AUTOMATIC SYNTHESIS OF SAFE COUNTERFACTUALS FOR HARMFUL IMAGES USING LLMS AND DIFFUSION MODELS

Sebin Lee

A THESIS

in

Master in Computer and Information Science

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Master of Science in Engineering

2025

Supervisor of Thesis

Signature: _____

Chris Callison-Burch

Professor, Computer and Information Science

Co-Supervisor of Thesis

Signature: _____

Artemis Panagopoulou

Ph.D Candidate, Computer and Information Science

CIS Master's Program Director

Signature: _____

Swapneel Sheth, Practice Associate Professor, Computer and Information Science

Committee

Signature: _____

Lyle Ungar, Professor, Computer and Information Science

Chris Callison-Burch, Professor, Computer and Information Science

Artemis Panagopoulou, Ph.D. Candidate, Computer and Information Science

ABSTRACT

AUTOMATIC SYNTHESIS OF SAFE COUNTERFACTUALS FOR HARMFUL IMAGES USING LLMS AND DIFFUSION MODELS

Sebin Lee

Chirs Callison-Burch

Artemis Panagopoulou

Despite recent progress in multimodal learning, vision–language systems remain vulnerable to harmful biases embedded in training data. Image–text models, from CLIP to text-to-image generators, can amplify social stereotypes—depicting men in leadership roles, over-representing lighter skin tones, or reinforcing gendered occupational tropes. While prior work has proposed fairness-aware training objectives and bias-sensitive evaluation benchmarks, existing methods often fall short in addressing unsafe visual content directly. In particular, there is a lack of automated, scalable methods for generating semantically equivalent yet mitigated visual counterfactuals—images that preserve the core meaning of a scene while altering biased attributes. This thesis introduces a framework for identifying harmful visual patterns and synthesizing safe counterfactuals that differ only in protected attributes, leveraging recent advances in diffusion models and vision–language alignment. In addition to building this pipeline, the work provides empirical and conceptual insights into the complexity of visual safety through evaluation on different image safety datasets along with human annotation. We demonstrate its effectiveness in reducing biased feature correlations through data augmentation and category-specific processing which yields better results than general purpose approaches. Finally, the thesis highlights a fundamental challenge in the field: the lack of consensus across safety datasets, where inconsistent definitions and standards hinder the development of robust, generalizable mitigation methods.

TABLE OF CONTENTS

CHAPTER 1

Introduction

Disclaimer: This paper contains content that may be disturbing to some readers.

Growing use of generative AI, especially Vision-Language-Models, has increased the public use of synthetic images. This vast advancement and access made the creation and spreading of harmful contents easier, emphasizing the importance of developing method to mitigate the generation of harmful contents. (Ramesh et al., 2021)

However, VLMs suffer from safety challenges such as its lack of defense mechanism when tasked with generating adversarial or harmful images. (Liu et al., 2024) Also, combined with jailbreak attempts to bypass the ethics safeguard, using VLMs to generate adversarial images has become easier. (Jin et al., 2024; Qi et al., 2024) Commercial and open-source models are vulnerable to such attacks it is crucial to find method to address this problem especially considering its wide accessibility and how young population can easily get affected by such technology. (Dong et al., 2023)

Currently, most methods used for safe image generation are focused on post-processing steps after training model. One popular approach is employing post-hoc classifier such as Q16 (Schramowski et al., 2022) and SD filter (Rando et al., 2022) that filters harmful images. However, these classifiers may introduce bias itself, leading to disproportionate filtering of certain group or topics. One example is association of attributes related to feminine or non-masculine features more likely flagging NSFW classifiers. Because majority of sexual, inappropriate images collected on Internet display female, this data distribution is reflected on final trained models, associating the presence of female as "harmful" even when the image displays female doing standard activities. (Leu et al., 2024)

These undesirable content generation originates from the model learning harmful concepts as training dataset from the Internet contains content on adversarial contents. In this paper, we propose a scalable pipeline that detects and sanitizes unsafe images via harmful aspect detection and mit-

igation using Large Language Models and Diffusion Models. Our pipeline attempts to recreate images by removing its harmful content, while preserving other features to maximize information retainment that is useful for model training. This approach allows us to generate safe counterfactual images, effectively addressing harmful content. We leverage the capabilities of LLMs and Diffusion Models in our pipeline to exploit their ability to efficiently adapt pre-trained concepts, enabling the generation of multiple diverse pairs for a single unsafe image. Because our pipeline is explainable by nature, we can easily probe how the image is getting processed and refined, allowing transparent insights into the modifications and facilitating human intervention when necessary.

