

785 A Proof

786 **Theorem 3.2.** Suppose that $\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t]$, $\mathbb{E}_{\text{safe}}[\mathbf{x}|\mathbf{x}_t]$, and $\mathbb{E}_{\text{unsafe}}[\mathbf{x}|\mathbf{x}_t]$ are the data denoiser, the safe
787 denoiser, and the unsafe denoiser. Then,

$$\begin{aligned}\mathbb{E}_{\text{safe}}[\mathbf{x}|\mathbf{x}_t] &= \mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] \\ &\quad + \beta^*(\mathbf{x}_t)(\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] - \mathbb{E}_{\text{unsafe}}[\mathbf{x}|\mathbf{x}_t])\end{aligned}$$

788 for a weight is defined by

$$\beta^*(\mathbf{x}_t) = \frac{Z_{\text{unsafe}}p_{\text{unsafe},t}(\mathbf{x}_t)}{Z_{\text{safe}}p_{\text{safe},t}(\mathbf{x}_t)},$$

789 where $Z_{\text{safe}} := \int 1_{\text{safe}}(\mathbf{x})p_{\text{data}}(\mathbf{x}) d\mathbf{x}$ and $Z_{\text{unsafe}} := \int 1_{\text{unsafe}}(\mathbf{x})p_{\text{data}}(\mathbf{x}) d\mathbf{x}$ are normalizing constants
790 of safe and unsafe distributions, respectively.

791 *Proof.* Using the relationships

$$p_{\text{safe}}(\mathbf{x}) = \frac{1}{Z_{\text{safe}}}1_{\text{safe}}(\mathbf{x})p_{\text{world}}(\mathbf{x}) \text{ and } p_{\text{unsafe}}(\mathbf{x}) = \frac{1}{Z_{\text{unsafe}}}1_{\text{unsafe}}(\mathbf{x})p_{\text{world}}(\mathbf{x}),$$

792 we derive the safe denoiser by

$$\begin{aligned}\mathbb{E}_{\text{safe}}[\mathbf{x}|\mathbf{x}_t] &= \int \mathbf{x}p_{\text{safe},t}(\mathbf{x}|\mathbf{x}_t) d\mathbf{x} \\ &= \frac{\int \mathbf{x}p_{\text{safe}}(\mathbf{x})q_t(\mathbf{x}_t|\mathbf{x}) d\mathbf{x}}{p_{\text{safe},t}(\mathbf{x}_t)} \\ &= \frac{\int \mathbf{x}1_{\text{safe}}(\mathbf{x})p_{\text{data}}(\mathbf{x})q_t(\mathbf{x}_t|\mathbf{x}) d\mathbf{x}}{Z_{\text{safe}}p_{\text{safe},t}(\mathbf{x}_t)} \\ &= \frac{\int \mathbf{x}(1(\mathbf{x}) - (1(\mathbf{x}) - 1_{\text{safe}}(\mathbf{x})))p_{\text{data}}(\mathbf{x})q_t(\mathbf{x}_t|\mathbf{x}) d\mathbf{x}}{Z_{\text{safe}}p_{\text{safe},t}(\mathbf{x}_t)} \\ &= \frac{\int \mathbf{x}(1(\mathbf{x}) - 1_{\text{unsafe}}(\mathbf{x}))p_{\text{data}}(\mathbf{x})q_t(\mathbf{x}_t|\mathbf{x}) d\mathbf{x}}{Z_{\text{safe}}p_{\text{safe},t}(\mathbf{x}_t)} \\ &= \frac{\int \mathbf{x}p_{\text{data}}(\mathbf{x})q_t(\mathbf{x}_t|\mathbf{x}) d\mathbf{x} - \int \mathbf{x}1_{\text{unsafe}}(\mathbf{x})p_{\text{data}}(\mathbf{x})q_t(\mathbf{x}_t|\mathbf{x}) d\mathbf{x}}{Z_{\text{safe}}p_{\text{safe},t}(\mathbf{x}_t)} \\ &= \frac{\int \mathbf{x}p_{\text{data}}(\mathbf{x})q_t(\mathbf{x}_t|\mathbf{x}) d\mathbf{x} - Z_{\text{unsafe}} \int \mathbf{x}p_{\text{unsafe}}(\mathbf{x})q_t(\mathbf{x}_t|\mathbf{x}) d\mathbf{x}}{Z_{\text{safe}}p_{\text{safe},t}(\mathbf{x}_t)} \\ &= \frac{p_{\text{data},t}(\mathbf{x}_t)}{Z_{\text{safe}}p_{\text{safe},t}(\mathbf{x}_t)} \frac{\int \mathbf{x}p_{\text{data}}(\mathbf{x})q_t(\mathbf{x}_t|\mathbf{x}) d\mathbf{x}}{p_{\text{data},t}(\mathbf{x}_t)} - \frac{Z_{\text{unsafe}}p_{\text{unsafe},t}(\mathbf{x}_t)}{Z_{\text{safe}}p_{\text{safe},t}(\mathbf{x}_t)} \frac{\int \mathbf{x}p_{\text{unsafe}}(\mathbf{x})q_t(\mathbf{x}_t|\mathbf{x}) d\mathbf{x}}{p_{\text{unsafe},t}(\mathbf{x}_t)} \\ &= \frac{p_{\text{data},t}(\mathbf{x}_t)}{Z_{\text{safe}}p_{\text{safe},t}(\mathbf{x}_t)} \mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] - \frac{Z_{\text{unsafe}}p_{\text{unsafe},t}(\mathbf{x}_t)}{Z_{\text{safe}}p_{\text{safe},t}(\mathbf{x}_t)} \mathbb{E}_{\text{unsafe}}[\mathbf{x}|\mathbf{x}_t].\end{aligned}$$

