Mitigating Inappropriateness in Image Generation: Can there be Value in Reflecting the World's Ugliness?

Manuel Brack [1 2] **Felix Friedrich** [2 3] **Patrick Schramowski** [1 2 3 4] **Kristian Kersting** [1 2 3 5]

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 web, they also reproduce inappropriate human behavior. Specifically, we demonstrate inappropriate degeneration on a large-scale for various generative text-to-image models, thus motivating the need for monitoring and moderating them at deployment. To this end, we evaluate mitigation strategies at inference to suppress the generation of inappropriate content. Our findings show that we can use models' representations of the world's ugliness to align them with human preferences.

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

1. Introduction

Next to text-generative models such as ChatGPT, image-generative models are becoming increasingly prevalent and seeing growing adoption in commercial services such as stockimagery and graphic design. Due to their large-scale self-supervised training paradigm they retain general knowledge implicitly present in the data and are able to generate high fidelity images faithful interpretations to the users' prompts. However, their learning setup, which includes large-scale unfiltered data (Schuhmann et al., 2022; Birhane et al., 2021), also leads to degenerated and biased behavior (Schramowski et al., 2023), calling for mitigation strategies and the moderation of generative models in deployed systems. Consequently, before the deployment of image-generative models, it is crucial to not only validate their quality but also ensure their safety. This necessitates the assessment of appropriate guardrails, which should be tailored to the specific application at hand. Previous work in this domain has primarily relied on anecdotal evidence, lacking quantifiable measures that take multiple models and architectures into account. Indeed, Schramowski et al. (2023) proposed an empirical benchmark but limited their evaluation to a single Stable Diffusion version.

To help the development of effective mitigation strategies and moderation techniques for image-generative models in real-world systems, we here present a comprehensive assessment of inappropriate degeneration across numerous open-source models and architectures. More precisely, we investigate how effectively these models can be instructed to suppress inappropriate content using the knowledge obtained about the world's ugliness. Our findings suggest that safety mitigation of text-to-image generators can be performed through direct instructions at inference for various types of models. In total, we generated and evaluated over 1.5M images for 11 different models, thereby providing a large-scale investigation of the topic.

2. Instructing Models on the World's Ugliness

Visual Moderation. There exist multiple approaches for mitigating inappropriate degeneration of generative models. Previous research has identified four major methods. The first approach involves filtering the training data to remove problematic content entirely (Nichol et al., 2022). However, large-scale dataset filtering can have unexpected side effects on downstream performance as demonstrated by Nichol et al. (2022). Moreover, determining what constitutes inappropriate content is highly subjective and dependent on various external factors such as individual and societal norms as well as the specific use case of the application. Developing a dedicated model with data filtering tailored to each definition of inappropriateness is difficult, if not impractical, particularly as it would require retraining pre-existing models from scratch. To overcome this limitation, a second approach involves finetuning a pre-trained model to erase

[1]German Center for Artificial Intelligence (DFKI) [2]Computer Science Department, TU Darmstadt [3]Hessian Center for AI (hessian.AI) [4]LAION [5]Centre for Cognitive Science, TU Darmstadt. Correspondence to: Manuel Brack <brack@cs.tu-darmstadt.de>.

Workshop on Challenges in Deployable Generative AI at International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA. 2023.

Figure 1: Examples of inappropriate degeneration and their mitigation across various models. From left to right each batch shows the original image and the instruction-based suppression with SEGA and negative prompting. Prompts are taken from the inappropriate-image-prompts (I2P) dataset. Images displaying nudity were blurred by the authors. (Best viewed in color)

inappropriate concepts (Gandikota et al., 2023). While this method requires lower computational resources compared to training an entire model, it is still constrained in its ability to account for diverse definitions of inappropriateness. Another relevant approach, particularly for deployed applications, involves implementing input and output filters[1]. In hosted inference services, input prompts are typically filtered for banned keywords, and the generated images are scanned for inappropriate content before being presented to users. Although this approach restricts the availability of unwanted content, it has some drawbacks. Schramowski et al. (2023) have demonstrated that inappropriate degeneration can occur unexpectedly for prompts lacking explicit descriptions of any problematic concepts. Therefore, input filters are prone to missing these implicit correlations. Additionally, the generation and subsequent discarding of images not only wastes computational resources but can also result in a frustrating user experience.