CHAPTER 2

Related Works

Several approaches create counterfactual images that preserve semantic content while altering biased attributes. (Zhang et al., 2024) propose GAMA, a two-stage vision-language model that generates gender-neutral narratives to reduce bias in downstream tasks. (Yang et al., 2024a) introduce a masking mechanism to suppress gender-specific information in image captioning. Safe-CLIP (Poppi et al., 2025) generates harmful textual counterfactuals for safe images and unsafe image counterfactuals for safe text, and finetunes CLIP to retrieve safe text for unsafe images, and unsafe images for safe text. While effective for harmful retrieval mitigation, this method poses the risk of inducing unnecessary noise in the semantic space of the model. Another line of work focuses on reducing model reliance on biased spurious correlations. For example, BiaSwap (Kim et al., 2021) uses class activation maps and image-to-image translation to swap out bias-relevant regions, generating bias-conflicting training examples. In a similar vain, BiasAdv (Lim et al., 2023) produces adversarial examples that confuse biased models but preserve core class semantics. Finally, DFA (Lee et al., 2021) augments training with feature-level counterfactuals that disentangle class and bias factors. While such methods have proven effective in reducing distribution biases in downstream tasks, they do not address the problem of eliminating biased content from training representations. More relevant to our approach, SocialCounterfactuals (Howard et al., 2024) leverages cross-attention control in diffusion models to generate intersectional counterfactuals (e.g., flipping race/gender while preserving scene context), producing high-fidelity, demographically balanced data. However, this method requires a prior knowledge of the attributes to flip, which is not always possible. Our method attempts to address this issue via the incorporation of LLMs for harmful content detection in images.

Fairness in Image Generation Text-to-image generation models, including Stable Diffusion, frequently reflect demographic and cultural stereotypes. (Jha et al., 2024) introduce the ViSAGe benchmark, showing that prompts like "a [nationality] person" yield highly stereotyped and sometimes offensive images, particularly for the Global South groups. Previous work has explored prompt

engineering (Yang et al., 2024b) or distribution-guided sampling to reduce such biases, but few methods explicitly generate counterfactual images to enable fair training or evaluation.

Diffusion-based methods have also emerged for debiasing generative models. FairDiffusion (Friedrich et al., 2023) introduces fairness-aware guidance to steer Stable Diffusion toward gender-balanced generations for occupation prompts. Distribution Guidance (Parihar et al., 2024) uses an attribute predictor in the latent space to balance demographic distributions during generation without altering the prompts. Studies also choose to sanitize the image by fine-tuning diffusion model on specific style or concept to remove those targeted aspects from the image. (Gandikota et al., 2023) Pinpoint Counterfactuals (Sirotkin et al., 2024) edit images using localized diffusion inpainting guided by text prompts, preserving all context while changing a protected attribute—key for fairness evaluation and training. Yet, such methods typically require apriori specification of protected attributes.

Evaluation Benchmarks and Fairness Testing A growing body of work aims to evaluate how biased models behave when presented with counterfactual inputs. (Qiu et al., 2023) show that popular vision-language evaluation metrics like CLIPScore can be biased toward stereotypical content, while traditional n-gram metrics remain more neutral. (Jha et al., 2024) present ViSAGe, a benchmark exposing nationality-based stereotypes in T2I models, highlighting the need for fairness-aware image generation. In vision, ConBias (Chakraborty et al., 2024) introduces concept graphs to detect and rebalance biased co-occurrence patterns, and OpenBias (D'Inca et al., 2024) uses LLM+VQA pipelines to automatically discover model biases across open sets of prompts and outputs. These efforts illustrate the increasing need for high-quality, semantically controlled counterfactuals—a gap our method aims to address by providing automated, scalable generation of safe image variants for bias analysis and mitigation.

CHAPTER 3

Methodology

In this work, we present an automated, scalable text-image editing framework that that extracts and mitigates harmful content and biases. The overall process consists of three main components:

1. **Image captioning**: To harness the knowledge of LLMs of biases, we generate a detailed caption of the image.

2. **Harmful Content Detection and Mitigation**: Given the image caption, an LLM is probed to identify harmful content and biases in the image, and then rewrite a 'debiased caption' that maintains the safe semantics of the image, without the unwanted content.

3. **Safe Image Recreation**: A semantically similar yet safe image is created via a diffusion model conditioned on both the caption and the original image.

We experiment with two variations on the vanilla framework: Chain of Thought Prompting (Wei et al., 2022) and Inpainting (Yu et al., 2018). The prompts used for each step are available in the Appendix.

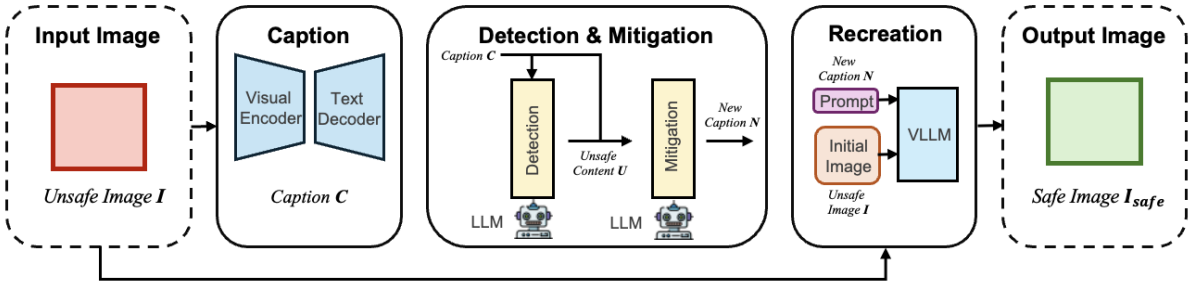


Figure 3.1: Overview of our framework that takes three-step procedure to build safe image from unsafe image

3.1. Baseline Procedure

We represent each unsafe image as $I \in R^{h \times w \times c}$ where h, w, and c refers to the height, width, and number of color channels, in our case RGBA.

