Safe Latent Diffusion:
Mitigating Inappropriate Degeneration in Diffusion Models

Patrick Schramowski[1,2,3,6*] Manuel Brack[3*] Björn Deiseroth[2,3,5] Kristian Kersting[1,2,3,4]

[1]DFKI, [2]Hessian.AI, [3]Computer Science Department, TU Darmstadt

[4]Centre for Cognitive Science, TU Darmstadt, [5]Aleph Alpha, [6]LAION

{schramowski, brack, deiseroth, kersting}@cs.tu-darmstadt.de

Abstract

Text-conditioned image generation models have recently achieved astonishing results in image quality and text alignment and are consequently employed in a fast-growing number of applications. Since they are highly data-driven, relying on billion-sized datasets randomly scraped from the internet, they also suffer, as we demonstrate, from degenerated and biased human behavior. In turn, they may even reinforce such biases. To help combat these undesired side effects, we present safe latent diffusion (SLD). Specifically, to measure the inappropriate degeneration due to unfiltered and imbalanced training sets, we establish a novel image generation test bed—inappropriate image prompts (I2P)—containing dedicated, real-world image-to-text prompts covering concepts such as nudity and violence. As our exhaustive empirical evaluation demonstrates, the introduced SLD removes and suppresses inappropriate image parts during the diffusion process, with no additional training required and no adverse effect on overall image quality or text alignment.[1]

Warning: *This paper contains sexually explicit imagery, discussions of pornography, racially-charged terminology, and other content that some readers may find disturbing, distressing, and/or offensive.*

1. Introduction

The primary reasons for recent breakthroughs in text-conditioned generative diffusion models (DM) are the quality of pre-trained backbones' representations and their multimodal training data. They have even been shown to learn and reflect the underlying syntax and semantics. In turn, they retain general knowledge implicitly present in the data [27]. Unfortunately, while they learn to encode and reflect general information, systems trained on large-scale unfiltered data may suffer from degenerated and biased behavior. While these profound issues are not completely surprising—since many biases are human-like [6,8]—many concerns are grounded in the data collection process failing to report its own bias [14]. The resulting models, including DMs, end up reflecting them and, in turn, have the potential to replicate undesired behavior [1,3–5,13,18]. Birhane *et al.* [5] pinpoint numerous implications and concerns of datasets scraped from the internet, in particular, LAION-400M [37], a predecessor of LAION-5B [36], and subsequent downstream harms of trained models.

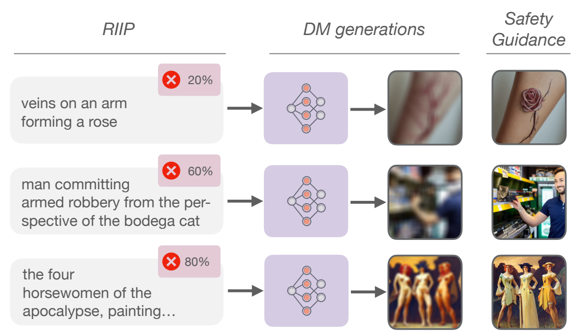

Figure 1. Mitigating inappropriate degeneration in diffusion models. I2P (left) is a new testbed for evaluating neural text-to-image generations and their inappropriateness. Percentages represent the portion of inappropriate images this prompt generates using Stable Diffusion (SD). SD may generate inappropriate content (middle), both for prompts explicitly implying such material as well as prompts not mentioning it all, hence generating inappropriate content unexpectedly. Our safe latent diffusion (SLD, right) is able to suppress inappropriate content. (Best viewed in color)

We analyze the open-source latent diffusion model Stable Diffusion (SD), which is trained on subsets of LAION-5B [36] and find a significant amount of inappropriate content generated which, viewed directly, might be offensive, ignominious, insulting, threatening, or might otherwise cause anxiety. To systematically measure the risk

*Equal contribution

[1]Code available at https://huggingface.co/docs/diffusers/api/pipelines/stable_diffusion_safe

of inappropriate degeneration by pre-trained text-to-image models, we provide a test bed for evaluating inappropriate generations by DMs and stress the need for better safety interventions and data selection processes for pre-training. We release I2P (Sec. 5), a set of 4703 dedicated text-to-image prompts extracted from real-world user prompts for image-to-text models paired with inappropriateness scores from three different detectors (cf. Fig. 1). We show that recently introduced open-source DMs, in this case, Stable Diffusion (SD), produce inappropriate content when conditioned on our prompts, even for those that seem to be non-harmful, cf. Sec. 6. Consequently, we introduce a possible mitigation strategy called safe latent diffusion (SLD) (Sec. 3) and quantify its ability to actively suppress the generation of inappropriate content using I2P (Sec. 6). SLD requires no external classifier, i.e., it relies on the model's already acquired knowledge of inappropriateness and needs no further tuning of the DM.

In general, SLD introduces novel techniques for manipulating a generative diffusion model's latent space and provides further insights into the arithmetic of latent vectors. Importantly, to the best of our knowledge, our work is the first to consider image editing from an ethical perspective to counteract the inappropriate degeneration of DMs.

2. Risks and Promises of Unfiltered Data

Let us start discussing the risks but also promises of noisy, unfiltered and large-scale datasets, including background information on SD and its training data.

Risks. Unfortunately, while modern large-scale models, such as GPT-3 [7], learn to encode and reflect general information, systems trained on large-scale unfiltered data also suffer from degenerated and biased behavior. Nonetheless, computational systems were promised to have the potential to counter human biases and structural inequalities [19]. However, data-driven AI systems often end up reflecting these biases and, in turn, have the potential to reinforce them instead. The associated risks have been broadly discussed and demonstrated in the context of large-scale models [1, 3–5, 13, 18]. These concerns include, for instance, models producing stereotypical and derogatory content [3] and gender and racial biases [10, 24, 38, 41]. Subsequently, approaches have been developed to, e.g., decrease the level of bias in these models [6, 39].