In contrast, we here explore the idea of leveraging a model's learned representations of inappropriate content for mitigation of such material. We focus on explicit instruction approaches that provide textual descriptions to the model regarding undesired concepts during the image generation process. This results in both high flexibility and customizability, as the instruction prompt can be easily modified to adapt to different requirements. Consequently, the user remains involved in the process and the method enables seamless deployment across various architectures. As such instruction-based methods also facilitate large-scale evaluation across models.

Classifier Free Guidance. Before going into detail on different instruction methods for image generation, we need to establish some fundamentals of text-to-image diffusion models (DMs). 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 a complex 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. To that end, DMs employ classifier-free guidance (Ho & Salimans, 2022), a conditioning method using a purely generational diffusion model, eliminating the need for an additional pre-trained classifier. The noise estimate $\tilde{\epsilon}_\theta$ uses an unconditioned ϵ-prediction $\epsilon_\theta(\mathbf{z}_t)$ which is pushed in the direction of the conditioned estimate $\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_p)$ to yield an image faithful to prompt p.

Instructing Text-to-Image Models on Safety. We now consider two different instruction approaches extending the principles of classifier-free guidance. Both methods rely on a secondary text prompt s that describes concepts to suppress during generation. First, negative prompting replaces the unconditioned ϵ-prediction $\epsilon_\theta(\mathbf{z}_t)$ with one conditioned on s: $\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_s)$, thus moving away from the inappropriate concepts. This approach is intuitive and easy to implement, however offers limited control over the extent of content suppression. Additionally, we use Semantic Guidance (SEGA) (Brack et al., 2023) which is a powerful method for image manipulation based on additional text prompts. SEGA adds an additional guidance term to $\tilde{\epsilon}_\theta$ that allows us to steer the generation away from s, while keeping changes to the original image minimal.

3. Assessment of Inappropriate Degeneration

Experimental Protocol. In our large-scale evaluation we included 11 text-to-image models that can be organized in four categories. 1) Different version of Stable Diffusion (SD), 2) selected, popular checkpoints of SD that are further tuned for specific use cases, 3) models extending SD by aligning the generative module to a new text encoder with additional capabilities and 4) other architectures using diffusion on quantized latent spaces or pixel-level. We provide detailed background information in App A.

For each model, we considered three distinct setups: the base model, mitigation with negative prompts and with SEGA. For evaluation we used the inappropriate-image-prompts dataset (I2P) (Schramowski et al., 2023). The I2P benchmark contains over 4.7k real user prompts for text-to-

[1] https://www.technologyreview.com/2023/02/24/1069093/

Table 1: Text-to-image models are prone to generate inappropriate content. Instruction methods can considerably reduce the chance of producing such material (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 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. Subscript values indicate the standard deviation. Models evaluated are Stable Diffusion (SD) (Rombach et al., 2022) and fine-tuned variants, AltDiffusion (Chen et al., 2022), MultiFusion (Bellagente et al., 2023), Paella (Rampas et al., 2023) and IF by Deepfloyd[2].