793 Now,

$$\begin{aligned}1 + \frac{Z_{\text{unsafe}}p_{\text{unsafe},t}(\mathbf{x}_t)}{Z_{\text{safe}}p_{\text{safe},t}(\mathbf{x}_t)} &= \frac{Z_{\text{safe}}p_{\text{safe},t}(\mathbf{x}_t) + Z_{\text{unsafe}}p_{\text{unsafe},t}(\mathbf{x}_t)}{Z_{\text{safe}}p_{\text{safe},t}(\mathbf{x}_t)} \\ &= \frac{Z_{\text{safe}} \int p_{\text{safe}}(\mathbf{x})q_t(\mathbf{x}_t|\mathbf{x}) d\mathbf{x} + Z_{\text{unsafe}} \int p_{\text{unsafe}}(\mathbf{x})q_t(\mathbf{x}_t|\mathbf{x}) d\mathbf{x}}{Z_{\text{safe}}p_{\text{safe},t}(\mathbf{x}_t)} \\ &= \frac{\int (Z_{\text{safe}}p_{\text{safe}}(\mathbf{x}) + Z_{\text{unsafe}}p_{\text{unsafe}}(\mathbf{x}))q_t(\mathbf{x}_t|\mathbf{x}) d\mathbf{x}}{Z_{\text{safe}}p_{\text{safe},t}(\mathbf{x}_t)} \\ &= \frac{\int (1_{\text{safe}}(\mathbf{x})p_{\text{data}}(\mathbf{x}) + 1_{\text{unsafe}}(\mathbf{x})p_{\text{data}}(\mathbf{x}))q_t(\mathbf{x}_t|\mathbf{x}) d\mathbf{x}}{Z_{\text{safe}}p_{\text{safe},t}(\mathbf{x}_t)} \\ &= \frac{\int p_{\text{data}}(\mathbf{x})q_t(\mathbf{x}_t|\mathbf{x}) d\mathbf{x}}{Z_{\text{safe}}p_{\text{safe},t}(\mathbf{x}_t)} = \frac{p_{\text{data},t}(\mathbf{x}_t)}{Z_{\text{safe}}p_{\text{safe},t}(\mathbf{x}_t)},\end{aligned}$$

794 which completes the proof. \square

B Limitations and Broader Impacts

Limitations We have addressed significant safety challenges in DMs, particularly concerning the generation of NSFW content and the inadvertent reproduction of sensitive data. We introduce the *safe denoiser*, a novel approach that modifies the sampling trajectories of DMs to adhere to theoretically safe distributions, thereby ensuring the generation of appropriate and authorized content.

However, this approach necessitates the introduction of an additional hyperparameter, β_t , as outlined in Theorem 3.2. While we demonstrate that this parameter is theoretically derived and straightforward to implement, it may not be optimal for realistic scenarios due to its assumption of access to numerous data points sampled from an unsafe distribution. In practice, we present evidence in Figure 4b that this parameter influences the performance of the model.

Despite the challenges, we have developed a novel training-free method that effectively guides the sampling trajectories of DMs towards safe distributions. Ultimately, this work provides a robust and scalable solution for mitigating safety risks in generative AI, paving a way for their responsible and ethical applications.

Broader Impacts This paper presents a work whose goal is to build a reliable and trustworthy Generative AI. There are many potential societal consequences of our work, particularly in addressing ethical risks associated with generative models. Our research is focused on preventing the generation of NSFW content, including nudity and violence, and mitigating the risk of models memorizing and reproducing private information, such as human face, from training datasets. We believe the presented work contributes to the responsible use of generative AI, reinforcing ethical safeguards and promoting AI systems that align with societal values and human rights.