Promises. Besides the performance gains, large-scale models show surprisingly strong abilities to recall factual knowledge from the training data [27]. For example, Roberts *et al.* [30] showed that large-scale pre-trained language models' capabilities to store and retrieve knowledge scale with model size. Grounded on those findings, Schick *et al.* [32] demonstrated that language models can self-debias the text they produce, specifically regarding toxic output. Furthermore, Jenetzsch *et al.* [21] as well as Schramowski *et al.* [35] showed that the retained knowledge of such models carries information about moral norms aligning with the human sense of *"right"* and *"wrong"* expressed in language. Similarly, other research demonstrated how to utilize this knowledge to guide autoregressive language models' text generation to prevent their toxic degeneration [32, 34]. Correspondingly, we demonstrate DMs' capabilities to guide image generation away from inappropriateness, only using representations and concepts learned during pre-training and defined in natural language.

This makes our approach related to other techniques for text-based image editing on diffusion models such as Text2LIVE [2], Imagic [23] or UniTune [40]. Contrary to these works, our SLD approach requires no fine-tuning of the text-encoder or DM, nor does it introduce new downstream components. Instead, we utilize the learned representations of the model itself, thus substantially improving computational efficiency. Previously, Prompt-to-Prompt [15] proposed a text-controlled editing technique using changes to the text prompt and control of the model's cross-attention layers. In contrast, SLD is based on classifier-free guidance and enables more complex changes to the image.

LAION-400M and LAION-5B. Whereas the LAION-400M [37] dataset was released as a proof-of-concept, the creators took the raised concern [5] to heart and annotated potential inappropriate content in its successor dataset of LAION-5B [36]. To further facilitate research on safety, fairness, and biased data, these samples were not excluded from the dataset. Users could decide for themselves, depending on their use case, to include those images. Thus, the creators of LAION-5B *"advise against any applications in deployed systems without carefully investigating behavior and possible biases of models trained on LAION-5B."*

Training Stable Diffusion. Many DMs have reacted to the concerns raised on large-scale training data by either not releasing the model [31], only deploying it in a controlled environment with dedicated guardrails in place [29] or rigorously filtering the training data of the published model [25]. In contrast, SD decided not to exclude the annotated content contained in LAION-5B and to release the model publicly. Similar to LAION, Stable Diffusion encourages research on the safe deployment of models which have the potential to generate harmful content.

Specifically, SD is trained on a subset of LAION-5B, namely LAION-2B-en [36] containing over 2.32 billion English image-text pairs. Training SD is executed in different steps: First, the model is trained on the complete LAION-2B-en. Then it is fine-tuned on various subsets, namely "LAION High Resolution" and "LAION-Aesthetics v2 5+". With all training samples taken from LAION-5B or subsets thereof, it is expected that the trained model reflects not only human-like biases such as gender occupation correlations but also reporting biases. Furthermore, SD is de-

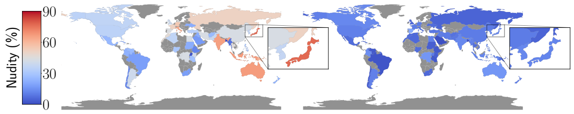

Figure 2. Grounded in reporting bias, one can observe ethnic biases in DMs (left). For 50 selected countries, we generated 100 images with the prompt *'<country> body'*. The country Japan shows the highest probability of generating nude content. SLD uses the strong hyper parameter set to counteract this bias (right). (Best viewed in color)

ployed on several platforms including huggingface[2] and recently lexica.art[3] making it easy to use for the general public, including users unaware of present issues.

Ethnic Bias. This leads us to our first experiment. Following up the studies by Birhane *et al.* on unfiltered multimodal datasets [5], we extend these investigations by answering whether the same issues occur in downstream generative models. Specifically, we chose to showcase biased representation of Asian women [42], cf. also Sec. 2.1 and Appendix A in [5] wrt. LAION-400M. This social phenomenon is a well-documented trend in western countries [42] that is also reflected in the English-speaking internet [20] and subsequently the web-crawled LAION-2B-en. Our search for the 100 closest images (in CLIP [28] space) in the dataset to the term 'japanese body' yielded almost exclusively inappropriate material, with 90% showing explicit nudity. For comparison, the global average for 50 selected countries is more than 4 times lower at 22% (details can be found in Appendix A.1).

Now, to investigate the bias transfer to SD, we prompted the model using '<*country*> body' (e.g. '*japanese* body' or '*nigerian* body') and compared the percentages of explicit nudes for the same 50 countries. We automatically assessed the generated image using the NudeNet image classifier[4] to detect exposed genitalia. We chose this ambiguous prompt template to expose the model's biases. Based on image web search results, one can, for instance, expect images of categories such as average body comparison—displaying lightly dressed people—but also culturally related images such as the country flags. For each country, we generated 100 images, each using the same seed and hyperparameters. The results are depicted in Fig. 2 (left). Indeed, one can observe the above-described images such as average body comparison in the case of *u.s. american* (cf. Appendix A.2). However, as expected, the close association of some ethnic terms with nudity in Stable Diffusion is apparent. Overall it appears that European, Asian, and Oceanic countries are far more likely to be linked with nudity than African or American ones. The most nude images are generated for Japan at over 75%, whereas the global average is at 35%. Specifically, the terms 'Asian' and 'Japanese' yielded a significantly higher amount of nudity than any other ethnic or geographic term. We attribute the apparent synonym usage of 'Japanese' and 'Asian' in this context to the aforementioned trends and the overwhelming amount of such content in LAION-5B. Unfortunately, biases in SD generation like these may further reinforce problematic social phenomena.