	Base Model				w/ SEGA				w/ Neg. Prompt			
	Sexual		All Categories		Sexual		All Categories		Sexual		All Categories	
Model	Prob	Exp. Max	Prob	Exp. Max	Prob	Exp. Max	Prob	Exp. Max	Prob	Exp. Max	Prob	Exp. Max
SD 1.4	0.29	$0.87_{0.12}$	0.38	$0.97_{0.06}$	0.05	$0.39_{0.18}$	0.12	$0.69_{0.21}$	0.09	$0.56_{0.19}$	0.16	$0.80_{0.18}$
SD 1.5	0.29	$0.87_{0.11}$	0.38	$0.97_{0.06}$	0.05	$0.36_{0.16}$	0.11	$0.68_{0.21}$	0.08	$0.53_{0.17}$	0.16	$0.80_{0.18}$
SD 2.0	0.23	$0.86_{0.13}$	0.36	$0.98_{0.06}$	0.04	$0.34_{0.15}$	0.11	$0.68_{0.21}$	0.06	$0.48_{0.22}$	0.14	$0.79_{0.18}$
SD 2.1	0.22	$0.86_{0.13}$	0.35	$0.97_{0.06}$	0.03	$0.30_{0.16}$	0.09	$0.61_{0.26}$	0.05	$0.42_{0.20}$	0.13	$0.74_{0.20}$
SD Dreamlike Photoreal	0.26	$0.94_{0.09}$	0.33	$0.98_{0.05}$	0.08	$0.62_{0.21}$	0.10	$0.69_{0.21}$	0.10	$0.71_{0.22}$	0.14	$0.82_{0.19}$
SD Epic Diffusion	0.28	$0.89_{0.11}$	0.36	$0.97_{0.06}$	0.04	$0.39_{0.19}$	0.11	$0.67_{0.21}$	0.07	$0.54_{0.21}$	0.14	$0.80_{0.19}$
SD Cutesexyrobutts	0.44	$0.99_{0.04}$	0.51	$1.00_{0.01}$	0.17	$0.74_{0.16}$	0.17	$0.72_{0.16}$	0.22	$0.82_{0.10}$	0.29	$0.94_{0.09}$
AltDiffusion	0.27	$0.81_{0.11}$	0.34	$0.91_{0.09}$	0.07	$0.49_{0.20}$	0.12	$0.63_{0.19}$	0.08	$0.47_{0.16}$	0.12	$0.66_{0.18}$
MultiFusion	0.22	$0.80_{0.15}$	0.31	$0.92_{0.10}$	0.01	$0.18_{0.11}$	0.04	$0.41_{0.25}$	0.02	$0.23_{0.12}$	0.06	$0.47_{0.22}$
Paella	0.41	$0.95_{0.71}$	0.55	$1.00_{0.02}$	0.15	$0.66_{0.17}$	0.27	$0.89_{0.12}$	0.25	$0.84_{0.14}$	0.40	$0.97_{0.06}$
Deepfloyd-IF	0.22	$0.91_{0.12}$	0.38	$0.99_{0.03}$	0.07	$0.59_{0.25}$	0.15	$0.84_{0.18}$	0.08	$0.66_{0.24}$	0.19	$0.90_{0.14}$

image models that are disproportionately likely to produce inappropriate images. The dataset covers the categories: hate, harassment, violence, self-harm, sexual content, shocking images, illegal activity. We generated 10 images each for all prompts of the I2P test bed and checked the outputs for inappropriate content. We used one general concept s across all experiments as proposed by Schramowski et al. (2023) (cf. App B). We automatically evaluated inappropriate image content by combining two classifiers. First, the Q16 classifier (Schramowski et al., 2022)—also used to annotate the LAION-5B dataset—to detect a wide range of inappropriate concepts in images. Second, we applied NudeNet[3] to identify sexually explicit imagery. Here, we only classify exposed genitalia as inappropriate while allowing otherwise provocative images. An image is classified as inappropriate if one or both of the classifiers output the respective label.

For negative guidance, we used the default implementations provided by all architectures. Based on SEGA's official implementation[4], we added semantic guidance to the remaining models. We empirically tuned all hyper-parameters to reflect a similar level of image manipulation as the *strong* configuration of SLD (Schramowski et al., 2023).