Step 1. Caption We first feed the image I to function $f_{caption}$ through caption model to gain caption C, which is sequence of words w_i:

$$\begin{aligned} C &= f_{caption}(I) \\ &= (w_1, w_2, ...w_n) \end{aligned} \tag{3.1}$$

The generated sequence $(w_1, w_2, ...w_n)$ corresponds to a textual description of the image. This transformation helps to leverage large language models into identifying problematic aspects of the image. In addition, the decision to operate in text space for the intermediate steps of the image recreation pipeline allows to leverage transformed captions as input for diffusion models.

Step 2. Detection and Mitigation In this step our goal is to create a safe caption N that is semantically close to the original caption C. We first construct a harmful content object U by passing caption C as input to a harmful content detection function $f_{detection}$. The function $f_{detection}$ is implemented as an LLM prompted accordingly to extract harmful content. Then we define another function $f_{mitigation}$, also implemented as an LLM, tasked with reconstructing the original caption C but without the harmful content in U. This function naturally takes both C and U as inputs and returns the safe caption N. Formally, we present this procedure as follows:

$$U = f_{detection}(C) \tag{3.2}$$

$$N = f_{mitigation}(C, U) \tag{3.3}$$

Step 3. Recreation Finally, we recreate the image without harmful content. We define a function $f_{recreate}$ implemented as a diffusion model which takes as input both the original image I and the

new caption N and generates the safe counterfactual image $I_{safe} \in R^{h \times w \times c}$. Formally, this is represented as follows:

$$I_{safe} = f_{recreate}(I, N) \tag{3.4}$$

3.2. Chain-of-Thought

We employ a Chain-of-Thought (CoT) (Wei et al., 2022) approach to improve the LLM's performance in understanding which harmful method to mitigate in Step 2. Because CoT reasoning generates intermediate reasoning steps, we leverage this rationale to better align our goal of generating counterfactuals while preserving safe semantics by instructing the LLM to target specific harmful phrases of the initial caption C in the mitigation step.

In this variant of our method we apply another function f_{phrase} after $f_{detection}$. The function f_{phrase} is implemented as an LLM and is tasked to find harmful phrases P in C that contain U. In this variant, we adapt the mitigation function to $f'_{mitigation}$ which also considers the harmful phrases in constructing the safe function N. Formally, this variant can be summarized as follows:

$$\begin{aligned} P &= f_{phrase}(C, U) \\ &= (p_1, p_2, \ldots p_n) \\ p &= (w_i, w_{i+1}, \ldots, w_j) \end{aligned} \tag{3.5}$$

$$N = f'_{mitigation}(C, U, P) \tag{3.6}$$

We go through same process as baseline for Step and output final image I_{safe}.

3.3. Inpainting

We also explore a variant of Step 3, where we employ inpainting for safe image recreation. Image inpainting (Yu et al., 2018) is a method that edits only a specified region of an image using diffusion. The intuition behind exploring this method is that it could potentially lead to better preservation of the semantics of the image by only applying targeted edits on its harmful aspects.

The inpainting variant of our method operates as follows: we leverage a different prompt in Step 1

to generate an object-based caption C' that prompts the model to focus on individual objects of the image. To generate object caption C', we use function f_{object} through the image captioning model to output a list of objects O comprising a series of object o depicted in image I.

$$\begin{aligned} O &= f_{object}(I) \\ &= (o_1, o_2, \dots o_n) \end{aligned} \tag{3.7}$$

For each object o_i, we input it to function $f_{describe}$ implemented as caption model tasked to output description d_i of each object. As a result we collect descriptions D. Using these outputs, we construct object caption C'.

$$\begin{aligned} D &= f_{describe}(I, O) \\ &= (d_1, d_2, \dots d_n) \end{aligned} \tag{3.8}$$

$$C' = \{(o_i, d_i) | o_i \in O, d_i \in D\}. \tag{3.9}$$

Next, we define additional factor $O'_{initial}$ by defining function f_{find} after $f_{detection}$. Function f_{find} is implemented as LLM to find series of objects o' that contains unsafe contents based on defined U. Note that o' and o do not necessarily overlap.

$$\begin{aligned} O'_{initial} &= f_{find}(C, U) \\ &= (o'_1, o'_2, \dots, o'_n) \end{aligned} \tag{3.10}$$

Using O', we get sanitized objects S using function $f'_{mitigate}$, implemented as an LLM instructed to to (1) exclude $o' \notin O$ to prevent hallucination and (2) get rewritten description d' with featured U removed.

$$\begin{aligned} S &= f'_{mitigate}(C, U, O') \\ &= \{(o'_i, d'_i) | o'_i \in O'_{final}\} \end{aligned} \tag{3.11}$$

Finally, the image recreation process is adapted to incorporate individual objects as follows: we define mask M using function f_{mask} implemented as object detector model to find featured region of each object on given image I. We generate mask for each object to limit the region that function $f_{inpainting}$ gets activated, enabling gradual changes on I for each object.

$$\begin{aligned} M &= \{(o'_i, m_i) | o'_i \in S\} \\ m_i &= f_{mask}(I, o'_i) \end{aligned} \tag{3.12}$$

Employment of function $f_{inpainting}$ tasks a diffusion model to recreate image I' by retrieving corresponding d'_i from S and updating the image sequentially for every o'_i and m_i in M.

$$I' = f_{inpainting}(I, M, S) \tag{3.13}$$

CHAPTER 4

Evaluation

Using the framework we proposed in the previous section, we now study how effectively our pipeline can generate counterfactauls of inputted unsafe images. We explore two avenues for such an evaluation.

