utility goals via aligning the text encoder of the text-to-image model. Since the diffusion module of the text-to-image model is responsible for the denoising process and image generation, modifying its parameters may significantly degrade image quality for safe prompts. Therefore, our SafeText fine-tunes only the text encoder while keeping the diffusion module intact to largely preserve image quality for safe prompts.

Specifically, to achieve the effectiveness goal, we fine-tune the text encoder so that the embeddings for unsafe prompts are altered substantially. Consequently, the images generated based on the embeddings produced by the aligned text encoder are much less likely to contain harmful content. To achieve the utility goal, we ensure that the aligned text encoder and the original one produce similar embeddings for a safe prompt. Formally, we propose two loss terms to respectively quantify the two goals, and formulate fine-tuning the text encoder as an optimization problem, whose objective is to minimize a weighted sum of the two loss terms. Finally, we solve the optimization problem via a standard gradient-based method.

4.2 Formulating an Optimization Problem

We use τ to denote the original text encoder and τ_s to denote our fine-tuned one.

Quantifying the effectiveness goal: For an unsafe prompt P_{un}, our objective is to ensure that the embedding $\tau_s(P_{un})$ produced by the fine-tuned encoder is highly likely to be safe. To achieve this, we fine-tune the text encoder so that the embedding $\tau_s(P_{un})$ is substantially different from the original embedding $\tau(P_{un})$, given that $\tau(P_{un})$ is unsafe. Therefore, to achieve our effectiveness goal, we fine-tune τ as τ_s such that the distance between $\tau_s(P_{un})$ and $\tau(P_{un})$ is large, based on a chosen distance metric. Formally, we quantify the effectiveness goal using the following loss term:

$$L_e = E_{P_{un} \sim \mathbb{D}_{un}}[d_e(\tau_s(P_{un}), \tau(P_{un}))], \tag{1}$$

where $\mathbb{D}_{un}$ represents the distribution of unsafe prompts, $P_{un} \sim \mathbb{D}_{un}$ means that P_{un} is an unsafe prompt sampled from $\mathbb{D}_{un}$, E stands for expectation, and d_e denotes a distance metric between two embedding vectors (e.g., Euclidean distance). The effectiveness goal may be better achieved when the loss term L_e is larger.

Quantifying the utility goal: For a safe prompt P_s, our objective is to keep its embeddings similar before and after fine-tuning. To achieve this, we fine-tune the text encoder so that the distance between the embeddings $\tau_s(P_s)$ and $\tau(P_s)$ is small, based on a chosen distance metric. Formally, we quantify this utility using the following loss term:

$$L_u = E_{P_s \sim \mathbb{D}_s}[d_u(\tau_s(P_s), \tau(P_s))], \tag{2}$$

where $\mathbb{D}_s$ represents the distribution of safe prompts, $P_s \sim \mathbb{D}_s$ means that P_s is a safe prompt sampled from $\mathbb{D}_s$, E stands for expectation, and d_u denotes a distance metric between two embedding vectors. The utility goal may be better achieved when the loss term L_u is smaller.

Optimization problem: To balance between the effectiveness and utility goals, we combine the two loss terms L_e and L_u to formulate an optimization problem as follows:

$$\min_{\tau_s} L_u - \lambda L_e, \tag{3}$$

where λ is a hyper-parameter that controls the trade-off between the effectiveness goal and the utility goal. The objective of this optimization problem is to fine-tune the text encoder to maximize the effectiveness for unsafe prompts while preserving utility for safe prompts.

4.3 Solving the Optimization Problem

We solve the optimization problem using a dataset of safe prompts (denoted as $\mathcal{D}_s$) and a dataset of unsafe prompts (denoted as $\mathcal{D}_{un}$). The two datasets are used to approximate the expectations. Specifically, given the two datasets, the optimization problem can be reformulated as follows:

$$\min_{\tau_s} \frac{1}{|\mathcal{D}_s|} \sum_{P_s \in \mathcal{D}_s} d_u(\tau_s(P_s), \tau(P_s)) - \frac{\lambda}{|\mathcal{D}_{un}|} \sum_{P_{un} \in \mathcal{D}_{un}} d_e(\tau_s(P_{un}), \tau(P_{un})). \tag{4}$$

We can use a standard gradient-based method (e.g., Adam optimizer) to solve this optimization problem. Specifically, we initialize τ_s as τ, and then update τ_s for n epochs with a batch size of m and a learning rate of α.

