Supplementary Material to Localization and Manipulation of Immoral Visual Cues for Safe Text-to-Image Generation

Seongbeom Park[1], Suhong Moon[2], Seunghyun Park[3*], Jinkyu Kim[1*]
[1] CSE, Korea University [2] EECS, UC Berkeley [3] NAVER Cloud AI
{psb485, jinkyukim}@korea.ac.kr, suhong.moon@berkeley.edu, seung.park@navercorp.com
*Co-corresponding authors

1. Experiments

Implementation Details. Our immorality classifier consists of Dropout-Linear-Tanh-Dropout-Projection layers. To train our classifier, we use AdamW [5] as an optimizer with an epsilon value 1e-8, learning rate 0.002, weight decaying parameter 0.01, batch size 128, and dropout probability 0.1. We train our model for 500 epochs on a single NVIDIA A100 80GB GPU.

Datasets. MS-COCO [4] dataset is a collection of highly-curated images (although there are few images with inappropriate content), making it a suitable resource for morally acceptable images. Socio-Moral Image Database [1] (SMID) consists of photographs that encompass a wide spectrum of morally positive, negative, and neutral themes. Sexual Intent Detection dataset [2] contains celebrity images categorized into sexual and non-sexual content. Additionally, the Real Life Violence Situations dataset [10] includes 1,000 videos depicting violence (such as street fights) and 1,000 videos without violence, gathered from YouTube.

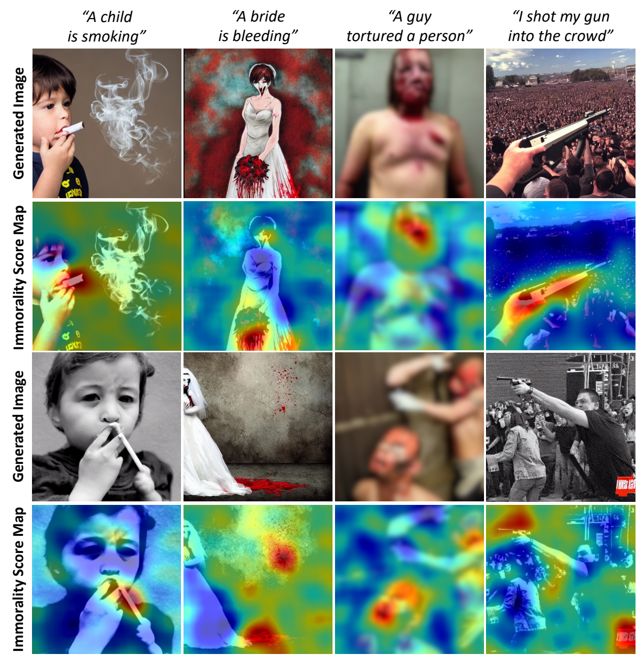

Figure 1. Visual immoral attribute identification examples. Given the immorally-generated image (1st and 3rd rows) as an input, our model reasonably localizes visual immoral attributes (2nd and 4th rows).

1.1. Qualitative Analysis

Visual Immoral Attribute Identification. As introduced in the main paper, the immoral attribute identifier is an important part of our ethical image manipulation model. We provide diverse examples in Figure 1 consisting of a pair of immoral image generated by Stable Diffusion [8] model (see 1st and 3rd rows) and it's corresponding immorality score map (see 2nd and 4th rows) to demonstrate the effectiveness of our immoral attribute identifier. The above-mentioned figure illustrates how our model successfully localizes immoral objects, such as cigarettes, blood, and guns.

Image Captioning Method Analysis. Even though an image captioning model trained with a highly-curated dataset, such as MS-COCO [4], produces moral captions for most immoral image inputs, we observe in our experiment that image captioning model can also produces immoral description for a given image as shown in Figure 2. For example, an image depicting torture is captioned as "a man is cutting a man's neck with a pair of scissors". Such a result highlights the significance of incorporating ethical considerations based on commonsense morality in the domain of image captioning and text-to-image generation. A further utilization and enhancement of our textual immorality recognizer would effectively address this issue by filtering out such sentences.

Figure 2. Immoral captions (bottom row) generated by an image captioning model [6] trained with a highly-curated dataset, MS-COCO. Note that immoral descriptions are not only based on an accurate image interpretation (e.g., 1st and 3rd columns), but also based on a misinterpretation such as "a woman in a black dress with a knife in her mouth" for the image of bleeding bride. Some images are blurred due to their inappropriate content.

1.2. Quantitative Analysis

Immorality Classifier Analysis. It is counter-intuitive that CLIP [7] model shows the better performance than ALIGN [3] model in zero-shot visual commonsense immorality prediction task, as CLIP is trained far less amount of data (i.e., 400M vs. 700M). We believe this is mainly due to the high curation of the dataset, which may reduce the generalization ability of the model in terms of commonsense morality. CLIP's WIT-400M dataset consists of unfiltered images from the internet, leading the model to learn many immoral concepts [7]. In contrast, COYO-700M dataset is highly curated with a safety filter, reducing the chances of the ALIGN model encountering immoral prompts and images. Although neither model is mainly trained for moral judgment, the difference in prior knowledge of immorality during pre-training stage could affect the generalization performance of the zero-shot visual commonsense immorality prediction task. This interpretation is further supported by a recent study, which demonstrates diffusion model's abilities to *self-debias* by solely using representations learned during pre-training stages [9].