C Experimental Details : Text-to-Image Generation

As outlined in the manuscript, we conduct the Text-to-Image experiment using SD-v1.4, following the same model as the baselines for generating images from text, as referenced in [2, 51, 12, 1]. To ensure consistency, we adopt the generation procedure described in each baseline. Preliminary observing the sensitivity of nudity-related content, we employ the DDPM scheduler [17]. For a fair comparison, we maintain the same number of inference steps, specifically 50, aligning with the official implementations of both SLD and SAFREE, which also use 50 inference steps.

Regarding the *Safe Denoiser*, the proposed model computes the transition kernel with an RBF kernel. The RBF kernel function is defined as follows:

$$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) \quad (\text{C.1})$$

For the bandwidth parameter σ , we set a value of 1.0 for SLD and 3.15 for SAFREE. Additionally, in case of SAFREE, we apply a scaling factor $\eta = 0.33$, whereas for SLD, we use $\eta = 0.03$ to regulate the strength of the repency in Eq. (7). Empirically, we introduce a heuristic in which the proposed *Safe Denoiser* is applied within critical timesteps $C = [780, \dots, 1000]$. In the early stages of diffusion, denoising process primarily establishes global structures rather than intricate details, while the later stages focus on refining fine-grained features. Since our approach aims to prevent the generation of globally harmful images rather than enhancing image quality or detail, we apply the denoiser at these later timesteps.

For reference images, we provide a detailed explanation of how they are obtained. To ensure safe generation against nudity prompts, we utilize a total of 515 images sourced from the I2P dataset [2]. These images were generated using SD-v1.4. As mentioned in the manuscript, these reference images meet the criterion, where a nude class probability exceeds 0.6, as determined by Nudenet. Sample images are presented in Figure C.1. On the other hand, for the inappropriate probability task with the CoPro dataset, we attempt to use the total images from the I2P dataset. However, our computational resources allow us to use only 3,000 reference images. To select these 3,000 images, we randomly choose them out of the 4,703 images available in the I2P dataset. All images used in this task are also generated using SD-v1.4. Sample images are presented in Figure C.2. It’s important to note that all experiments conducted in this study use the same set of reference images across all baselines. This ensures a fair comparison.

844 Additionally, the reference images we used in the task to generate safer images against nudity prompts
 845 are included as attachments in our supplementary materials. On the other hand, due to space
 846 constraints, we cannot include the attachment in the inappropriate probability task. We ensure that
 847 the reference images used in this task will be included in the public repository upon the acceptance of
 848 the paper.