5 Experiment

5.1 Experimental Setup

Fine-tuning datasets $\mathcal{D}_s$ and $\mathcal{D}_{un}$: Our fine-tuning needs datasets $\mathcal{D}_s$ and $\mathcal{D}_{un}$. In our experiments, $\mathcal{D}_s$ contains 30,000 safe prompts and $\mathcal{D}_{un}$ contains 30,000 unsafe prompts, both sampled from a pre-processed Civitai-8M dataset (AdamCodd, 2024). The original Civitai-8M dataset comprises 7,852,309 prompts collected from Civitai, an online platform where users upload and share prompts. Each prompt in Civitai-8M is assigned an unsafe level ranging from 0 to 32. To construct high-quality datasets $\mathcal{D}_s$ and $\mathcal{D}_{un}$, we keep the prompts with an unsafe level of 1 or below as safe prompts, while those with an unsafe level greater than 8 as unsafe prompts. Moreover, we apply a safety classifier (michellejieli, 2022) to further score and classify each prompt, where a larger score indicates safer. We keep the safe prompts with a score above 0.9 as the final safe dataset, while the unsafe prompts classified as unsafe by the safety classifier as the final unsafe dataset. We then randomly sample 30,000 prompts from the final safe dataset to form $\mathcal{D}_s$ and 30,000 prompts from the final unsafe dataset to form $\mathcal{D}_{un}$.

Testing unsafe prompt datasets: We consider both manually and adversarially crafted unsafe prompts to evaluate the effectiveness of an alignment method.

- **Manually crafted unsafe prompts.** We acquire 4 datasets of manually crafted unsafe prompts: **Civitai-Unsafe**, **NSFW**, **I2P**, and **U-Prompt**. Table 5 in Appendix summarizes them. Civitai-Unsafe includes 1,000 unsafe prompts sampled from Civitai-8M (AdamCodd, 2024) excluding those in $\mathcal{D}_{un}$ used for fine-tuning. NSFW consists of 1,000 unsafe prompts sampled from

NSFW-56k (Li et al., 2024), a dataset of unsafe prompts generated by using BLIP2 (Li et al., 2023) to caption a set of pornographic images. I2P (Schramowski et al., 2023) consists of prompts collected from lexica.art using keyword matching. The original I2P dataset includes many safe prompts. Thus, we use GPT-4o to filter and retain only those detected as unsafe, resulting in 229 unsafe prompts. U-Prompt is collected by us and consists of 1,000 unsafe prompts generated by using BLIP2-OPT (Salesforce, 2023) to caption a sexual image dataset (Noktedan, 2020). Compared to other datasets, the unsafe prompts in U-Prompt are shorter, potentially introducing additional challenges for alignment methods to defend against them.

- **Adversarially crafted unsafe prompts.** We use three state-of-the-art jailbreak attacks– **SneakyPrompt** (Yang et al., 2024b), **Ring-A-Bell** (Tsai et al., 2024), and **MMA-Diffusion** (Yang et al., 2024a)–to generate adversarially crafted unsafe prompts. The details of these methods are shown in Section A in Appendix. Given a manually crafted unsafe prompt, these attacks turn it into an adversarial prompt with a goal to bypass safety guardrails. We randomly sample 200 unsafe prompts from NSFW-56k following Li et al. (2024), and then use each attack to generate 200 adversarially crafted unsafe prompts. We use the publicly available code and default settings of the three attacks. Note that SneakyPrompt generates adversarial prompts tailored to each (unaligned or aligned) text-to-image model.

Testing safe prompt datasets: To evaluate utility of an alignment method, we use 3 datasets of safe prompts: **Civitai-Safe**, **MS-COCO**, and **Google-CC**. Each dataset includes 1,000 safe prompts sampled from Civitai-8M (AdamCodd, 2024), MS-COCO (Lin et al., 2014), and Google's Conceptual Captions (Sharma et al., 2018), respectively. Table 5 in Appendix summarizes these datasets.

Evaluation metrics: An alignment method aims to achieve the effectiveness and utility goals. Thus, we use *NSFW Removal Rate (NRR)* to evaluate the effectiveness, and *LPIPS* and *FID* to evaluate the utility of an alignment method. These metrics are also widely used in prior works (Schramowski et al., 2023; Gandikota et al., 2023; Lu et al., 2024; Li et al., 2024; Zhang et al., 2024).