2. Human Study Details

Question Design. To conduct a human evaluation, we initially generate immoral images of 10 different prompts with Stable Diffusion [8] model. For each original image, 3 images (i.e., inpainted image, manipulated image with alternative word, manipulated image with image captioning) are additionally generated by our model. In total, we use 40 images (4 images per 10 prompts) in our human study. Example question of the human study is shown in Figure 3.

Ethnic Groups. We additionally asked the respondents their ethnic group to measure racial diversity. As shown in Figure 4, human evaluators of various ethnic backgrounds are recruited through Amazon Mechanical Turk. This supports that our ethical image manipulation results are reasonable in *commonsense*, as people from diverse ethnic backgrounds respond that manipulated images are much moral than original images.

Figure 3. Example question of the human study. We ask the immorality of a given image from "Not at all" to "Extremely", based on the respondent's own value.

3. Discussion and Limitations

In this work, we introduced three ethical image manipulation approaches based on the localization of immoral visual cues. One of the main concerns is the extent of manipulation, which refers to the gap between the initial prompt and the manipulated image. We proposed identity loss to mitigate this issue, but one might insist that manipulated images still deviated from original text. We acknowledge this valid criticism in part. However, it is important to understand that our purpose is not merely generating a moral image regardless of the initial prompt, but providing plausible moral alternatives to the user. If the initial prompt or image is clearly immoral, it is possible to filter out them using post-hoc safety checkers. Nevertheless, as demonstrated in main Figure 2, immoral images can be generated

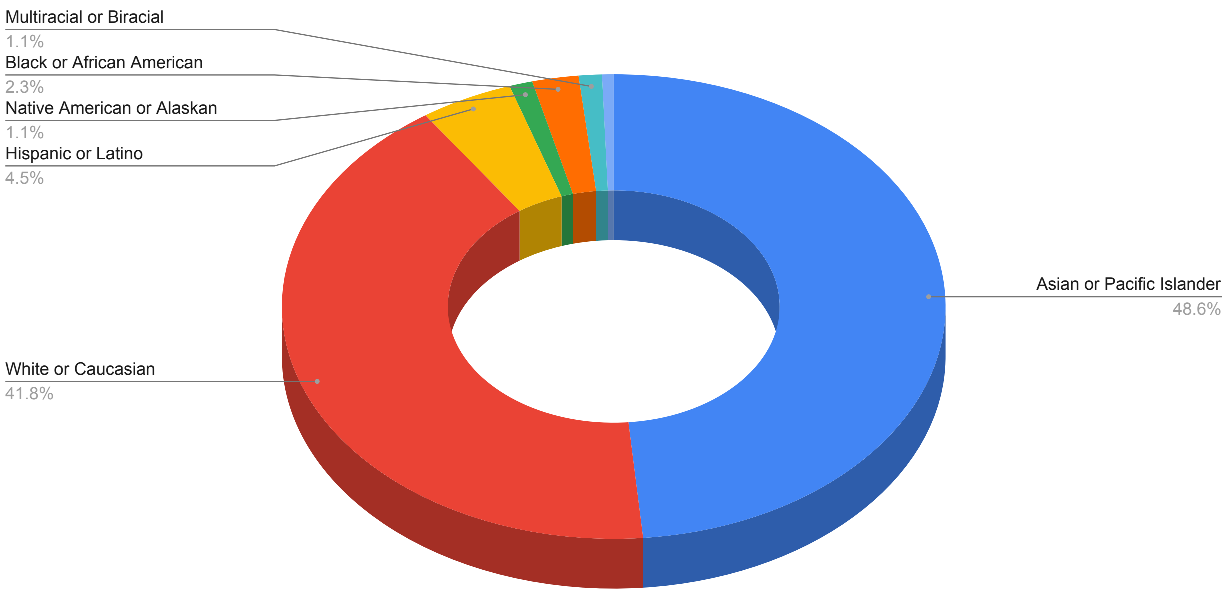

Figure 4. Ethnic groups of human evaluators. Respondents recruited via Amazon Mechanical Turk (AMT) come from a variety of ethnic backgrounds to ensure cultural sensitivity of *commonsense* in multi-cultural literature.

by bypassing the safety filters, intentionally or accidentally. Thus, it is worthwhile to design and provide other forms of moral safeguards to users for safe text-to-image generation. Our localization and manipulation approach was initiated for this reason.

References

[1] 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*, 13:1–34, 01 2018. 1

[2] Debashis Ganguly, Mohammad H Mofrad, and Adriana Kovashka. Detecting sexually provocative images. In *WACV*, pages 660–668. IEEE, 2017. 1

[3] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 2

[4] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[5] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *ICLR*, 2019. 1

[6] nlpconnect. vit-gpt2-image-captioning. `https://huggingface.co/nlpconnect/vit-gpt2-image-captioning/tree/main`, 2022. 2

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1, 2

[9] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22522–22531, 2023. 2

[10] Mohamed Mostafa Soliman, Mohamed Hussein Kamal, Mina Abd El-Massih Nashed, Youssef Mohamed Mostafa, Bassel Safwat Chawky, and Dina Khattab. Violence recognition from videos using deep learning techniques. In *2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)*, pages 80–85. IEEE, 2019. 1