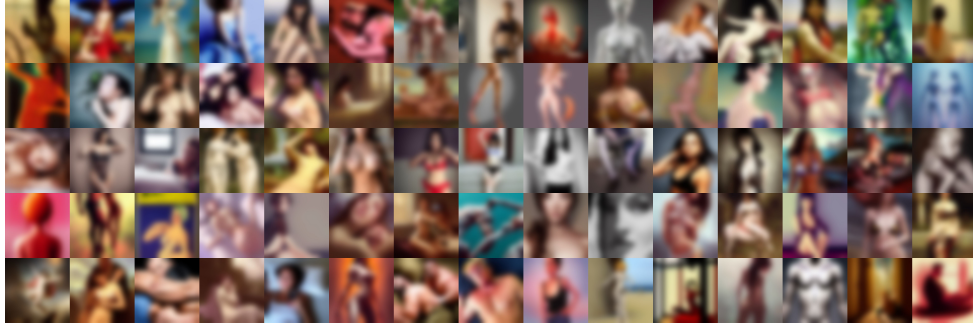


Figure C.1: Reference images for safe generation against nudity prompts

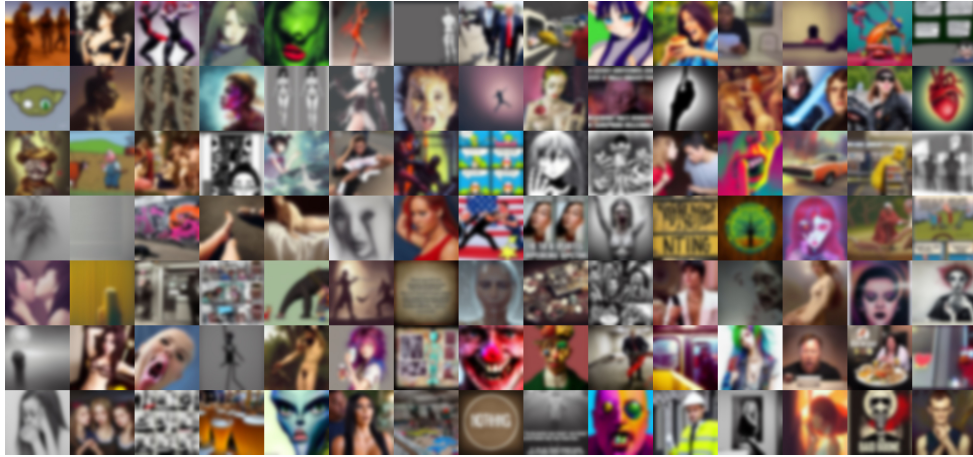


Figure C.2: Reference images for inappropriate probability

849 Next, we briefly introduce the baseline models used in our experiments. The first two approaches
 850 serve as comparisons for unlearning-based safe diffusion models [9, 12]. Specifically, we evaluate
 851 Erased Stable Diffusion (ESD) [9] as a representative method. More recently, reliably trained safe
 852 diffusion (RECE) models have demonstrated improved performance, particularly in reducing the
 853 attack success rate [12]. In addition to these unlearning-based approaches, we also include SLD and
 854 SAFREE as training-free safe diffusion models [2, 1]. While both methods employ negative prompts,
 855 their underlying mechanisms differ significantly. In SLD, the set of unsafe prompts, denoted as c_{US} ,
 856 is designed to mitigate globally harmful image generation [2]. In contrast, SAFREE focuses on more
 857 precise negative prompts specifically tailored to nudity-related content [1]. Beyond negative prompts,
 858 SAFREE further enhances safety by applying an orthogonal projection technique in Euclidean space
 859 to shift text embeddings away from predefined toxic regions. In the following, we provide an overview
 860 of the datasets used in our experiments.

861 C.1 Inappropriate Prompt Datasets

862 **I2P** The I2P dataset consists of prompts related to seven unsafe concepts: hate, harassment, violence,
 863 self-harm, sexual content, shocking content, and illegal activity [2]. It contains a total of 4,703
 864 prompts and was introduced in earlier stages of research, with subsequent studies primarily focusing
 865 on this dataset [12, 1]. In this work, we utilize the I2P dataset as a source of reference data points
 866 rather than for additional training. The dataset was obtained from <https://huggingface.co/datasets/AI-MIL-TUDA/i2p>
 867

CoPro Compared to I2P [2], the CoPro dataset offers a more extensive dataset comprising a total of 226,104 prompts, each associated with 723 concepts that span both safe and unsafe scenarios. This expansion enhances the dataset’s suitability for rigorous evaluation [42]. Particularly, it also offers super-concept information, following the same framework of I2P [2]. All text prompts are categorized into {hate, harassment, violence, self-harm, sexual content, shocking content, and illegal activities}. This ensures that they align with the corresponding category information in the I2P dataset. To efficiently evaluate models, we randomly sample 10,000 prompts, ensuring a uniform distribution across all categories. We validated that the average inappropriate probability of SD-v1.4 in the randomly sampled dataset, presented in Table 3, closely matches the numerical information provided in [52]. In this work, we evaluate safe image generation performance across baselines on the CoPro dataset using reference data points from the I2P dataset. This dataset was obtained from https://github.com/rt219/LatentGuard/blob/main/dataset/CoPro_v1.0.json

C.2 Nudity in NSFW Prompt Datasets

Ring-A-Bell The Ring-A-Bell dataset was developed through a red-teaming approach that evaluates text-to-image diffusion models using black-box methods [37]. The original dataset Chia15/RingABell-Nudity contains 285 prompts; however, we use a curated subset of 79 prompts, following prior baselines [12, 1]. This selection ensures a more equitable comparison of our method. The curated Ring-A-Bell dataset was obtained from either <https://github.com/CharlesGong12/RECE> or <https://github.com/jaehong31/SAFE>.

MMA-Diffusion MMA-Diffusion is another dataset generated via a red-teaming approach [26]. Unlike other datasets, it consists of adversarial prompts designed to include potentially harmful contexts without explicit expressions. Similar to the Ring-A-Bell dataset, we use a curated set of 1,000 prompts, consistent with prior baselines [12, 1]. The dataset was obtained from <https://github.com/CharlesGong12/RECE> or <https://github.com/jaehong31/SAFE>.

UnlearnDiff The UnlearnDiff dataset contains various harmful text prompts that can potentially generate NSFW images [36]. Among its categories, we specifically focus on nudity-related prompts. The dataset includes a total of 116 nudity-related prompts, derived from an initial set of 143 prompts, from which 27 were excluded as they contained other NSFW categories such as self-harm and shocking content. This selection ensures that our numerical metrics remain unaffected by unrelated factors. The dataset was obtained from <https://github.com/CharlesGong12/RECE> or <https://github.com/jaehong31/SAFE>.