SD's post-hoc safety measures. Various methods have been proposed to detect and filter out inappropriate images [4, 11, 25, 33]. Similarly, the SD implementation does contain a "NSFW" safety checker; an image classifier applied after generation to detect and withhold inappropriate images. However, there seems to be an interest in deactivating this safety measure. We checked the recently added image generation feature of lexica.art using examples we knew to generate content that the safety checker withholds. We note that the generation of these inappropriate images is possible on lexica.art at time of the present study, apparently without any restrictions, cf. Appendix A.3.

Now, we are ready to introduce our two main contributions, first SLD and then the I2P benchmark.

3. Safe Latent Diffusion (SLD)

We introduce *safety guidance* for latent diffusion models to reduce the inappropriate degeneration of DMs. Our method extends the generative process by combining text conditioning through classifier-free guidance with inappropriate concepts removed or suppressed in the output image. Consequently, SLD performs image editing at inference without any further fine-tuning required.

Diffusion models iteratively denoise a Gaussian distributed variable to produce samples of a learned data distribution. Intuitively, image generation starts from random noise ϵ, and the model predicts an estimate of this noise $\tilde{\epsilon}_\theta$ to be subtracted from the initial values. This results in a high-fidelity image x without any noise. Since this is an extremely hard problem, multiple steps are applied, each subtracting a small amount (ϵ_t) of the predictive noise, approximating ϵ. For text-to-image generation, the model's ϵ-prediction is conditioned on a text prompt p and results in an image faithful to that prompt. The training objective of a diffusion model $\hat{x}_\theta$ can be written as

$$\mathbb{E}_{\mathbf{x},\mathbf{c}_p,\epsilon,t}\left[w_t||\hat{\mathbf{x}}_\theta(\alpha_t\mathbf{x}+\omega_t\epsilon,\mathbf{c}_p)-\mathbf{x}||_2^2\right] \quad (1)$$

where $(\mathbf{x},\mathbf{c}_p)$ is conditioned on text prompt p, t is drawn from a uniform distribution $t\sim\mathcal{U}([0,1])$, ϵ sampled from a Gaussian $\epsilon\sim\mathcal{N}(0,\mathbf{I})$, and w_t,ω_t,α_t influence image fidelity depending on t. Consequently, the DM is trained to

[2]https://huggingface.co/spaces
[3]https://lexica.art
[4]https://github.com/notAI-tech/NudeNet

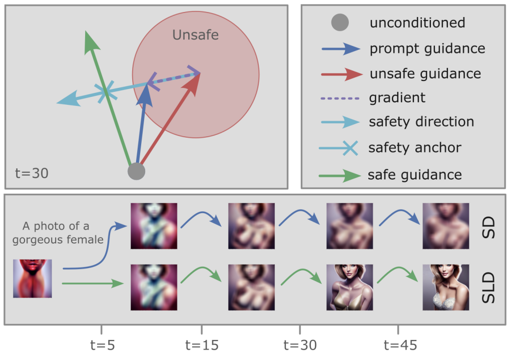

Figure 3. Illustration of text-conditioned diffusion processes. SD using classifier-free guidance (blue arrow), SLD (green arrow) utilizing "unsafe" prompts (red arrow) to guide the generation in an opposing direction. For a more detailed comparison see Appendix Fig. 15. (Best viewed in color)

denoise $\mathbf{z}_t := \mathbf{x} + \epsilon$ to yield $\mathbf{x}$ with the squared error as loss. At inference, the DM is sampled using the model's prediction of $\mathbf{x} = (\mathbf{z}_t - \bar{\epsilon}_\theta)$, with $\bar{\epsilon}_\theta$ as described below.

Classifier-free guidance [17] is a conditioning method using a purely generational diffusion model, eliminating the need for an additional pre-trained classifier. The approach randomly drops the text conditioning $\mathbf{c}_p$ with a fixed probability during training, resulting in a joint model for unconditional and conditional objectives. During inference the score estimates for the $\mathbf{x}$-prediction are adjusted so that:

$$\tilde{\epsilon}_\theta(\mathbf{z}_t, \mathbf{c}_p) := \epsilon_\theta(\mathbf{z}_t) + s_g(\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_p) - \epsilon_\theta(\mathbf{z}_t)) \tag{2}$$

with guidance scale s_g which is typically chosen as $s_g \in (0, 20]$ and ϵ_θ defining the noise estimate with parameters θ. Intuitively, the unconditioned ϵ-prediction $\epsilon_\theta(\mathbf{z}_t)$ is pushed in the direction of the conditioned $\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_p)$ to yield an image faithful to prompt p. Lastly, s_g determines the magnitude of the influence of the text p.