1. **Downstream Performance**: We substitute a portion of safe images in the training set of a classifier with data generated through our pipeline and evaluate performance.

2. **Human Evaluation**: We conduct a human evaluation to judge the quality of the generated images.

4.1. Models and Dataset

To evaluate our proposed framework, we experiment with UnSafeBench (Qu et al., 2024), a comprehensive image dataset for image safety research. The dataset consists of 8,110 training images out of which 4,048 are classified as unsafe, and 2,040 test images on which we evaluate downstream performance. These images are grouped into 11 different safety categories and include both natural and synthetic images. We exclude the "Spam" category from our analysis because it primarily contains text-heavy images. Diffusion models struggle to modify embedded text, which would introduce noise and obscure the impact of our detection and mitigation steps, leading to a disproportionate number of failure cases due to model limitations. Moreover, to evaluate out-of-domain robustness, we use UnSafeDiffusion(Qu et al., 2023), a synthetic image dataset consisting of harmful images generated by various text-to-image editing tools, for downstream evaluation.

We implement the various components of our pipeline using 4 high-performing open-source models: Molmo-7B (Deitke et al., 2024) for captioning, OLMo-2-13B (Groeneveld et al., 2024) for bias detection and mitigation, and stable-diffusion-xl-refiner-1.0 (Podell et al.) for image recreation. For the inpainting branch of our pipeline, we additionally incorporate GroundingDINO for open-set object detection (Liu et al., 2023). The LLM and VLM are served via vllm Kwon et al. (2023) and

the diffusion and detector models using huggingface.[1] Experiments are conducted on 2-A6000 40GB GPUs.

4.2. Evaluations

We compare the performance of a classifier trained with the original UnsafeBench training data, and a variant training dataset consisting of unsafe images from UnSafeBench and safe images recreated using our pipeline. We randomly sample 1000 images labeled "Unsafe" from train split of UnsafeBench and create sanitized version of those 1000 images for Our Pipeline Dataset. For comparison, we grab 1000 random safe images from train split of UnsafeBench, using total 2000 images in each case. Inspired by previous studies that show benefits in model robustness with mixed use of synthetic and real images(Singh et al., 2024), we also propose an Augmentation setting. In this setting, we add subset of images from our pipeline to original UnSafeBench and use it as a train dataset.

To evaluate, we use CLIP-ViT image encoder to gain image features of all train dataset. Then, we train linear regression classifier using these image features. We sample 700 images from test split of UnsafeBench to evaluate the classifier performance by calculating accuracy and F1 score.

Evaluation 1: Performance by Categories We conduct experiments in 3 different prompting strategies: Zero-shot, Few-shot and Few-shot by category. In Zero-shot setting, we directly instruct LLM to find bias from the caption and rewrite the caption. In Few-shot setting, we provide 4 examples both to the bias detection and bias mitigation steps. Few-shot by Category contains tailored examples based on the category of image, which is one of 10 different safety categories defined by UnSafeBench. We also evaluate the possibility of our pipeline providing a more diverse safe image set that the classifier can learn from by comparing the score of a model trained with augmented dataset where we add certain percent of safe images from our pipeline to UnSafeBench.

Evaluation 2: Out-Of-Distribution Evaluation Because our pipeline incorporates a diffusion model in the final generation of the safe counterfactual to the original image, all of safe images

[1] https://huggingface.co

generated are inherently synthetic images, which could potentially result in an imbalanced distribution of features compared to other classifiers trained on both real and synthetic data. As a result, we incorporate an Out-Of-Distribution evaluation by comparing the performance of classifiers on UnSafeDiffusion(Qu et al., 2023).

Evaluation 3: t-SNE Analysis We perform a t-SNE visualization analysis using both the training datasets and the resulting misclassified images. First, we employ CLIP (Radford et al., 2021) Image Encoder to get Image Embeddings of target images and visualize them. For visualization, we employ t-Distributed Stochastic Neighbor Embedding (t-SNE) (Van der Maaten and Hinton, 2008), a widely used non-linear dimensionality reduction technique. We keep the number of dimensionality as 2 and perplexity as 10 for all visualizations. With the t-sne graph, we perform two analyses:

1. **Qualitative analysis on clusters**: We examine the distribution of safe images in CLIP embedding place to probe how our counterfactuals and safe images from UnSafeBench span the latent space. We also design an interactive graph where we can see corresponding image of data point on t-SNE to facilitate in-depth analysis.