C.3 Ann Graham Lotz for Data Memorization

In Figure 1b, we demonstrate that SD-v1.4 exhibits training dataset memorization, as it is capable of regenerating an identical images using the text prompt, (*‘Living in the light with Ann Graham Lotz <|startoftext|> lad mans’*). In this example, our method is applied with a bandwidth $\sigma = 13.15$ and scaling factor of 0.69. To construct a reference data for this case, we collected a total of 10 images from the internet. These are presented in Figure C.3



Figure C.3: Reference images for Ann Graham Lotz case

D Experimental Details : Class-Conditional and Unconditional Generation

In this section, we use our safe denoiser in the DMs without text inputs. Specifically, we employ experiments on FFHQ [49] and ImageNet [48] in the 256×256 resolution. We utilize the pretrained diffusion models from FFHQ [53]⁶ and ImageNet [7]⁷. For the experiments, we use a DDPM solver [54] with 100 steps.

Unconditional Generation For unconditional generation, we utilize the FFHQ dataset to evaluate whether the proposed method effectively mitigates sexual bias, using our method. Although FFHQ dataset does not include explicit label information, Table 6 illustrates that the generated images exhibit a noticeable bias toward female images over male ones. In this experiment, we select 1K female images from CelebA-HQ [55]⁸ validation split to serve as unseen negative data, thereby establishing the negative dataset $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(1000)}\}$. Then, we employ our safe denoiser to generate 1K images. While both FFHQ and CelebA-HQ are designed to capture similar distribution, they are not completely aligned. This distinction provides an advantageous experimental setup, where we assess the controllability of image generation using reference images. For performance evaluation, we compute FID [43] score using 1,000 male images from the CelebA-HQ dataset. For classification performance, we train a ResNet18 model, as implemented in the PyTorch framework⁹ using the training dataset in the CelebA-HQ. In this experiment, we chose $\sigma = 1.0$ and $\eta = 0.05$, and employ *Safe denoiser* across the entire denoising timesteps.

Conditional Generation For conditional ImageNet [48] experiments at 256×256 resolution, we use a diffusion model trained on the full ImageNet-256 dataset guided by a classifier [7]. The diffusion backbone follows a linear noise schedule and is constructed with 1,000 diffusion time-steps. We condition on class labels by scaling the classifier guidance at 5.0, creating a strong pull towards the desired class during the sampling process. Each experiment generates 50 samples per class across all 1000 ImageNet classes, producing 50,000 samples that are then evaluated with a pretrained ImageNet classifier for precision, recall, and classification accuracy measurements [56]. Our metrics include (i) **Precision**: the fraction of generated samples that match the designated ImageNet label when conditioned on the class, (ii) **Recall**: aims to evaluate the diversity and coverage of the targeted class distribution, and (iii) **Classification Accuracy**: the rate at which generated images are correctly identified as their conditioned label among the 999 classes (excluding the negated target class, i.e., Chihuahua). The classification accuracy on the hold-out negated class is also calculated, to evaluate how well the respective method does not generate the negated target class. As illustrated in Table 4, we focus on the Chihuahua class to investigate how effectively our proposed safe denoiser can repel a target class while preserving generative quality for other classes in this experiment. To avoid unintended Chihuahua generation, aforementioned metrics aim to make sure that samples do not drift toward distinct Chihuahua-like features. For instance, when we generate an image based on a reference dataset sampled from a Chihuahua, the resulting sample may resemble a Golden Retriever, but it won't resemble a Chihuahua.

To compare our approach, we implement three variants of the conditional diffusion process: vanilla classifier-guided diffusion model without repellency mechanisms, the Sparse Repellency (SR) [38] technique applied to the classifier-guided diffusion model, and our safe denoiser technique applied to the same diffusion process. For the reference dataset, we select the validation set of Chihuahua class as the negative images. In this experiment, the safe denoiser technique is applied on the 200 to 800 timesteps of the diffusion process. $\eta = 0.02$ was chosen as to control the strength of the repellency away from the Chihuahua target class. In the SR variant of the experiment, a repellency scale of 0.01 is combined with a large radius of 300 to push generated samples out of regions resembling the negated target class.

⁶<https://github.com/DPS2022/diffusion-posterior-sampling>

⁷<https://github.com/openai/guided-diffusion>

⁸<https://www.kaggle.com/datasets/badasstechie/celebahq-resized-256x256>

⁹<https://pytorch.org/vision/stable/index.html>

E Additional Experimental Results

In this section, we share extra experimental results. Both numeric and visual results are included, which are not presented in the main text. These results highlight the empirical gains in terms of safe generation and the preservation of the global context of the generated samples simultaneously. Specifically, this ensures that the samples remain faithful to their original meanings while enabling us to negate specific concepts we intended.