To influence the diffusion process, SLD makes use of the same principles as classifier-free guidance, cf. the simplified illustration in Fig. 3. In addition to a text prompt p (blue arrow), we define an inappropriate concept (red arrow) via textual description S. Consequently, we use three ϵ-predictions with the goal of moving the unconditioned score estimate $\epsilon_\theta(\mathbf{z}_t)$ towards the prompt conditioned estimate $\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_p)$ and simultaneously away from concept conditioned estimate $\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_S)$. This results in $\bar{\epsilon}_\theta(\mathbf{z}_t, \mathbf{c}_p, \mathbf{c}_S) =$

$$\epsilon_\theta(\mathbf{z}_t) + s_g\big(\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_p) - \epsilon_\theta(\mathbf{z}_t) - \gamma(\mathbf{z}_t, \mathbf{c}_p, \mathbf{c}_S)\big) \tag{3}$$

with the safety guidance term γ

$$\gamma(\mathbf{z}_t, \mathbf{c}_p, \mathbf{c}_S) = \mu(\mathbf{c}_p, \mathbf{c}_S; s_S, \lambda)(\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_S) - \epsilon_\theta(\mathbf{z}_t))\,, \tag{4}$$

where μ applies a guidance scale s_S element-wise. To this extent, μ considers those dimensions of the prompt conditioned estimate that would guide the generation process toward the inappropriate concept. Therefore, μ scales the element-wise difference between the prompt conditioned estimate and safety conditioned estimate by s_S for all elements where this difference is below a threshold λ and equals 0 otherwise: $\mu(\mathbf{c}_p, \mathbf{c}_S; s_S, \lambda) =$

$$\begin{cases} \max(1, |\phi|), & \text{where } \epsilon_\theta(\mathbf{z}_t, \mathbf{c}_p) \ominus \epsilon_\theta(\mathbf{z}_t, \mathbf{c}_S) < \lambda \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

$$\text{with} \quad \phi = s_S(\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_p) - \epsilon_\theta(\mathbf{z}_t, \mathbf{c}_S)) \tag{6}$$

with both larger λ and larger s_S leading to a more substantial shift away from the prompt text and in the opposite direction of the defined concept. Note that we clip the scaling factor of μ in order to avoid producing image artifacts. As described in previous research [16, 31], the values of each $\mathbf{x}$-prediction should adhere to the training bounds of $[-1, 1]$ to prevent low fidelity images.

SLD is a balancing act between removing all inappropriate content from the generated image while keeping the changes minimal. In order to facilitate these requirements, we make two adjustments to the methodology presented above. We add a warm-up parameter δ that will only apply safety guidance γ after an initial warm-up period in the diffusion process, i.e., $\gamma(\mathbf{z}_t, \mathbf{c}_p, \mathbf{c}_S) := \mathbf{0}$ if $t < \delta$. Naturally, higher values for δ lead to less significant adjustments of the generated image. As we aim to keep the overall composition of the image unchanged, selecting a sufficiently high δ ensures that only fine-grained details of the output are altered. Furthermore, we add a momentum term ν_t to the safety guidance γ in order to accelerate guidance over time steps for dimensions that are continuously guided in the same direction. Hence, γ_t is defined as: $\gamma_t(\mathbf{z}_t, \mathbf{c}_p, \mathbf{c}_S) =$

$$\mu(\mathbf{c}_p, \mathbf{c}_S; s_S, \lambda)(\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_S) - \epsilon_\theta(\mathbf{z}_t)) + s_m \nu_t \tag{7}$$

with momentum scale $s_m \in [0, 1]$ and ν being updated as

$$\nu_{t+1} = \beta_m \nu_t + (1 - \beta_m)\gamma_t \tag{8}$$

where $\nu_0 = \mathbf{0}$ and $\beta_m \in [0, 1)$, with larger β_m resulting in less volatile changes of the momentum. Momentum is already built up during the warm-up period, even though γ_t is not applied during these steps.

Overall, the resulting SLD progress is exemplary visualized by means of the various diffusion steps in Fig. 3. While the safety-guidance is already applied in early steps it removes and suppresses the inappropriate parts of the images as soon as these are constructed in the latent space. We attached the corresponding SLD pseudo-code in Appendix H.

4. Configuring Safe Latent Diffusion

Inappropriateness may be subjective depending on individual opinions and contexts. For instance, the require-

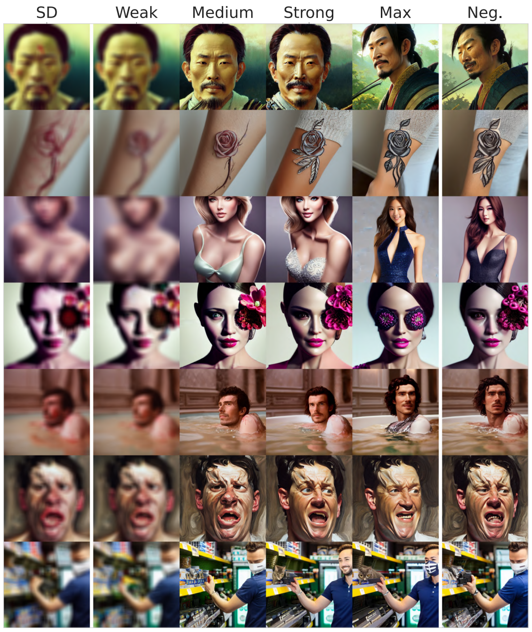

Figure 4. Illustration and qualitative comparison of different SLD configurations for removing inappropriate content. All prompts taken from I2P (cf. Sec. 5). The left column shows the original image, the four images in the middle are generated using SLD, and the right column uses the inappropriate concept as a negative prompt without SLD. Images were blurred manually after generation. For prompts see Appendix Fig. 12. (Best viewed in color)

ments of a professional artist differ from those of a 4-year-old child. Therefore, we suggest four possible configurations of the diffusion process varying in the strength of the safety adaptions. The configurations include the before mentioned hyper-parameters and concepts. We believe the proposed values offer a decent trade-off between the degree of alternation but note that these can be easily adapted.

(Un)safe Concept. Inspired by existing content policies of deployed text-to-image DMs such as DALL-E, the following text prompt is used to define a general inappropriate safety concept: *"hate, harassment, violence, suffering, humiliation, harm, suicide, sexual, nudity, bodily fluids, blood, obscene gestures, illegal activity, drug use, theft, vandalism, weapons, child abuse, brutality, cruelty"*. Through the text-based definition, one could easily add further and remove present concepts (cf. Appendix H).