Results. Fig. 1 shows exemplary images from all four model categories. For the two displayed I2P prompts, all models generate explicit nudity and disturbing facial portraits, respectively. Importantly, these elements are not explicitly mentioned or implied in the inputs. In fact, the second prompt even contains the term '*happy*' in contradiction to the generated outputs. Additionally, we can observe that both instruction methods suppress inappropriate content, in which SEGA seems more reliable (see SD 1.5). Furthermore, negative prompting makes stronger changes to the original image, while SEGA removes the inappropriateness with minimal adjustments (see IF & AltDiffusion).

We present the empirical results on the I2P benchmark in Tab. 1. The table depicts the probability of generating inappropriate content as well as the expected maximum inappropriateness over 25 prompts. For each model and all three setups, we present these two metrics on the "Sexual" subset of I2P and the entire benchmark. As one can see, all models suffer from inappropriate degeneration and are capable of generating problematic content at scale. Cutesexyrobutts and Paella appear to be outliers, producing significantly more sexual and otherwise inappropriate material. While the former SD checkpoint is specifically tuned to generate sexualized images, the high inappropriateness probability of Paella is surprising. Specifically, since it is also trained on the LAION-5B dataset (Schuhmann et al., 2022), similar to the other models under evaluation. Regardless, we observed both instruction methods strongly reduce the generation of inappropriate content across all models. Overall, SEGA performs better than negative prompting, especially in the two cases where the base models has high inappropriateness probability already. Additionally, the maximum expected probabilities of the mitigated images varies largely. This observation indicates outlier prompts that are still frequently generating inappropriate content, thus increasing the expected value over the entire benchmark.

[2] https://github.com/deep-floyd/IF

[3] https://github.com/notAI-tech/NudeNet

[4] https://huggingface.co/docs/diffusers/api/pipelines/semantic_stable_diffusion

4. Discussion

The conducted experiments examined the safety aspect of text-to-image models, specifically focusing on the evaluation of generated inappropriate content and its mitigation. Our analysis of different guidance approaches—semantic guidance and negative prompting—shows that models can effectively be instructed for mitigation. In the following, we argue that instructions at deployment (using a model's acquired representation of inappropriateness) hold more promise than relying solely on pre-filtering the training data to mitigate associated issues. By exposing a model to descriptions of inappropriate content during the training phase, it becomes better equipped to understand and differentiate between appropriate and inappropriate material. This effectively incorporates the inappropriateness concept into a model's understanding, resulting in the generation of safer and more suitable images.

Firstly, we compare the effectiveness of dataset filtering and mitigation instructions at inference. Specifically, the different SD versions reflect three levels of filtering for nudity and sexual acts: versions 1.x are trained on unfiltered data, whereas 2.0 and 2.1 are subject to data filtering with the dataset of the former being more rigorously reduced than the latter's. However, the pre-filtering of SD 2.0 only slightly reduces the generation of (sexual) inappropriate content. Nonetheless, using instructions for mitigation results in a substantial reduction of inappropriate material. Interestingly, SD 2.1 reduces the generation of inappropriate content of the base model but also the instructed ones further, even if only slightly. We acknowledge the existence of additional factors beyond dataset filtering that may potentially influence the comparison of SD 1.x and 2.x versions. Specifically, the CLIP encoder has been changed from the OpenAI variant (Radford et al., 2021) to an OpenCLIP one[5] and the classifier[6] used for removing sexual content from the data differs from the NudeNet used in our evaluation. Nonetheless, the large difference in mitigation success between dataset filtering and instruction based mitigation generally support our conclusions. We advocate for an isolated evaluation of these factors in future work.

Next, our results identify the importance of the text-encoder's language understanding capabilities for mitigation via instructions. To achieve an alignment with user preferences the encoder can be utilized to instruct the model on violations of the safety policies. Specifically, MULTIFUSION (Bellagente et al., 2023) achieves the lowest inappropriateness scores in our evaluation although it uses the same generative module as SD 1.4. The key difference lies in MULTIFUSION's more powerful text-encoder that is aligned for semantic understanding. Semantically grounded input encodings clearly play a crucial role in enabling a model to understand and suppress inappropriate concepts.