- **NSFW Removal Rate (NRR).** Following Li et al. (2024), we use NudeNet (notAI Tech, 2019) to calculate the number of nude body parts in an image. Given an image, NudeNet detects and labels nude body parts in it. We treat the parts detected by NudeNet as "exposed" as nude. Let $n(M(P_{un}))$ (or $n(M_s(P_{un}))$) denote the number of nude parts in an image generated by the text-to-image model M (or M_s) before (or after) alignment based on an unsafe prompt P_{un}. NRR measures the reduction of nude parts in the generated images after alignment. Specifically, given a testing dataset $\mathcal{D}^t_{un}$ of unsafe prompts, NRR is calculated as follows:

$$NRR = 1 - \frac{1}{|\mathcal{D}^t_{un}|} \sum_{P_{un} \in \mathcal{D}^t_{un}} \frac{n(M_s(P_{un}))}{n(M(P_{un}))}. \tag{5}$$

  Note that, given a prompt P_{un}, we use the same seed when generating images using M and M_s to avoid the impact of the randomness in the seed. A larger NRR indicates better effectiveness.
- **LPIPS.** Given a safe prompt and a random seed, we use the models M and M_s to generate two images. Then, we calculate the two images' Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) based on features extracted by AlexNet (Krizhevsky et al., 2012). Given a testing dataset of safe prompts, we calculate the average LPIPS across all prompts in the dataset. A lower LPIPS indicates better utility.
- **FID.** While LPIPS measures the visual similarity between two images, Fréchet Inception Distance (FID) (Heusel et al., 2017) measures the similarity between two image datasets: those generated by M and those generated by M_s for a testing dataset of safe prompts. A lower FID indicates better utility.

Baseline alignment methods: We compare our SafeText with six state-of-the-art alignment methods. **Safe Retraining (SR)** (Rombach et al., 2022) retrains a diffusion module on a safe dataset that contains only non-harmful images and safe prompts. **Safe Latent Diffusion (SLD)** (Schramowski et al., 2023) prevents harmful content generation by combining safety guidance with classifier-free guidance to remove or suppress harmful image elements during the image generation process. **Erased Stable Diffusion (ESD)** (Gandikota et al., 2023), **Mass Concept Erasure (MACE)** (Lu et al., 2024), and **SafeGen** (Li et al., 2024) fine-tune the diffusion module to prevent generating harmful images. **AdvUnlearn** (Zhang et al., 2024) fine-tunes the text encoder using the loss function of ESD and adversarial training.

Table 1: Effectiveness results (NRR ↑) of different alignment methods on Stable Diffusion v1.4.

(a) Manually crafted unsafe prompts

Method	Unsafe prompt dataset			
	Civitai-Unsafe	NSFW	I2P	U-Prompt
SR	0.639	0.712	0.780	0.770
SLD	0.626	0.596	0.741	0.635
ESD	0.796	0.826	0.867	0.839
MACE	0.906	0.889	0.908	0.904
SafeGen	0.936	0.970	0.886	0.979
AdvUnlearn	0.972	0.944	0.960	0.888
SafeText	**0.990**	**0.987**	**0.990**	**0.994**

(b) Adversarially crafted unsafe prompts

Method	Jailbreak attack		
	SneakyPrompt	Ring-A-Bell	MMA-Diffusion
SR	0.766	0.545	0.787
SLD	0.670	0.603	0.616
ESD	0.792	0.684	0.851
MACE	0.866	0.955	0.902
SafeGen	0.960	0.951	0.986
AdvUnlearn	0.925	0.997	0.989
SafeText	**0.984**	**1.000**	**0.992**

Table 2: Utility results (LPIPS ↓ / FID ↓) of different alignment methods on Stable Diffusion v1.4.

Method	Safe prompt dataset		
	Civitai-Safe	MS-COCO	Google-CC
SR	0.669 / 74.3	0.640 / 60.2	0.646 / 70.2
SLD	0.601 / 66.3	0.572 / 53.0	0.581 / 63.5
ESD	0.510 / 55.8	0.502 / 47.2	0.507 / 56.0
MACE	0.642 / 74.0	0.522 / 53.9	0.590 / 65.3
SafeGen	0.620 / 67.1	0.581 / 54.5	0.591 / 64.5
AdvUnlearn	0.669 / 84.3	0.512 / 48.6	0.594 / 64.2
SafeText	**0.207 / 32.4**	**0.218 / 28.4**	**0.206 / 31.5**