E.1 Quantitative Results

ImageNet Case for Negating Chihuahua Class In ImageNet, we focus on negating a specific Chihuahua class during generation. We select the validation set of Chihuahua class as the negative images. We generate 50 samples per class and classify samples from 999 classes by a classifier [7] and report the accuracy by Top-1. Also, we measure the Top-1 accuracy of 50 samples from Chihuahua class, reporting it by Top-1* in Table E.1. From the result, we note that our method excels generating other 999 classes, while SR cannot generate images from those 999 classes. To evaluate the overall quality, Table E.1 further report the precision (sample accuracy) and recall (sample diversity) [50] over 50K samples, indicating that our method is better than SR in negating a specific class.

Table E.1: Experiments on ImageNet for the specific class (Chihuahua) negation task. Top-1 is the classification accuracy of the generated samples on 999 classes, and Top-1* indicates the accuracy on the specific class.

Method	Prec \uparrow	Rec \uparrow	Top-1 \uparrow	Top-1* \downarrow
Baseline (B)	0.72	0.63	0.76	0.68
B + SR	0.59	0.54	0.01	0.0
B + Ours	0.62	0.58	0.14	0.0

Aesthetic Scores for Long Text Prompts We identified that our method, which incorporates negative prompts, effectively reduces the risk of generating unsafe data and maintains alignment with the given text prompts. However, the text prompts in these cases span various lengths. Therefore, it is necessary to quantify whether our method excessively applies to remove unsafe contents, leading to unfavorable images in extreme cases. We sample the most complex cases from the I2P datasets and compare the generated images across baselines.

Table E.2: Aesthetic scores for long text prompts in the I2P dataset.

Method	LAION-aesthetic V2 \uparrow
SD-v1.4	5.97 ± 0.534
SAFREE	6.03 ± 0.540
SAFREE+Ours	5.94 ± 0.529

To test how our method works when long and complex prompts are given, we use LAION-aesthetic V2 score¹⁰ as a metric and use top 10% longest prompts (475 prompts, avg. word_count=54) selected from the I2P dataset. We choose this score as it is known to be correlated with human perception of quality of images (higher the better). As shown Table E.2, our method maintains aesthetic quality comparable to baselines, even with complex prompts. We prove that the proposed method does not struggle to create appropriate images even when asked with long text prompts.

E.2 Qualitative Results

We present additional qualitative results across three experimental scenarios: (1) *Text-to-Image Generation for preventing nudity and inappropriate images*, (2) *Sexual Debiasing in unconditional generation for facial images*, and (3) *Class-Conditional Generation, where reference images serve as constraints not to generate*. To systematically demonstrate the effectiveness of our approach, we

¹⁰<https://github.com/christophschuhmann/improved-aesthetic-predictor>

present the results in sequence, beginning with text-to-image generation followed by unconditional generation and concluding with conditional generation. To facilitate straightforward understanding, we include as many figures and qualitative comparisons as possible.

E.3 Text-to-Image Generation

Safe Generation against Nudity Prompts We present a qualitative comparison across baselines and ours. All figures are generated using the same text prompts. We decide to exclude the case of MMA-Diffusion since prompts in this dataset generate pornographic images by baselines, which is not suitable for academic purpose. Hence, we select text prompts from Ring-A-Bell [37] and UnlearnDiff [36]. From Figure E.4 to Figure E.5 we observe that our model effectively eliminates nudity information while preserving textual information.



Figure E.4: Generated images by baselines and ours on Ring-A-Bell [37]



Figure E.5: Generated images by baselines and ours on UnlearnDiff [36]

Inappropriate Probability in CoPro Dataset In our evaluation on the CoPro dataset [42], we apply our method with images from the I2P dataset [2], which includes a wide range of sensitive categories: {hate, harassment, violence, self-harm, sexual content, shocking content, and illegal activities}. Among these, we focus on the 'Self-Harm' category. Self-harm content is suitable for graphical illustration, distinguishing it an appropriate and interpretable case for homogenous visual inspection in the public domain. Unlike other categories—where perceptions of appropriateness can vary widely across cultural and personal contexts—'Self-harm' is typically associated with somber or distressing imagery that is broadly and publicly recognized as unsafe.

As illustrated in Figure E.6, our Safe Denoiser effectively reduces the generation of implicit unsafe content while preserving coherence with the provided prompts. This underscores its ability to both detect and suppress sensitive content without compromising the semantic alignment of textual prompts. From this figure, it is evident that negative prompts do not yield significant results in mitigating the generation of sad and gloomy atmospheres, particularly for conveying feelings of collapse. Conversely, our *Safe Denoiser* tends to generate images that more accurately reflect the literal meanings of the textual prompts. This tendency contributes to a reduction in the likelihood of generating content that evokes feelings of 'Self-harm'.