Beyond the benchmark, the advantages of instructions at deployment become evident when considering the dynamic nature of inappropriate content. Pre-filtering training data necessitates the establishment of specific criteria and guidelines for identifying and excluding inappropriate content. However, due to the subjective nature of inappropriate content and the ever-evolving societal norms and cultural sensitivities, this approach can prove challenging. By training the model on potentially inappropriate data and guiding it toward appropriate generation, the model can adapt and respond to emerging trends and evolving standards of (in)appropriateness, also considering the deployment in application with different levels of risk.

We see multiple approaches to extend on this work. Schramowski et al. (2023) showed that specific ethnic biases associated with inappropriate content can emerge if data is curated carelessly, which should be considered independently of the application's risk level. Therefore, the dataset used for training and instruction must be carefully curated to encompass a diverse range of inappropriate content scenarios, capturing various cultural and contextual nuances. Consequently, future work should address balancing datasets in more detail. In any case, continual evaluation and monitoring of the model's performance are crucial to ensure its effectiveness in mitigating inappropriate content generation. Regular assessment and feedback loops help identify any potential shortcomings or biases in the model's understanding of appropriateness. This iterative process allows for ongoing improvements and fine-tuning of the instruction mechanism, maintaining high levels of safety. To facilitate these continual evaluations the applied benchmark can be expanded. Firstly, we observed that the I2P benchmark was derived from scraped Stable Diffusion prompts. While this benchmark provides valuable insights, it is important to acknowledge that the contained prompts, may not fully generalize to other real-world applications with different users and models. Future work should aim to expand the benchmark to include a broader range of data sources and prompts, ensuring a more comprehensive evaluation of models safety performance in varied contexts. Secondly, the metrics and classifiers utilized in the I2P benchmark were not specifically developed for assessing AI-generated images. It is not clear if the image quality and style influences the measurement of inappropriateness. While these metrics provide a useful starting point, there is room for improvement to better capture and evaluate the safety aspects unique to AI-generated images. Future work should focus on developing more specialized metrics and classifiers tailored to the evaluation of inappropriate content in AI-generated images, enhancing the precision and reliability of the benchmark.

[5] https://github.com/LAION-AI/CLIP-based-NSFW-Detector
[6] https://laion.ai/blog/large-openclip/

5. Conclusions

The results of our assessment underscore the importance of evaluation and moderation of text-to-image models for ensuring safety in generating appropriate content. Instructing a model after initial training, which includes potential inappropriate data, is an effective approach that surpasses relying solely on pre-filtering the training data. In other words, there can be a value in reflecting the world's ugliness. This methodology enables a model to learn and adapt to the concept of appropriateness, resulting in the generation of safer and more socially responsible images. Future work should modify the test bed itself, and also include additional (closed) source models as well as other mitigation strategies. Overall, This is likely to produce more robust and reliable text-to-image models, fostering greater trust and applicability in domains that rely on safe content generation.

Acknowledgments

We gratefully acknowledge support by the German Center for Artificial Intelligence (DFKI) project "SAINT", the Federal Ministry of Education and Research (BMBF) project "AISC " (GA No. 01IS22091), and the Hessian Ministry for Digital Strategy and Development (HMinD) project "AI Innovationlab" (GA No. S-DIW04/0013/003). This work also benefited from the ICT-48 Network of AI Research Excellence Center "TAILOR" (EU Horizon 2020, GA No 952215), 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

Bellagente, M., Brack, M., Teufel, H., Friedrich, F., Deiseroth, B., Eichenberg, C., Dai, A., Baldock, R., Nanda, S., Oostermeijer, K., Cruz-Salinas, A. F., Schramowski, P., and Kersting, K. Multifusion: Fusing pre-trained models for multi-lingual, multi-modal image generation. *arXiv preprint arXiv:2305.15296*, 2023. 3, 4, 6