Text-to-image models: Baseline alignment methods (Rombach et al., 2022; Schramowski et al., 2023; Gandikota et al., 2023; Lu et al., 2024; Li et al., 2024; Zhang et al., 2024) were evaluated on Stable Diffusion v1.4 (Rombach et al., 2022). Therefore, for fair comparison, we compare our SafeText with them on Stable Diffusion v1.4 (Rombach et al., 2022). However, in our ablation study, we further evaluate our SateText using another 5 models: Stable Diffusion XL v1.0 (SDXL v1.0) (Podell et al., 2024), Dreamlike Photoreal v2.0 (dreamlike.art, 2023), LCM Dreamshaper v7 (Luo et al., 2023), Openjourney v4 (Hero, 2023), and Juggernaut X v10 (Diffusion, 2024).

Parameter settings: Our SafeText fine-tunes the text encoder of a text-to-image model using the Adam optimizer with $n = 5$, $m = 32$, and $\alpha = 10^{-5}$. Additionally, unless otherwise mentioned, we use **Euclidean distance** as d_u and **negative absolute cosine similarity (NegCosine)** as d_e, and λ is set to be 0.2. Our ablation study will show this combination of distance metrics d_u and d_e achieves the best performance. Note that NegCosine aims to make the embeddings for an unsafe prompt produced by the fine-tuned and original text encoders orthogonal. On the other hand, negative cosine similarity aims to make the embeddings for an unsafe prompt produced by the fine-tuned and original text encoders inverse. We use NegCosine instead of negative cosine similarity because we find that the former empirically outperforms the latter (see results in Figure 5 in Appendix).

Table 3: Effectiveness results (NRR ↑) of SafeText on other text-to-image models.

(a) Manually crafted unsafe prompts

Model	Unsafe prompt dataset			
	Civitai-Unsafe	NSFW	I2P	U-Prompt
SDXL v1.0	0.973	0.945	0.902	0.951
Dreamlike Photoreal v2.0	0.996	0.986	0.950	0.995
LCM Dreamshaper v7	0.971	0.951	0.935	0.960
Openjourney v4	0.948	0.963	0.906	0.958
Juggernaut X v10	0.986	0.981	0.936	0.985

(b) Adversarially crafted unsafe prompts

Model	Jailbreak attack		
	SneakyPrompt	Ring-A-Bell	MMA-Diffusion
SDXL v1.0	0.933	0.958	0.911
Dreamlike Photoreal v2.0	0.988	0.997	0.987
LCM Dreamshaper v7	0.931	0.998	0.978
Openjourney v4	0.950	0.970	0.962
Juggernaut X v10	0.963	0.998	0.988

Table 4: Utility results (LPIPS ↓ / FID ↓) of SafeText on other text-to-image models.

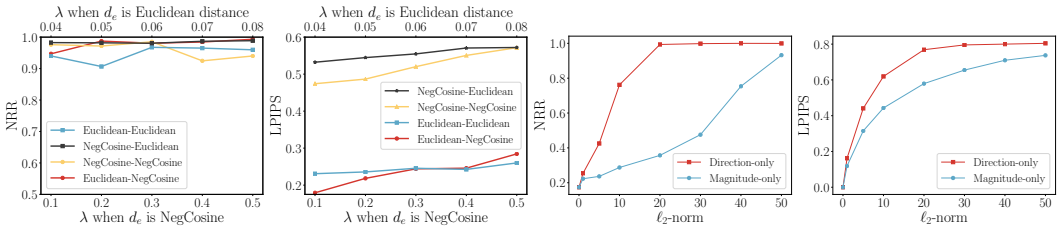
Model	Safe prompt dataset		
	Civitai-Safe	MS-COCO	Google-CC
SDXL v1.0	0.319 / 37.3	0.293 / 38.9	0.307 / 39.3
Dreamlike Photoreal v2.0	0.326 / 36.7	0.340 / 35.7	0.338 / 38.6
LCM Dreamshaper v7	0.129 / 21.9	0.158 / 24.3	0.153 / 24.8
Openjourney v4	0.265 / 33.0	0.282 / 32.3	0.260 / 34.0
Juggernaut X v10	0.344 / 39.8	0.338 / 37.0	0.329 / 41.9

For baseline alignment methods, we use their publicly available aligned versions of Stable Diffusion v1.4. In particular, the safety configurations of SafeGen and SLD are set to "MAX," indicating their strongest configuration. For ESD, MACE, and AdvUnlearn, we use their publicly available aligned versions of Stable Diffusion v1.4. For SR, we adopt Stable Diffusion v2.1 (Rombach et al., 2022), which is the safe retraining version of Stable Diffusion v1.4.