Alignment of Textual Prompts We present uncensored generated images from the CoCo dataset. This dataset encompasses a wide range of textual prompts that cover various lengthy and diverse



Figure E.6: Generated images by baselines and ours on CoPro [42]. All textual prompts are labeled as 'Self-Harm'. This dataset also provides safe alternatives, and we present both.

1011 situations. As shown in Figure E.7 and Table 2 we conclude that our approach does not compromise
 1012 the performance of generating normal images. Instead, it focuses on addressing the challenge of
 1013 generating potentially unsafe images.



Figure E.7: Uncurated generated images by SAFREE+Ours on CoCo30K

1014 E.4 Unconditional Generation: Sexual Debiasing

1015 We present uncured generated images created by the DM trained on the FFHQ dataset [49]. This
 1016 dataset lacks explicit labels indicating sexual information. However, we observe a tendency for this
 1017 model to generate female images more frequently than male images, as shown in Table 6. On the
 1018 other hand, when we utilize *Safe denoiser* with female images, we mitigate the potential bias towards
 1019 female images and achieves generating images uniformly distributed across sexual information.
 1020 Figure E.8 illustrates that the generated images align with the numerical results presented in Table 6



Figure E.8: Comparison of *Safe Denoiser* against existing approaches when negation on female.

1021 E.5 Class-conditional Generation : Negation of Specific Class

1022 We present uncured images that focus on negating a specific Chihuahua class. Here are two
 1023 experimental setups. First, we employ class conditional guidance on the ‘Chihuahua’ class and
 1024 simultaneously use the *Safe denoiser* to work with negative images sampled from the ‘Chihuahua’
 1025 class in the validation split. We observe that the SR does not follow homogeneous images that
 1026 align with the superclass, ‘Dog’, but our method produces similar small dogs but not matched with
 1027 ‘Chihuahua’ as shown in Figure E.9

1028 Second, we qualitatively evaluate that our method with negative images from the ‘Chihuahua’ class
 1029 works when class guidance is applied to classes other than ‘Dogs’, for example, ‘Tench’ and ‘Truck’.
 1030 This result is shown in Figure E.10. We observe that the SR sometimes depicts different classes
 1031 even when class guidance is applied, but our method aligns with homogeneous classes following
 1032 class guidance even when the *Safe denoiser* works with ‘Chihuahua’ images. This indicates that our
 1033 method effectively tackles specific concepts and preserves the original performance when it is not
 1034 mutually correlated.



Figure E.9: Generated samples when negating the Chihuahua class, primarily producing visually similar small dog breeds.



Figure E.10: Comparison of *Safe Denoiser* against existing approaches when negating on Chihuahua. This comparison includes non-dog related ImageNet classes, which include Tench, Garbage Truck, Church, Spoonbill, and Great White Shark.

1035 Additional graphical illustrations are presented in the following figures from Figure [E.11](#) to Fig-
 1036 ure [E.13](#).