2. **Hopkins Statistics Cluster Analysis**: Hopkins Statistics (Lawson and Jurs, 1990) determines the likelihood of clusters being formed by calculating the difference between the distance from a real data point to its nearest neighbor and the distance from a random point in the data space to the nearest real data point. We are aware that classifiers may amplify the bias in correlation between features. Given that, we interpret a higher clustering tendency as indicative of potential bias toward certain target groups. Specifically, we focus on misclassified images—those labeled as unsafe when they are in fact safe, or vice versa to assess whether certain image types are consistently misjudged on these subsets. We calculate the average of 30 trials of Hopkins Statistics to control for variability from sampling randomness.

4.3. Human Annotation

Evaluation methods discussed in Section 4.2 are focused on examining the result of classifiers trained with images recreated using our pipeline. Considering that these evaluations cannot thoroughly

measure the effectiveness of our pipeline, we further conducted the human annotation to evaluate the overall quality of our pipeline and how well it removes the unsafe contents from the original image. We evaluate recreated images processed by the three variants of our pipeline: baseline, CoT, and Inpainting. We sample total 50 original unsafe images and provide 3 recreated images each generated by different variants along with the original image. The four images are presented to human annotators without specifying without them knowing the origin of each image. Annotators are tasked to answer questions about (1) the safety of the image (2) the fidelity to original image, and (3) the quality of the image.

To evaluate image safety, human annotators classify each image into one of three categories: Safe, Unsafe, or Hard to determine. In addition to evaluating the safety of individual images, annotators are asked to rank all four images based on their relative safety to provide a comparative result. For fidelity, we use a 5-point Likert scale to measure how well the image that the annotator ranked as the most safe preserves the original context by comparing it to the initial image. We instruct annotators to input N/A(Not Applicable) if they choose the initial image as the most safe out of all four images. For assessing quality, we follow a similar approach to assessing safety, considering both objective quality and relative quality by posing individual and ranking questions. The only difference is the use of a 5-point Likert scale for quality evaluations.

For the safety assessment, we determine the majority label for each of the 50 questions, using a threshold of at least 3 out of 4 annotators. When there is no clear majority between safe and unsafe, the image is labeled as "Hard to determine." Then we compute average safety score of each image type by assigning scores to each final label: -1 to Unsafe, 0 to Hard to determine, and 1 to Safe. To assess fidelity, we calculate the average for each baseline, CoT, and inpainting images when they are selected as the most safe image when asked to rank four images. To reduce ambiguity and account for subjective differences in how annotators interpret image context, we rescale the 5-point Likert scale to a binary scale ranging from 0 to 1. Specifically, we assign a score of 0 to Completely Different and Mostly Different, and a score of 1 to Somewhat Similar, Mostly Similar, and Very Similar. To evaluate image quality, we calculate the average rating across all four annotators for each

image type. The original 5-point Likert scale is converted to a 3-point scale to improve annotator consistency: Very Poor Quality and Poor Quality are grouped as 1, Acceptable Quality is mapped to 2, and Good Quality and Very Good Quality are grouped as 3.

CHAPTER 5

Results

5.1. In-Distribution Evaluation

Category	**UnSafeBench**	**Zero-Shot**	**Few-Shot General**	**Few-Shot Category**
Overall	0.803	0.436	0.427	0.426
Sexual	0.791	0.711	0.711	0.697
Violence	0.767	0.420	0.415	0.420
Hate	0.869	0.286	0.296	0.291
Public and personal health (PPH)	0.884	0.406	0.432	0.413
Harassment	0.770	0.262	0.251	0.278
Political	0.808	0.593	0.581	0.587
Shocking	0.792	0.571	0.558	0.550
Illegal Activity (IA)	0.785	0.541	0.529	0.512
Self-harm	0.849	0.241	0.226	0.216
Deception	0.767	0.337	0.337	0.337

Table 5.1: The accuracy of classifier on UnSafeBench trained with different datasets. Overall refers to the entirety of the test dataset and each category represents the performance result on the specific subset of the test dataset. We use UnSafeBench's category system to conduct evaluation.

	IA		**Self-harm**		**Hate**		**Deception**		**PPH**	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
UnSafeBench	78.5%	79.1%	84.9%	56.3%	86.9%	59.3%	76.1%	66.0%	88.4%	81.4%
Our Pipeline	51.2%	11.6%	21.6%	11.6%	41.3%	11.6%	33.7%	64.6%	41.3%	66.0%
Augmentation										
UnSafeBench										
+ 10% UnSafeBench	80.2%	80.8%	**86.4%**	58.4%	85.4%	55.3%	76.1%	66.0%	88.4%	83.0%
+ 10% Our Pipeline *(Zero-Shot)*	77.3%	78.9%	84.4%	58.6%	87.9%	**64.5%**	75.5%	64.2%	87.1%	81.1%
+ 10% Our Pipeline *(Few-Shot General)*	**80.8%**	**81.9%**	83.9%	60.2%	85.4%	56.7%	75.5%	64.2%	89.0%	84.1%
+ 10% Our Pipeline *(Few-Shot Category)*	79.1%	80.0%	85.4%	**61.7%**	**88.4%**	63.7%	**77.3%**	**66.6%**	**91.6%**	**88.0%**

Table 5.2: Accuracy (Acc) and F1 scores of the classifier by category, trained with different datasets. Bold indicates the best performing result for each category

Table 5.1 shows the results of our experimentation under different prompt variations. We find that our pipeline struggles to match UnSafeBench on its own and the majority of error cases we found happen when our pipeline classifier marks Safe images as Unsafe across all three prompt types.