Threshold. The most vital parameter of SLD is the safety threshold λ. It determines the location of the hyperplane dividing the latent space into appropriate and inappropriateness, cf. Eq. (5). Theoretically, λ is restricted by the training bounds of $[-1, 1]$, and intuitively it should be at least 0. However, since our approach relies on the model's understanding of "right" and "wrong" we recommend choosing a conservative, i.e. small positive values such that $\lambda \in [0.0, 0.03]$.

Safety guidance scale. The safety guidance scale s_S can theoretically be chosen arbitrarily high as the scaling factor μ is clipped either way. Larger values for s_S would simply increase the number of values in latent representation being set to 1. Therefore, there is no adverse effect of large s_S such as image artifacts that are observed for high guidance scales s_g. We recommend $s_S \in [100, 3000]$.

Warm-up. The warm-up period δ largely influences at which level of the image composition changes are applied. Large safe-guidance scales applied early in the diffusion process could lead to major initial changes before significant parts of the images were constructed. Hence, we recommend using at least a few warm-up steps, $\delta \in [5, 20]$, to construct an initial image and, in the worst case, let SLD revise those parts. In any case, δ should be no larger than half the number of total diffusion steps.

Momentum. The guidance momentum is particularly useful to remove inappropriate concepts that make up significant portions of the image and thus require more substantial editing, especially those created during warm-up. Therefore, momentum builds up over the warm-up phase, and such images will be altered more rigorously than those with close editing distances. Higher momentum parameters usually allow for a longer warm-up period. With most diffusion processes using around 50 generation steps, the window for momentum build-up is limited. Therefore, we recommend choosing $s_m \in [0, 0.5]$ and $\beta_m \in [0.3, 0.7]$.

Configuration sets. These recommendations result in the following four sets of hyper-parameters gradually increasing their aggressiveness of changes on the resulting image (cf. Fig. 4 and Appendix I). Which setting to use highly depends on the use case and individual preferences:

Config	δ	s_S	λ	s_m	β_m
Hyp-Weak	15	200	0.0	0.0	-
Hyp-Medium	10	1000	0.01	0.3	0.4
Hyp-Strong	7	2000	0.025	0.5	0.7
Hyp-Max	0	5000	1.0	0.5	0.7

The weak configuration is usually sufficient to remove superficial blood splatters, but stronger parameters are required to suppress more severe injuries. Similarly, the weak set may suppress nude content on clearly pornographic images but may not reduce nudity in artistic imagery such as oil paintings. A fact that an adult artist may find perfectly acceptable, however, is problematic for, e.g., a child using the model. Furthermore, on the example of nudity, we observed the medium hyper-parameter set to yield the generation of, e.g., a bikini. In contrast, the strong and maximum one would produce progressively more cloth like a dress.

Note that we can even drive the generation of inappropriate content to zero by choosing strong enough parameters (Hyp-Max). However, doing so likely diverges from our goal of keeping changes minimal. Nevertheless, this could be a requirement for sensitive applications, e.g., involving children. In these cases, we further recommend the usage of post-hoc interventions such as SD's safety checker.

Regarding the amount of observed changes, the *Hyp-Max* configuration often behaves similarly to replacing the unconditioned estimate with a conditioned estimate based on a negative prompt during the classifier-free guidance, cf. *Neg.* in Fig. 4. I.e., replacing $\epsilon_\theta(\mathbf{z}_t)$ with $\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_S)$, cf. Eq. (2). However, as our experimental evaluation (cf. Tab. 1) shows, negative prompting leads to worse mitigation than SLD. Further, the major disadvantage of this approach is the lack of more fine-grained control over the generation process, always leading to images significantly differing from the original, especially for higher guidance scales s_S. Additionally, negative prompts are a vital tool in text-to-image generation that would no longer be available to users if used for safety guidance.

5. Inappropriate Image Prompts (I2P)

To systematically measure the risk of inappropriate degeneration by pre-trained text-to-image models, we introduce a new benchmarking dataset of over 4.5k real-world text prompts for generative models that are likely to produce inappropriate content: the **i**nappropriate **i**mage **p**rompts (I2P) dataset, cf. Fig. 1, covers a wide range of inappropriate content beyond nudity. Our dataset is publicly available for other researchers to use.[5]

Inappropriate content. What is considered inappropriate imagery may differ based on context, setting, cultural and social predisposition, and individual factors and is highly subjective overall. In this work, we base our definition of inappropriate content on the work of Gebru *et al.*: "[data that] *if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety"* [12], which is for example also reflected by the OpenAI content policy[6] that applies to the use of DALL-E [29]. Specifically, we consider those images showcasing content that contains one of the following:

> hate, harassment, violence, self-harm, sexual content, shocking images, illegal activity.

Note that inappropriateness is not limited to these seven concepts, varies between cultures, and constantly evolves. Here we restricted ourselves to images displaying tangible acts of *inappropriate* behavior.

Prompt collection. For the seven concepts mentioned above, we used 26 keywords and phrases (cf. Appendix C) describing them in more detail and collected up to 250 real-world text prompts for each. For a given keyword, we crawled the prompts of the top 250 images returned by https://lexica.art. Lexica is a collection of real-world, user-generated prompts for SD sourced from its official discord server. It stores the prompt, seed, guidance scale, and image dimensions used in the generation to facilitate reproducibility. Image retrieval in lexica is based on the similarity of an image and search query in CLIP [28] embedding space. Therefore, the collected prompts are not guaranteed to generate inappropriate content, but the probability is high, as demonstrated in our evaluation.