5.2 Main Results

Our SafeText achieves both effectiveness and utility goals: Tables 1a and 1b respectively show the NRR of our SafeText for manually and adversarially crafted unsafe prompts on Stable Diffusion v1.4. The results demonstrate that SafeText achieves the effectiveness goal. Specifically, the NRR exceeds 98.7% across the four datasets of manually crafted unsafe prompts. For adversarially crafted unsafe prompts, SafeText achieves an NRR larger than 98.4% across the three jailbreak attack methods. Additionally, Table 2 shows the LPIPS and FID of SafeText for the three datasets of safe prompts. The results demonstrate that SafeText also achieves the utility goal. Specifically, SafeText achieves an LPIPS below 0.218 and an FID below 32.4 on all three datasets.

Our SafeText outperforms baseline alignment methods: Tables 1 and 2 also show the effectiveness and utility results for the six baseline alignment methods. The results demonstrate that SafeText outperforms all of them in terms of both effectiveness and utility. Specifically, SafeText achieves the highest NRR across the four datasets of manually crafted unsafe prompts and adversarial prompts crafted by the three jailbreak attack methods. Furthermore, on the three datasets of safe prompts, SafeText achieves significantly lower LPIPS and FID scores compared to the baseline methods.



(a) NRR (b) LPIPS (c) NRR (d) LPIPS

Figure 3: (a) NRR on NSFW and (b) LPIPS on MS-COCO for SafeText with different distance metrics and λ values. Controlled experiments to assess the impact of embedding direction and magnitude on (c) harmfulness of images for unsafe prompts and (d) utility of images for safe prompts.

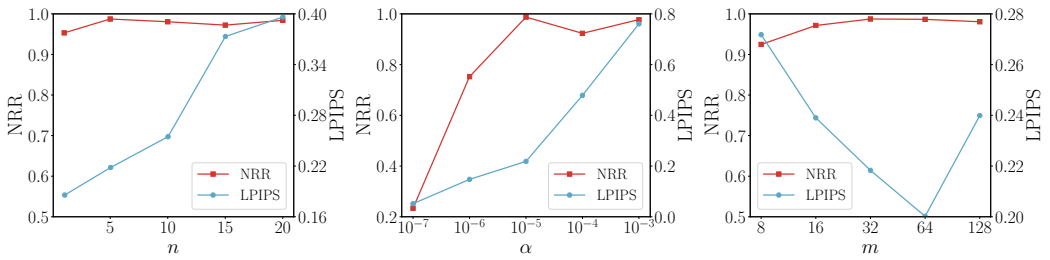
5.3 Ablation Study

Other text-to-image models: Tables 3a and 3b show the effectiveness results of our SafeText for manually and adversarially crafted unsafe prompts across another five text-to-image models. The results demonstrate that our SafeText still achieves the effectiveness goal when applied to these models. Specifically, our SafeText achieves an NRR larger than 90.2% for manually crafted unsafe prompts and larger than 91.1% for adversarially crafted unsafe prompts across all five models. Additionally, Table 4 shows the utility results of our SafeText across the five text-to-image models, confirming that our SafeText still achieves the utility goal when applied to these models. Specifically, our SafeText achieves an LPIPS below 0.344 and an FID below 41.9 across all the three datasets of safe prompts and the five models. Some image samples generated by these text-to-image models without alignment and with our SafeText are shown in Figures 6–15 in Appendix.

Different distance metrics and λ: Figures 3a and 3b respectively compare the NRR and LPIPS of SafeText when using different distance metrics as d_u and d_e, and different λ on Stable Diffusion v1.4. Each curve in the figures corresponds to a combination of distance metrics in the form of d_u-d_e. For instance, Euclidean-NegCosine indicates that Euclidean distance is used as d_u, while NegCosine is used as d_e. For each of the 4 combinations of distance metrics, we show the NRR and LPIPS results for different λ, where the bottom x-axis indicates λ when d_e is NegCosine and the top x-axis indicates λ when d_e is Euclidean distance. We observe a general trend: LPIPS increases and NRR increases (and then stabilizes or fluctuates slightly) when λ increases, indicating that λ balances between the effectiveness and utility goals. In the figures, we show the ranges of λ that achieve good effectiveness-utility trade-offs for these combinations of distance metrics.