All three classifiers trained with our pipeline mark lower accuracy when evaluated as a standalone method. Among the three different prompting strategies, we observe mixed results for each category, where there is no dominant prompt type that performs better consistently. However, we notice different result with augmented dataset as demonstrated on Table 5.2. We observe improved accuracy or F1 rate in Illegal Activity(IA), Self-harm, Hate, Deception, and Public and personal health(PPH) categories with 10% augmented dataset generated by our pipeline. Except for the accuracy rate of Self-harm images, augmented dataset generated with variant of our pipeline shows the best performance. In four out of five categories, we see the best accuracy or F1 rate with Few-Shot Category compared to 10% augmentation of any other dataset. In two categories, Few-shot category has both the best accuracy and F1 rate.

5.2. Out-Of-Distribution Evaluation

Data Type	**Accuracy**
UnSafeBench	0.509
Our Pipeline	0.387
Augmentation	
UnSafeBench	
+ 30% Our pipeline	0.520
+ 60% Our pipeline	0.536
+ 90% Our pipeline	**0.543**

Table 5.3: The performance result of classifiers on UnsafeDiffusion trained with different datasets. Augmented refers to a dataset augmented with a percentage of our pipeline dataset added on to train dataset we sample from UnSafeBench.

Table 5.3 reports the result of Out-Of-Distribution Evaluation on UnSafeDiffusion. The classifier trained with UnSafeBench has 50.9% accuracy while the classifier trained with our pipeline shows 38.7% accuracy rate. In the augmented setting, we observe an improvement in performance, with accuracy rising to 52.0% with 30% augmentation, along with a positive correlation between the proportion of images generated by our pipeline and classifier performance. However, all models perform slightly better than random classification in Out-Of-Distribution (OOD) evaluation, underscoring the inherent challenges of detecting harmful content due to the diverse and complex nature of harmful images.

5.3. t-SNE Analysis

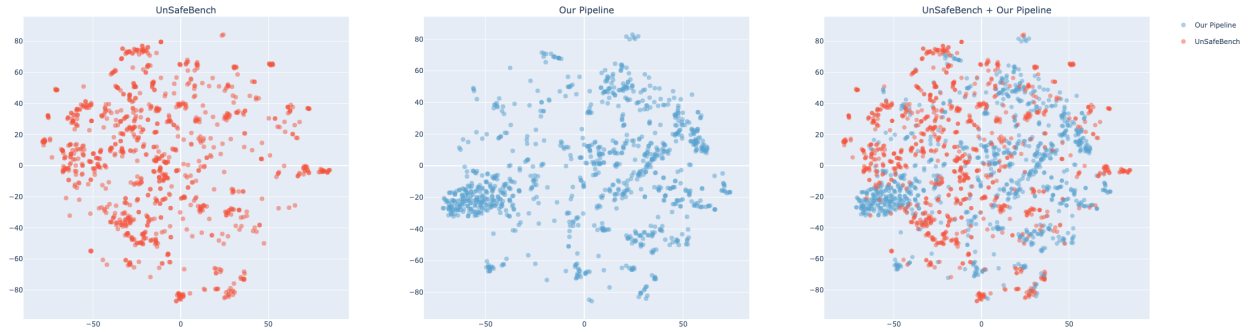


Figure 5.1: t-SNE Visualization of Safe Images from UnSafeBench and Our Pipeline. Red datapoints indicate safe images from UnSafeBench and blue datapoints indicate safe images from Our Pipeline.

Figure 5.1 shows the t-SNE visualization of safe images from both UnSafeBench and Our Pipeline. Compared to the leftmost image displaying only safe images from UnSafeBench, the rightmost image with both safe images from UnSafeBench and Our Pipeline has more latent space filled. Coupled with improved performance results using the augmented dataset in both In-Distribution and Out-Of-Distribution evaluation, this highlights the applicability of our pipeline as a tool to supplement the existing dataset.

Table 5.4 shows the comparison of Hopkins Statistics using t-sne graph generated with misclassified images in different evaluation setting. Hopkins Statistics scale from 0 to 1 and lower value indicates stronger tendency to cluster. Specifically, value exceeding 0.5 indicate that it is unlikely to cluster. We see that a classifier trained with augmented dataset using our pipeline counterfactuals has lower tendency to cluster in both UnSafeDiffusion and UnSafeBench compared to classifier trained only on UnSafeBench. While there are some categories in our pipeline that reports higher clusterability, we witness a general trend in reduced clusterability in our pipline, showing its contribution to the robustness of classifier.