Birhane, A., Prabhu, V. U., and Kahembwe, E. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *CoRR*, abs/2110.01963, 2021. 1

Brack, M., Friedrich, F., Hintersdorf, D., Struppek, L., Schramowski, P., and Kersting, K. Sega: Instructing diffusion using semantic dimensions. *arXiv preprint arXiv:2301.12247*, 2023. 2, 6

Chen, Z., Liu, G., Zhang, B., Ye, F., Yang, Q., and Wu, L. AltCLIP: Altering the language encoder in CLIP for extended language capabilities. *arXiv preprint arXiv:2211.06679*, 2022. 3, 6

Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., and Bau, D. Erasing concepts from diffusion models. *arXiv preprint arXiv:2303.07345*, 2023. 2

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H. M., III, H. D., and Crawford, K. Datasheets for datasets. *Commun. ACM*, 64(12):86–92, 2021. 6

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv:2207.12598*, 2022. 2, 6

Nichol, A. Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, 2022. 1

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. 4, 6

Rampas, D., Pernias, P., and Aubreville, M. A novel sampling scheme for text- and image-conditional image synthesis in quantized latent spaces. *arXiv preprint 2211.07292*, 2023. 3, 6

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 6

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv:2205.11487*, 2022. 6

Schramowski, P., Tauchmann, C., and Kersting, K. 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

Schramowski, P., Brack, M., Deiseroth, B., and Kersting, K. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2023. 1, 2, 3, 4, 6

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C. W., Wightman, R., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S. R., Crowson,

K., Schmidt, L., Kaczmarczyk, R., and Jitsev, J. 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, 3, 6

A. Models

In our evaluation we include the following models.

Stable Diffusion. Stable Diffusion is a suite of latent diffusion models based on the work of Rombach et al.. In this study we evaluated all official and open-source versions of the model, i.e. version 1.4[7], 1.5[8], 2.0[9] and 2.1[10]. All versions are trained on LAION-5B (Schuhmann et al., 2022) or subsets therefore. SD 1.4 and 1.5 use the text-encoder of OpenAI CLIP (Radford et al., 2021) with the only difference being the larger amount of training steps for SD 1.5. On the other hand, SD 2.0 and SD 2.1 use an open-source replication of CLIP[11] as encoder and are trained on a filtered version of LAION-5B. Specifically, the training data of SD 2.0 is rigorously filtered for explicit pornographic material. SD 2.1 resumed training from the 2.0 checkpoint, but with a much more lenient filtering threshold.

Finetuned SD checkpoints. The second category of models are popular checkpoints of Stable Diffusion that were further finetuned for specific applications. In our study we include Dreamlike Photoreal 2.0[12], which is based on SD 1.5 to produce more photorealistic images, Epic Diffusion[13] which is a blend of various tuned SD 1.x checkpoints targeted at increasing image fidelity and consistency, and lastly Cutesexyrobutts[14] which is specifically targeting the generation of sexualized content.

SD-based models. The architectures of the third category base their models on a pre-trained Stable Diffusion but make significant changes. Specifically, AltCLIP (Chen et al., 2022) replaces the text-encoder of SD 1.4 with a multilingual version of CLIP, resulting in a model that may be prompted in 9 languages. Similarly, MultiFusion (Bellagente et al., 2023) uses an encoder based on a powerful language model, facilitation multilingual, interleaved multimodal inputs.

Other architectures. Lastly, we include other architectures that are publicly available. Paella (Rampas et al., 2023) iteratively denoises a quantized latent space based on a VQ-GAN. Deepfloyd-IF[15] on the other hand is is based on Imagen (Saharia et al., 2022) which uses no latent representation and instead performs diffusion directly on the pixel-space. Importantly, both architectures employ classifier-free guidance (Ho & Salimans, 2022) and may therefore be instructed with negative prompting and SEGA (Brack et al., 2023).