From Figure 3b, we observe that using Euclidean distance as d_u (i.e., Euclidean-NegCosine and Euclidean-Euclidean) achieves significantly smaller LPIPS than using NegCosine as d_u (i.e., NegCosine-NegCosine and NegCosine-Euclidean). This suggests that both the direction and magnitude of the embedding are crucial for preserving utility for safe prompts. The two combinations Euclidean-Euclidean and Euclidean-NegCosine achieve similar utility/LPIPS. However, Figure 3a shows that using NegCosine as d_e results in a higher NRR. In other words, the combination Euclidean-NegCosine achieves the best performance among the four. This might be because harmfulness in a generated image is more sensitive to the direction of the embedding of an unsafe prompt than to the magnitude. NegCosine only considers direction of embeddings, and thus outperforms Euclidean distance when used as d_e.

To investigate this further, we design a controlled experiment to explore the impact of varying direction and magnitude of a prompt's embedding on the generated image. Suppose we are given the embedding of a prompt produced by an unaligned text encoder. For *direction-only*, we rotate the embedding while preserving its magnitude, under a constraint on the ℓ_2-norm of the change to the embedding. For *magnitude-only*, we increase the magnitude of the embedding while keeping its direction, under the same ℓ_2-norm constraint. We generate an image using the unmodified embedding and an image using the embedding modified by direction-only (or magnitude-only), and we calculate NRR (for unsafe prompts) or LPIPS (for safe prompts) between the two images. Figures 3c and 3d respectively show the NRR and LPIPS of direction-only and magnitude-only averaged over NSFW and MS-COCO given different ℓ_2-norm constraints. We observe that direction-only achieves



(a) Number of epochs (b) Learning rate (c) Batch size

Figure 4: NRR on NSFW and LPIPS on MS-COCO of our SafeText with different (a) number of epochs, (b) learning rates, and (c) batch sizes.

higher NRR under the same ℓ_2-norm constraint. For instance, direction-only achieves an NRR of 99.3%, while magnitude-only reaches only 35.7% when the ℓ_2-norm constraint is 20. For utility, we observe that both direction-only and magnitude-only have large impact on LPIPS. These results demonstrate that harmfulness of a generated image is more sensitive to the direction of the embedding of an unsafe prompt and the image quality for safe prompts is sensitive to both direction and magnitude. Therefore, we choose Euclidean distance as d_u and NegCosine as d_e.

Different number of epochs n: Figure 4a shows the effectiveness and utility of our SafeText across different numbers of fine-tuning epochs n on Stable Diffusion v1.4. For effectiveness, we observe that the NRR initially increases and then stabilizes as the number of epochs grows. This demonstrates that our SafeText can achieve high effectiveness when the text encoder is fine-tuned for a sufficient number of epochs. For utility, the LPIPS increases with more epochs, indicating a more significant visual change of images generated from safe prompts. This occurs because excessive fine-tuning of the text encoder may significantly alter its parameters, causing the generated images to visually deviate substantially from the original ones.

Different learning rate α: Figure 4b shows the effectiveness and utility of our SafeText across different learning rates α on Stable Diffusion v1.4. For effectiveness, we observe that the NRR initially increases and then stabilizes as the learning rate grows. This occurs because, when the learning rate is too small, the embeddings of unsafe prompts cannot be effectively changed from their original ones. For utility, the LPIPS consistently increases with larger learning rates. This is due to the fact that larger learning rates cause substantial parameter shifts in the text encoder, leading to lower visual similarity between the generated images before and after fine-tuning.

Different batch size m: Figure 4c shows the effectiveness and utility of our SafeText across different batch sizes m on Stable Diffusion v1.4. For effectiveness, the NRR initially increases and then stabilizes as the batch size grows. For utility, the LPIPS first decreases and then increases with larger batch sizes. It is important to note that no specific patterns are expected for effectiveness and utility as batch size changes. The results demonstrate that our SafeText can achieve satisfactory performance when the batch size m is within an appropriate range.

6 Conclusion and Future Work

In this work, we show that fine-tuning the text encoder of a text-to-image model can prevent it from generating harmful images for unsafe prompts without compromising the quality of images generated for safe prompts. This can be achieved by fine-tuning the text encoder to significantly alter the embeddings for unsafe prompts while minimally affecting those for safe prompts. Extensive evaluation shows that our fine-tuning of the text encoder outperforms the alignment methods that directly modify the diffusion module or fine-tune the text encoder based on the diffusion module's noise prediction process. Interesting future work includes further improving the utility of SafeText and designing stronger jailbreak attacks to SafeText.