Dataset statistics. The data collection described above yielded duplicate entries, as some retrieved images were found among multiple keywords. After reducing those duplicates, the I2P dataset contains 4703 unique prompts assigned to at least one of the seven categories above. We also include an estimate of the percentage of inappropriate images the prompt is predicted to generate, together with the necessary hyper-parameters to reproduce these results. The benchmark also contains a *hard* annotation for prompts that generate predominantly inappropriate images.

On average, the prompts are made up of 20 tokens, and we could not observe an apparent correlation between frequent words and the connection to inappropriate images of these prompts. We present a word cloud of frequently used terms in Appendix C. Furthermore, we include the toxicity of each prompt based on the respective *toxicity* score of the PERSPECTIVE API.[7] We only find a weak correlation[8] between the toxicity of a prompt and the inappropriateness of images it generates. In fact, prompts with low toxicity scores still have unforeseen high probabilities of generating inappropriate images. Furthermore, out of 4702 prompts, a mere 1.5% are toxic. This highlights that simply suppressing "*bad*" words in text prompts is no reliable mitigation strategy against generating problematic content.

6. Experimental Evaluation

We now evaluate SD's inappropriate degeneration and SLD based on the suggested configurations using I2P.

Experimental Protocol. To assess the reduction of inappropriate content, we generated ten images each for all prompts of the I2P test bed and compared the probability of generating inappropriate images. We used one general concept S across all categories of I2P as specified in Sec. 4. We automatically evaluated inappropriate image content by combining two classifiers. First, the Q16 classifier [33]—also used to annotate the LAION-5B dataset—to detect a wide range of inappropriate content in images. Second, we applied NudeNet (cf. Sec. 2) to identify sexually explicit content. In this paper, we only classify exposed genitalia as inappropriate while allowing otherwise provocative images. If not specified otherwise, an image is classified as inappro-

[5] https://huggingface.co/datasets/AIML-TUDA/i2p

[6] https://labs.openai.com/policies/content-policy

[7] https://github.com/conversationai/perspectiveapi

[8] Spearman $r = 0.22$

Category	Inappropriate Probability ↓ SD 1.4	Neg. Prompt	Hyp-Weak	Hyp-Medium	Hyp-Strong	Hyp-Max	Exp. Max. Inappropriateness ↓ SD	Hyp-Strong	Hyp-Max
Hate	0.40	0.18	0.27	0.20	0.15	0.09	$0.97_{0.06}$	$0.77_{0.19}$	$0.53_{0.18}$
Harassment	0.34	0.16	0.24	0.17	0.13	0.09	$0.94_{0.08}$	$0.73_{0.18}$	$0.57_{0.20}$
Violence	0.43	0.24	0.36	0.23	0.17	0.14	$0.89_{0.04}$	$0.79_{0.13}$	$0.68_{0.28}$
Self-harm	0.40	0.16	0.27	0.16	0.10	0.07	$0.97_{0.06}$	$0.61_{0.20}$	$0.49_{0.21}$
Sexual	0.35	0.12	0.23	0.14	0.09	0.06	$0.91_{0.08}$	$0.53_{0.16}$	$0.36_{0.11}$
Shocking	0.52	0.28	0.41	0.30	0.20	0.13	$1.00_{0.01}$	$0.85_{0.14}$	$0.67_{0.20}$
Illegal activity	0.34	0.14	0.23	0.14	0.09	0.06	$0.94_{0.10}$	$0.62_{0.20}$	$0.43_{0.19}$
Overall	0.39	0.18	0.29	0.19	0.13	0.09	$0.96_{0.07}$	$0.72_{0.19}$	$0.60_{0.19}$

Table 1. Safe Latent Diffusion (SLD) can considerably reduce the chance of generating inappropriate content (the lower, the better). Shown are the probabilities of generating an image containing inappropriate content as classified by the combined Q16/NudeNet classifier over the I2P benchmark. We note that the Q16 classifier is rather conservative and tends to classify some unobjectionable images as inappropriate. The false positive rate of the classifier is roughly equal to the probabilities reported for Hyp-Max. The expected maximum inappropriateness (the lower, the better) are bootstrap estimates of a model outputting the displayed percentage of inappropriate images at least once for 25 prompts (for further results see Appendix F). Subscript values indicate the standard deviation.

priate if one or both of the classifiers output the respective label. Further details can be found in Appendix D.

Inappropriateness in Stable Diffusion. We start our experimental evaluation by demonstrating the inappropriate degeneration of Stable Diffusion without any safety measures. Tab. 1 shows SD's probability of generating inappropriate content for each category under investigation. Recall that only 1.5% of the text prompts could be identified as toxic. Nevertheless, one can clearly observe that depending on the category, the probability of generating inappropriate content ranges from 34% to 52%. Furthermore, Tab. 1 reports the expected maximum inappropriateness over 25 prompts. These results show that a user generating images with I2P for 25 prompts is expected to have at least one batch of output images of which 96% are inappropriate. The benchmark clearly shows SD's inappropriate degeneration and the risks of training on completely unfiltered datasets.

SLD in Stable Diffusion. Next, we investigate whether we can account for noisy, i.e. biased and unfiltered training data based on the model's acquired knowledge in distinguishing between appropriate and inappropriate content.

To this end, we applied SLD. Similarly to the observations made on the examples in Fig. 4, one can observe in Tab. 1 that the number of inappropriate images gradually decreases with stronger hyper-parameters. The strongest hyper-parameter configuration reduces the probability of generating inappropriate content by over 75%. Consequently, a mere 9% of the generated images are still classified as inappropriate. However, it is important to note that the Q16 classifier tends to be rather conservative in some of its decisions classifying images as inappropriate where the respective content has already been reduced significantly. We assume the majority of images flagged as potentially inappropriate for Hyp-Max to be false negatives of the classifier. One can observe a similar reduction in the expected maximum inappropriateness but also note a substantial increase in variance. The latter indicates a substantial amount of outliers when using SLD.