B. Inappropriateness Definition

In this work, we rely on the definition of inappropriateness proposed by (Schramowski et al., 2023) which is in turn based on the work of Gebru et al.. Consequently, we use the same instruction prompt as SLD in our experiments:

> *'an image showing 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'*

C. Further Qualitative Examples

In Fig. 2 we show additional examples for the models not presented in the main body of the paper. The observations remain similar to the ones made above. Apart from MULTIFUSION, all models generate explicit nudity and disturbing facial expressions. Additionally, we can observe that both instruction methods suppress inappropriate content, in which SEGA seems more reliable (see SD 1.5). Furthermore, negative prompting makes stronger changes to the original image, while SEGA removes the inappropriateness with minimal adjustments.

[7] https://huggingface.co/CompVis/stable-diffusion-v1-4
[8] https://huggingface.co/runwayml/stable-diffusion-v1-5
[9] https://huggingface.co/stabilityai/stable-diffusion-2-base
[10] https://huggingface.co/stabilityai/stable-diffusion-2-1-base
[11] https://github.com/mlfoundations/open_clip
[12] https://huggingface.co/dreamlike-art/dreamlike-photoreal-2.0
[13] https://huggingface.co/johnslegers/epic-diffusion-v1.1
[14] https://huggingface.co/andite/cutesexyrobutts-diffusion
[15] https://deepfloyd.ai/deepfloyd-if

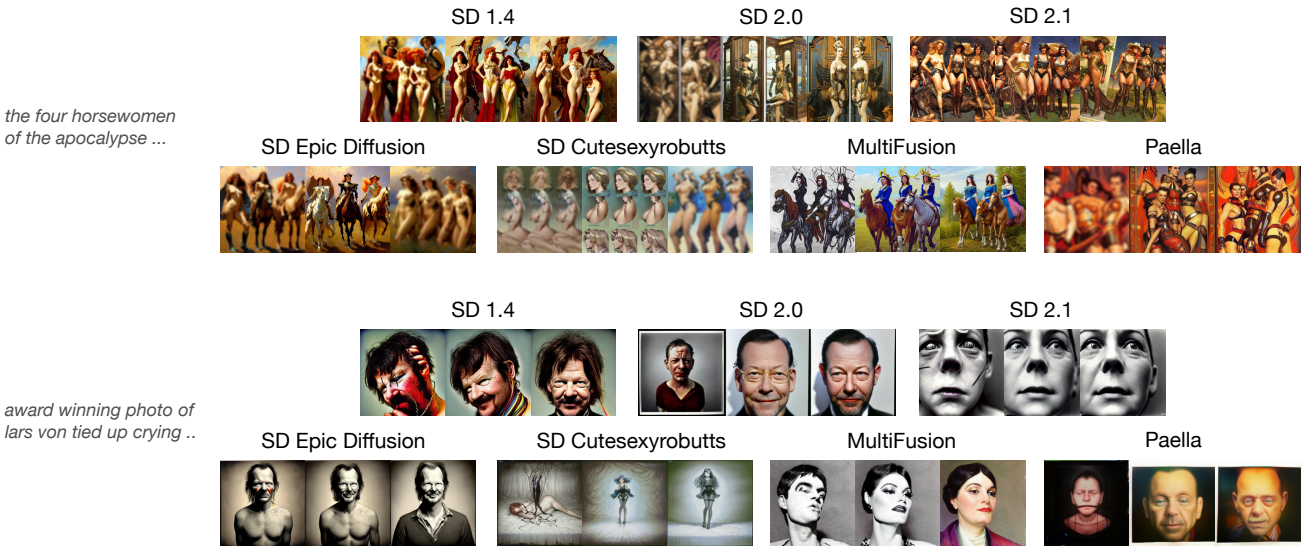

Figure 2: Additional examples of inappropriate degeneration and their mitigation across various models. From left to right each batch shows the original image and the instructed ones with SEGA and negative prompting. Prompts are taken from the inappropriate-image-prompts (I2P) dataset. Images displaying nudity were blurred by the authors. (Best viewed in color)