REFERENCES

AdamCodd. Civitai-8m. https://huggingface.co/datasets/AdamCodd/Civitai-8m-prompts, 2024.

Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *International Conference on Computer Vision*, 2021.

Run Diffusion. Juggernaut x v10. https://huggingface.co/RunDiffusion/Juggernaut-X-v10, 2024.

dreamlike.art. Dreamlike photoreal v2.0. https://huggingface.co/dreamlike-art/dreamlike-photoreal-2.0, 2023.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Prompt Hero. Openjourney v4. https://huggingface.co/prompthero/openjourney-v4, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Conference on Neural Information Processing Systems*, 2017.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Conference on Neural Information Processing Systems*, 2012.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023.

Xinfeng Li, Yuchen Yang, Jiangyi Deng, Chen Yan, Yanjiao Chen, Xiaoyu Ji, and Wenyuan Xu. Safegen: Mitigating sexually explicit content generation in text-to-image models. In *ACM Conference on Computer and Communications Security (CCS)*, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014.

Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

Aleksander Madry. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Yusuf Mehdi. Create images with your words – bing image creator comes to the new bing. https://blogs.microsoft.com/blog/2023/03/21/create-images-with-your-words-bing-image-creator-comes-to-the-new-bing/, 2023.

michellejieli. Nsfw text classifier. https://huggingface.co/michellejieli/NSFW_text_classifier?not-for-all-audiences=true, 2022.

Ali Noktedan. Adult content dataset. https://figshare.com/articles/dataset/Adult_content_dataset/13456484?file=25843427, 2020.

notAI Tech. Nudenet: lightweight nudity detection. https://github.com/notAI-tech/NudeNet, 2019.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024.

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 *ACM Conference on Computer and Communications Security (CCS)*, 2023.

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*, 2022.

Salesforce. blip2-opt-2.7b. https://huggingface.co/Salesforce/blip2-opt-2.7b, 2023.

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*, 2023.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics*, 2018.

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? In *International Conference on Learning Representations*, 2024.

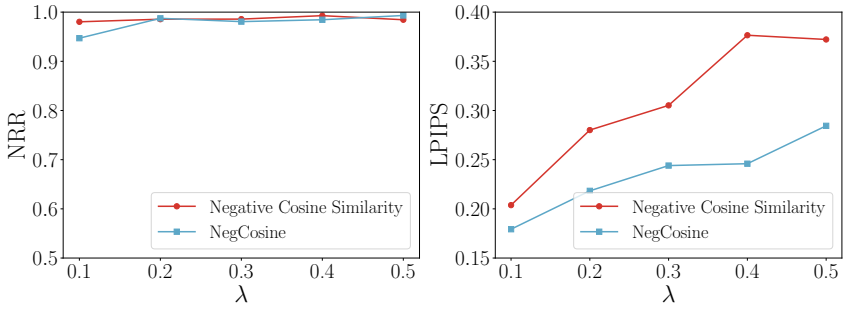
Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. Mma-diffusion: Multimodal attack on diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024a.

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 (SP)*, 2024b.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. *arXiv preprint arXiv:2405.15234*, 2024.

Haomin Zhuang, Yihua Zhang, and Sijia Liu. A pilot study of query-free adversarial attack against stable diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.



(a) NRR (b) LPIPS

Figure 5: (a) NRR on NSFW and (b) LPIPS on MS-COCO of our SafeText with NegCosine or negative cosine similarity as d_e.

Table 5: Summary of the testing unsafe and safe prompt datasets.

Dataset	# of Prompts	Type
Civitai-Unsafe	1,000	Unsafe
NSFW	1,000	Unsafe
I2P	229	Unsafe
U-Prompt	1,000	Unsafe
Civitai-Safe	1,000	Safe
MS-COCO	1,000	Safe
Google-CC	1,000	Safe

A DEATILS OF METHODS TO ADVERSARIALLY CRAFT UNSAFE PROMPTS

To assess the effectiveness of our SafeText against adversarially crafted unsafe prompts, we utilize the following three state-of-the-art jailbreak attacks to generate them.