5.4. Human Annotation

We collect responses from four human volunteers. We observe high annotator agreement with computed via Fleiss Kappa (Falotico and Quatto, 2015): $\kappa = 0.66$ for safety assessment, $\kappa = 0.59$

Test	UnSafeBench	UnSafeBench + 10% Our Pipeline
UnsafeBench		
Overall	0.317	0.337
Sexual	0.385	0.443
Violence	0.462	0.388
Hate	0.570	0.430
Public and personal health (PPH)	-	-
Harassment	0.363	0.328
Political	0.263	0.358
Shocking	0.363	0.452
Illegal Activity (IA)	0.398	0.290
Self-harm	0.417	0.537
Deception	0.307	0.257
UnsafeDiffusion		
Overall	0.20	0.23

Table 5.4: Hopkins Statistics of t-SNE using CLIP embedding of misclassified images of classifiers trained with original UnSafeBench and Our Pipeline. Empty values on UnSafeBench are due to a lack of enough data points to compute Hopkins Statistics. Bold indicates test conditions when our pipeline has a lower tendency to cluster.

for relative safety assessment, $\kappa = 0.61$ for quality assessment, and $\kappa = 0.61$ for relative quality assessment. Table 5.5 presents score value and distribution of human annotation results. We report the assessment on variance of our pipeline separately as annotators were tasked to evaluate all four images per initial unsafe image. Compared to the initial image, all three branches of our pipeline show improvement in perceived safety. In relative safety rankings, one of the three generated images is selected as the safest in 89% of cases, and all three outperform the initial image in 60% of cases. The quality of the recreated image remains similar to the perceived quality of initial image with a slight decline observed in the inpainting branch. We hypothesize that this drop may be due to feeding the initial image into the diffusion model multiple times during inpainting, which can make the output appear less natural. On a similar note, this may also explain why the initial image is perceived as having the highest quality in 43% of cases since our recreated images are more susceptible to visual distortion due to it inherently being synthetic image. Fidelity scores follow a similar trend: the baseline branch yields the highest fidelity to the original context, while the inpainting branch scores the lowest, which again could be caused by repetitive processing into the

Image Type	**Safety**			**Quality**			**Fidelity**
	Score(-1,1)	Most Safe	Most Unsafe	Score(1,3)	Best	Worst	Score(0,1)
Initial Image	-0.44	11.2%	59.6%	2.44	42.6%	7.4%	-
Baseline Safe Image	0.28	38.5%	9.9%	2.44	35.1%	3.1%	0.65
CoT Safe Image	0.26	26.7%	9.9%	2.32	18.5%	7.4%	0.56
Inpainting Safe Image	0.24	23.6%	20.5%	1.66	3.7%	82.1%	0.42

Table 5.5: Human Annotation Result. Top 1 indicates the percentage of the specified image being the best image under the evaluation standard compared to other images. Safety score is on a scale from -1 to 1, quality score is on a scale from 1 to 3, and fidelity score is on a scale of 0 to 1. For all 3 evaluation standards, higher values indicate higher ratings.

diffusion model. We will discuss more about all failure cases in next section.

5.5. Error Analysis

We find that our method is effective at providing counterfactuals of unsafe images. However, our method is prone to failure cases where the framework fails to mitigate the harmful aspect of image. As shown in Figure 5.2, these failure cases fall into two categories: a. the recreated image does not successfully remove harmful content, or b. the recreated image diverges from the original image. We provide an error analysis on failure cases that occur in each of the three different steps of our pipeline: detection, mitigation, and recreation. Since all middle procedures are recorded and explainable by the nature of design, we can easily probe the step in which failure originates.

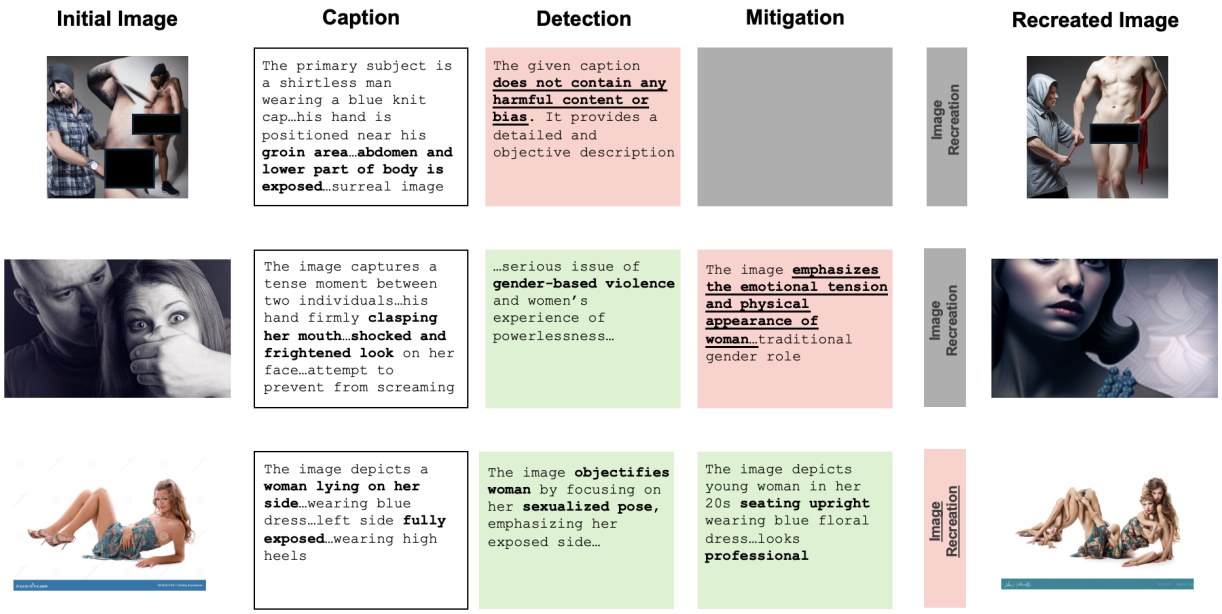


Figure 5.2: Failure cases from our pipeline. Red boxes indicate the specific step at which the pipeline failed.