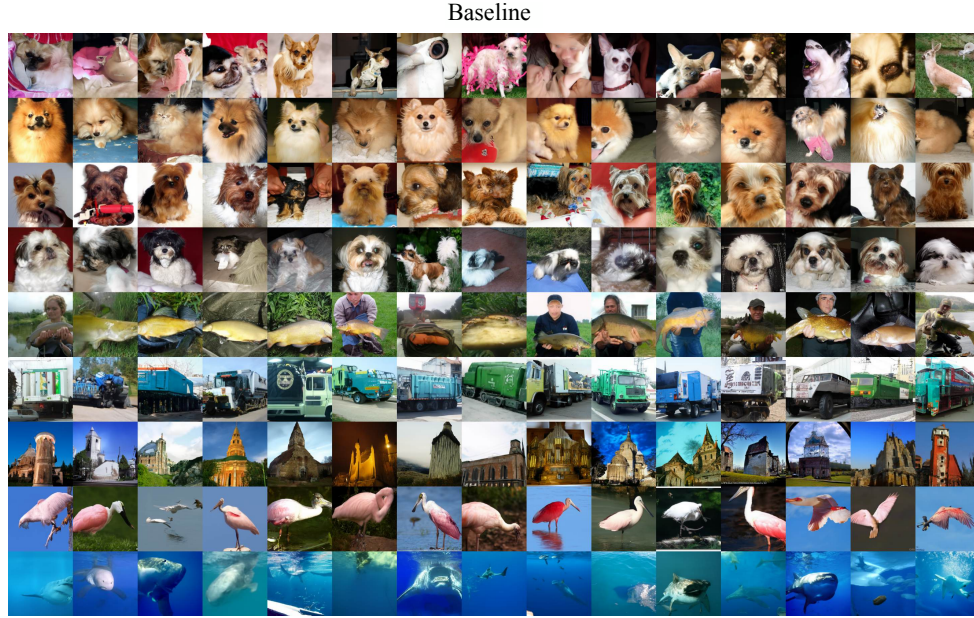


Figure E.11: Classifier guidance diffusion model generated samples when negating on Chihuahua. This comparison includes non-dog-related ImageNet classes mentioned in [E.10](#) along with the dog-related classes in Figure [E.9](#) which are Pomeranian, Yorkshire Terrier, and Shih Tzu.

Sparse Repellency

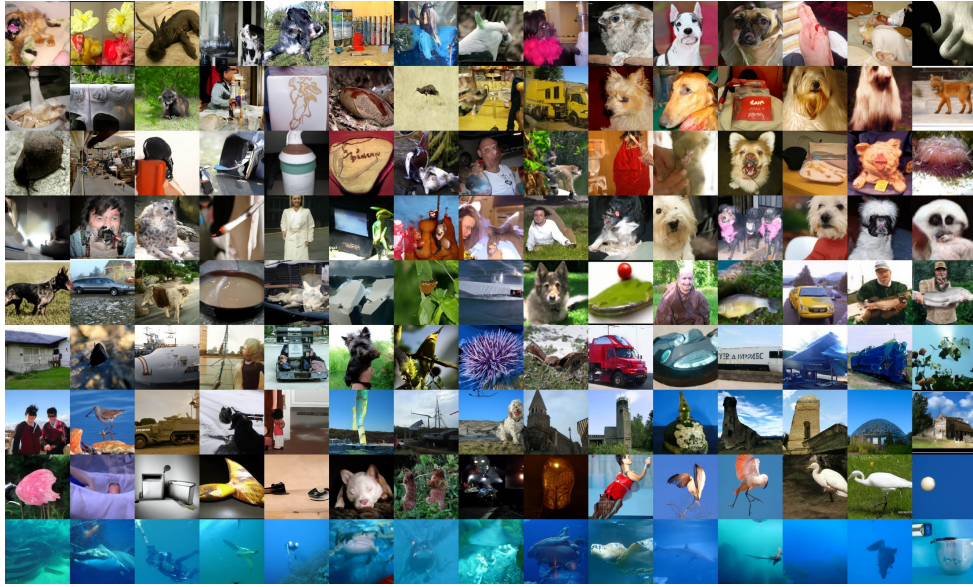


Figure E.12: *Sparse Repellency* generated samples when negating on Chihuahua. The same classes are selected as [E.11](#).

Ours

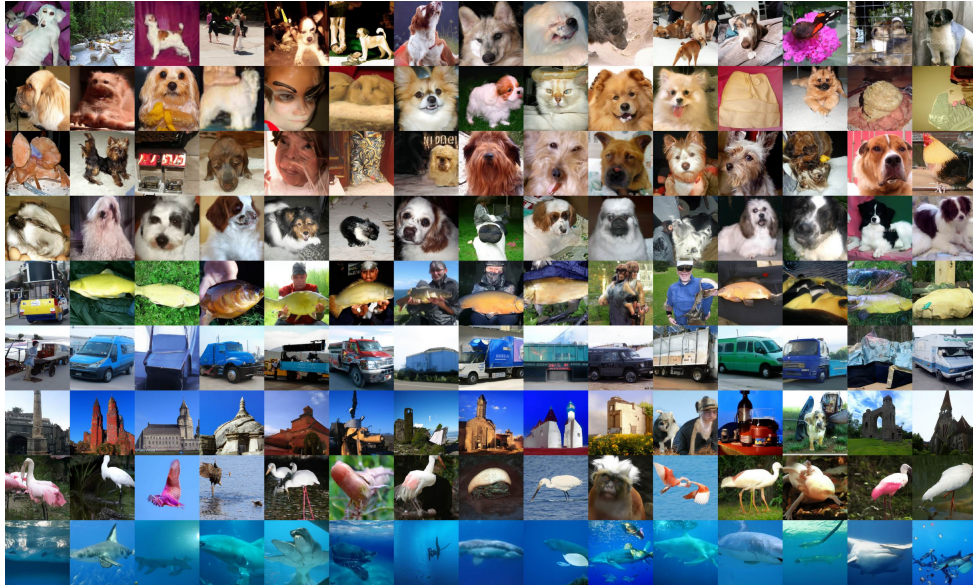


Figure E.13: *Safe Denoiser* generated samples when negating on Chihuahua. The same classes are selected as [E.11](#).