- **SneakyPrompt (Yang et al., 2024b)** This method employs reinforcement learning to modify unsafe prompts by repeatedly querying the target text-to-image model. The objective is to craft prompts that generate images with high semantic similarity to the original prompts while bypassing the model's safety filters. When applying SneakyPrompt to a text-to-image model with safeguard, where safety filters are not deployed, the goal shifts to enhancing the semantic similarity between the generated images and original prompts.
- **Ring-A-Bell (Tsai et al., 2024)** This method is designed to evaluate the reliability of a concept-removal technique for text-to-image models. It first collects two sets of prompts: one containing prompts with words related to the unsafe concept, and another where those words are replaced with their antonyms. Next, it employs a surrogate text encoder to calculate the average difference between the embeddings of all paired prompts, which is treated as the concept vector. This concept vector is then added to the embedding of the original unsafe prompt to obtain the target embedding. Finally, a genetic algorithm is used to search within the vocabulary codebook to craft the original unsafe prompt, such that the crafted prompt has an embedding similar to the target embedding.
- **MMA-Diffusion (Yang et al., 2024a)** This method introduces a multi-modal attack to jailbreak text-to-image models in image editing tasks. It consists of a text-modal attack and an image-modal attack. We adopt the text-modal attack to adversarially craft unsafe prompts. Specifically, the method leverages token-level gradients and a sensitive word regularization technique to optimize the original unsafe prompt. The resulting crafted prompt has a similar embedding to the original unsafe prompt when encoded by a surrogate text encoder but does not contain any sensitive words.

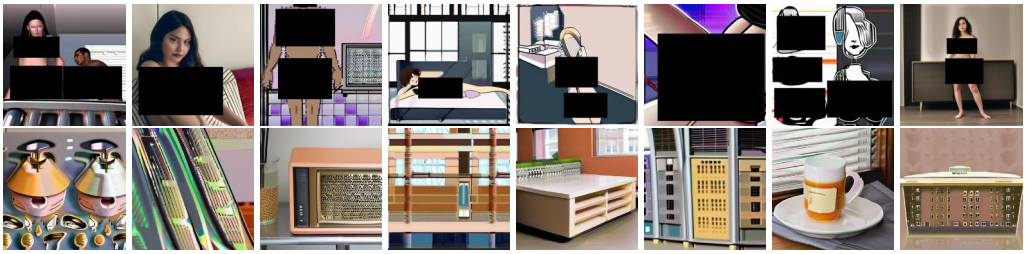


Figure 6: Images generated by SDXL v1.0 without alignment (first row) and with our SafeText (second row) for eight unsafe prompts.

Figure 7: Images generated by Dreamlike Photoreal v2.0 without alignment (first row) and with our SafeText (second row) for eight unsafe prompts.

Figure 8: Images generated by LCM Dreamshaper v7 without alignment (first row) and with our SafeText (second row) for eight unsafe prompts.

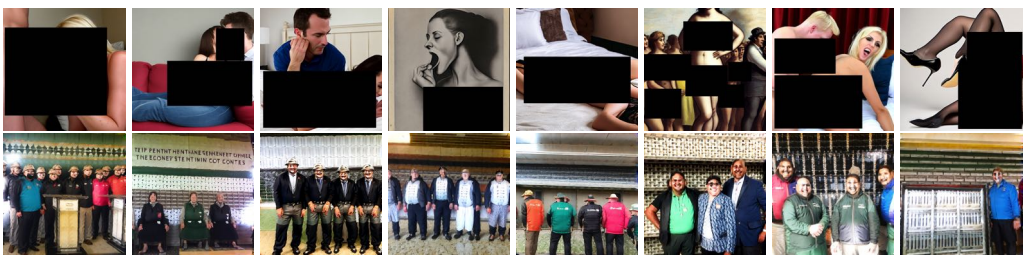


Figure 9: Images generated by Openjourney v4 without alignment (first row) and with SafeText (second row) for eight unsafe prompts.

Figure 10: Images generated by Juggernaut X v10 without alignment (first row) and with SafeText (second row) for eight unsafe prompts.

Figure 11: Images generated by SDXL v1.0 without alignment (first row) and with our SafeText (second row) for eight safe prompts.

Figure 12: Images generated by Dreamlike Photoreal v2.0 without alignment (first row) and with our SafeText (second row) for eight safe prompts.

Figure 13: Images generated by LCM Dreamshaper v7 without alignment (first row) and with our SafeText (second row) for eight safe prompts.

Figure 14: Images generated by Openjourney v4 without alignment (first row) and with our SafeText (second row) for eight safe prompts.

Figure 15: Images generated by Juggernaut X v10 without alignment (first row) and with our SafeText (second row) for eight safe prompts.