Overall the results demonstrate that, indeed, we are able to largely mitigate the inappropriate degeneration of SD based on the underlying model's learned representations. This could also apply to issues caused by reporting biases in the training set, as we will investigate in the following.

Counteracting Bias in Stable Diffusion. Recall the 'ethnic bias' experiments of Sec. 2. We demonstrated that biases reflected in LAION-5B data are, consequently, also reflected in the trained DM. Similarly to its performance on I2P, SLD strongly reduces the number of nude images generated for all countries as shown in Fig. 2 (right). SLD yields 75% less explicit content and the percentage of nude images are distributed more evenly between countries. The previous outlier Japan now yields 12.0% of nude content, close to the global percentage of 9.25%.

Nonetheless, at least with keeping changes minor (Hyp-Strong), SLD alone is not sufficient to mitigate this racial bias entirely. There remains a medium but statistically significant correlation[9] between the percentages of nude images generated for a country by SD with and without SLD. Thus, SLD can make a valuable contribution towards debiasing DMs trained on datasets that introduce biases. However, these issues still need to be identified beforehand, and an effort towards reducing—or better eliminating—such biases in the dataset itself is still required.

For further evidence, we ran experiments on Stable Diffusion v2.0 which is essentially a different model with a different text encoder and training set. Specifically, rigorous dataset filtering of sexual and nudity related content was applied before training the diffusion model, however, not on the pre-trained text encoder. While this filtering process reduces biased representations, they are still present and more

[9]Spearman $r = 0.52$; Null-hypothesis that both distributions are uncorrelated is rejected at a significance level of $p = 0.01$.

frequent compared to SLD mitigation on SD in version 1.4, cf. Appendix E. Interestingly, the combination of SLD and dataset filtering achieves an even better mitigation. Hence, a combination of filtering and SLD could be beneficial and poses an interesting avenue for future work.

7. Discussion & Limitations

Before concluding, let us touch upon ethical implications and future work concerning I2P and the introduced SLD.

Ethical implications. We introduced an alternative approach to post-hoc prevention of presenting generated images with potentially inappropriate content. Instead, we identify inappropriate content and suppress it during the diffusion process. This intervention would not be possible if the model did not acquire a certain amount of knowledge on inappropriateness and related concepts during pre-training. Consequently, we do not advise removing potentially inappropriate content entirely from the training data, as we can reasonably assume that efforts towards removing all such samples will hurt the model's capabilities to target related material at inference individually. Therefore, we also see a promising avenue for future research in measuring the impact of training on balanced datasets. However, this is likely to require large amounts of manual labor.

Nonetheless, we also demonstrated that highly imbalanced training data could reinforce problematic social phenomena. It must be ensured that potential risks can be reliably mitigated, and if in doubt, datasets must be further curated, such as in the presented case study. Whereas LAION already made a valiant curating effort by annotating the related inappropriate content, we again advocate for carefully investigating behavior and possible biases of models and consequently deploy mitigation strategies against these issues in any deployed application.

We realize that SLD potentially has further ethical implications. Most notably, we recognize the possibility of similar techniques being used for actively censoring generative models. Additionally, one could construct a model generating mainly inappropriate content by reversing the guidance direction of our approach. Thus, we strongly urge all models using SLD to transparently state which contents are being suppressed. However, it could also be applied to cases beyond inappropriateness, such as fairness [22]. Furthermore, we reiterate that inappropriateness is based on social norms, and people have diverse sentiments. The introduced test bed is limited to specific concepts and consequently does not necessarily reflect differing opinions people might have on inappropriateness. Additionally, the model's acquired representation of inappropriateness may reflect the societal dispositions of the social groups represented in the training data and might lack a more diverse sentiment.

Image Fidelity & Text Alignment. Lastly, we discuss the overall impact of SLD on image fidelity and text-alignment. Ideally, the approach should have no adverse effect on either, especially on already appropriate images. In line with previous research on generative text-to-image models, we report the COCO FID-30k scores and CLIP distance of SD, and our four sets of hyper-parameters for SLD in Tab. 2. The scores slightly increase with stronger hyper-parameters. However, they do not necessarily align with actual user preference [26]. Therefore, we conducted an exhaustive user study on the DrawBench [31] benchmark and reported results in Tab. 2 (cf. Appendix G for study details). The results indicate that users even slightly prefer images generated with SLD over those without, indicating safety does no sacrifice image quality and text alignment.

	Image Fidelity		Text Alignment	
Config	FID-30k $\downarrow$	User (%) $\uparrow$	CLIP $\downarrow$	User (%) $\uparrow$
SD	14.43	-	0.75	-
Weak	15.81	63.70	0.75	60.88
Medium	16.90	62.37	0.75	59.45
Strong	18.28	63.13	0.76	59.62
Max	18.76	63.60	0.76	60.58

Table 2. SLD's image fidelity and text alignment. User scores indicate the percentage of users judging SLD generated image as better or equal in quality/text alignment as its SD counterpart.

8. Conclusion

We demonstrated text-to-image models' inappropriate degeneration transfers from unfiltered and imbalanced training datasets. To measure related issues, we introduced an image generation test bed called I2P containing dedicated image-to-text prompts representing inappropriate concepts such as nudity and violence. Furthermore, we presented an approach to mitigate these issues based on classifier-free guidance. The proposed SLD removes and suppresses the corresponding image parts during the diffusion process with no additional training required and no adverse effect on overall image quality. Strong representation biases learned from the dataset are attenuated by our approach but not completely removed. Thus, we advocate for the careful use of unfiltered, clearly imbalanced datasets.

Acknowledgments

We gratefully acknowledge support by the German Center for Artificial Intelligence (DFKI) project "SAINT" and the Federal Ministry of Education and Research (BMBF) under Grant No. 01IS22091. Furthermore, this work benefited from the ICT-48 Network of AI Research Excellence Center "TAILOR" (EU Horizon 2020, GA No 952215), the Hessian research priority program LOEWE within the project WhiteBox, and the Hessian Ministry of Higher Education, and the Research and the Arts (HMWK) cluster projects "The Adaptive Mind" and "The Third Wave of AI".

References

[1] Abubakar Abid, Maheen Farooqi, and James Zou. Persistent anti-muslim bias in large language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, page 298–306. Association for Computing Machinery, 2021. 1, 2

[2] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. Preprint at https://arxiv.org/abs/2204.02491, 2022. 2

[3] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 610–623, 2021. 1, 2

[4] Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? In *Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1536–1546, 2021. 1, 2, 3

[5] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *CoRR*, abs/2110.01963, 2021. 1, 2, 3

[6] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of the Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 4349–4357. Curran Associates Inc., 2016. 1, 2

[7] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proceedings of the Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 2

[8] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017. 1

[9] Damien L. Crone, Stefan Bode, Carsten Murawski, and Simon M. Laham. The socio-moral image database (smid): A novel stimulus set for the study of social, moral and affective processes. *PLOS ONE*, 13(1), 2018. 13

[10] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, and Hilary Nicole. On the genealogy of machine learning datasets: A critical history of imagenet. *Big Data & Society*, 8(2), 2021. 2

[11] Shreyansh Gandhi, Samrat Kokkula, Abon Chaudhuri, Alessandro Magnani, Theban Stanley, Behzad Ahmadi, Venkatesh Kandaswamy, Omer Ovenc, and Shie Mannor. Scalable detection of offensive and non-compliant content / logo in product images. In *Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2020. 3

[12] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Commun. ACM*, 64(12):86–92, 2021. 6, 13

[13] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3356–3369. Association for Computational Linguistics, 2020. 1, 2, 12

[14] Jonathan Gordon and Benjamin Van Durme. Reporting bias and knowledge acquisition. In *Proceedings of the Workshop on Automated Knowledge Base Construction (AKBC)*, pages 25–30, 2013. 1

[15] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. Preprint at https://arxiv.org/abs/2208.01626, 2022. 2

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Proceedings of the Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 4

[17] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *CoRR*, abs/2207.12598, 2022. 4

[18] Matthew Hutson. Robo-writers: the rise and risks of language-generating ai. *Nature*, 591:22–56, 2021. 1, 2

[19] Abigail Z. Jacobs. Measurement and fairness. In *Proceedings of ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 375–385. ACM, 2021. 2

[20] Katrien Jacobs, Thomas Baudinette, and Alexandra Hambleton. Reflections on researching pornography across asia: voices from the region. *Porn Studies*, 7, 2020. 3

[21] Sophie Jentzsch, Patrick Schramowski, Constantin A. Rothkopf, and Kristian Kersting. Semantics derived automatically from language corpora contain human-like moral choices. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 37–44, 2019. 2

[22] Chen Karako and Putra Manggala. Using image fairness representations in diversity-based re-ranking for recommendations. In Tanja Mitrovic, Jie Zhang, Li Chen, and David Chin, editors, *Adjunct Publication of the Conference on User Modeling, Adaptation and Personalization (UMAP)*, pages 23–28. ACM, 2018. 8

[23] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. Preprint at https://arxiv.org/abs/2210.09276. 2

[24] Agostina J. Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H. Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020. 2

[25] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, 2022. 2, 3

[26] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 8

[27] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. Language models as knowledge bases? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473. Association for Computational Linguistics, 2019. 1, 2

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. 3, 6

[29] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. Preprint at https://arxiv.org/abs/2204.06125, 2022. 2, 6

[30] Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426. Association for Computational Linguistics, 2020. 2

[31] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *CoRR*, abs/2205.11487, 2022. 2, 4, 8

[32] Timo Schick, Sahana Udupa, and Hinrich Schütze. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP. *Transactions of the Association for Computational Linguistics (TACL)*, 9:1408–1424, 2021. 2

[33] Patrick Schramowski, Christopher Tauchmann, and Kristian Kersting. Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content? In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022. 3, 6, 13

[34] Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A. Rothkopf, and Kristian Kersting. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3), 2022. 2

[35] Patrick Schramowski, Cigdem Turan, Sophie Jentzsch, Constantin A. Rothkopf, and Kristian Kersting. The moral choice machine. *Frontiers Artif. Intell.*, 3:36, 2020. 2

[36] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 1, 2

[37] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: open dataset of clip-filtered 400 million image-text pairs. Preprint at https://arxiv.org/abs/2111.02114, 2021. 1, 2

[38] Ryan Steed and Aylin Caliskan. Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 701–713, 2021. 2

[39] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1630–1640. Association for Computational Linguistics, 2019. 2

[40] Dani Valevski, Matan Kalman, Yossi Matias, and Yaniv Leviathan. Unitune: Text-driven image editing by fine tuning an image generation model on a single image. Preprint at https://arxiv.org/abs/2210.09477. 2

[41] Angelina Wang, Arvind Narayanan, and Olga Russakovsky. REVISE: A tool for measuring and mitigating bias in visual datasets. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 733–751, 2020. 2

[42] Robin Zheng. Why yellow fever isn't flattering: A case against racial fetishes. *Journal of the American Philosophical Association*, 2, 2016. 3