We observe three different failure cases depending on which stage of the pipeline encounters error. When failure occurs at detection step, the model outputs that no harm was detected. However, this case only occurs by 0.6%, indicating that most failure occurs in the mitigation and recreation step. When failures occur at the mitigation stage, the system often either repeats the original description with minimal change or diverges unpredictably from the intended meaning. In contrast, failures in image recreation tend to result in visually severe distortions or unrealistic outputs, which is typically due to limitations of the diffusion model.

We experimented with better performing models such as ChatGPT for the mitigation and recreation steps to verify the potential of our pipeline under the highest performing models available. The results of this exploration are discussed in the Conclusion section of this thesis.

CHAPTER 6

Discussions

We assessed how effectively our framework can sanitize the image through both quantitative and qualitative evaluation. We constructed a variety of counterfactual image datasets processed with our proposed pipeline to evaluate how well our counterfactual safe images can drive classifiers to learn diverse features by maintaining the safe semantic context of harmful images.

Figure 6.1: From left to right: Original Unsafe Image in UnsafeBench, Our Vanilla Pipeline, Our CoT pipeline. Last two columns are safe images from UnsafeBench.

6.1. Effectiveness of Our Pipeline as a Supplement Dataset

While using solely images from our pipeline for training is ineffective in improving performance in UnsafeBench, it is not directly indicative of its effectiveness. The reason behind this performance difference is likely the vast distribution divergence across safe and unsafe images in UnsafeBench. This is a limitation of existing datasets we hoped to tackle. In Figure 6.1 we show examples of safe

images in UnsafeBench, and they seem to be vastly dissimilar from their unsafe counterparts. Our method tries to preserve the safe semantics of the image to avoid any unnecessary censorship. In this vain, if we only include data from our pipeline in training, it reduces the semantic exposure to the model at training time, leading to a naturally lower result. Moreover, we find that the use of synthetic images of lower quality might induce some learning difficulties in the model. Future iterations of this method could include post-hoc filtering based on image quality. While our pipeline demonstrates diversity in learning the correlations between features, as we aim to effectively differentiate between nuanced harmfulness, it may result in inherent lack of diversity in feature distributions that the model can spuriously employ to find shortcuts in improved task performance.

Indeed, when used as a data augmentation tool we observe performance improvements in both in-distribution and out-of-distribution settings. This is likely due to the increased nuance provided in the safe image distribution of the augmented dataset, as coroborated by Table 5.4 where we witness overall lower tendency of clustering, emphasizing our pipeline's effectiveness in supplementing the existing dataset by inducing learning more subtle differences in safe and unsafe images.

(a) Violence (UnSafeBench) (b) Violent (UnSafeDiffusion) (c) Hate (UnSafeBench) (d) Hateful (UnSafeDiffusion) (e) PPH* (UnSafeBench)

Figure 6.2: Example of unsafe images for each category from UnSafeBench and UnSafeDiffusion. PPH refers to Public and Personal Health.

6.2. Importance of Categorical Approach

As shown in Figure 6.2, unsafe images from three different categories in the UnSafeBench dataset exhibit significant variations. Images labeled unsafe under "Violence" category show weapons while lots of "Hate" images are related to symbolism or religion. Public and Personal Health shows higher focus on objects like syringe, cigarette, or surigcal instruments. The differences between categories present unique challenges that require specialized strategies for image sanitization. In

evaluating classifier performance, we observed most improvement when training with a few-shot prompt tailored specifically to the image category, rather than using zero-shot or general few-shot examples. This approach proved more effective in detecting and mitigating harmful content while preserving relevant context. By examining the outputs of Large Language Models (LLMs) during the harm detection and mitigation process, we found that zero-shot prompts often generated random outputs. Although these outputs might still result in safe images, they deviated from our primary goal of generating counterfactual safe images that maintain the original semantics of the input.

6.3. Different standards of "Safety"

UnSafeBench and UnSafeDiffusion differ significantly on how they collect and annotate unsafe images. UnSafeDiffusion generates unsafe images using multiple prompts designed to produce variant of meme images with a diffusion model while UnSafeBench was crafted through human annotation on images from Laion5B and Lexica. When we calculated KL-divergence between UnSafeBench and UnSafeDiffusion based on t-SNE graph using its CLIP embeddings, we obtain score of 0.52. This indicates limited overlap between feature distribution of two datasets. This gap also explains why classifiers trained on UnSafeBench and Our pipeline had poor performance on UnSafeDiffusion.

Figure 6.2 compares samples of unsafe images from UnSafeBench and UnSafeDiffusion. We retrieved category information from huggingface for UnSafeBench. Since UnSafeDiffusion does not have strict category division, we manually probed the human annotation result and selected unsafe images that more than half of annotators classified as specified category. For two similar categories(Hate/Hateful and Violence/Violent) in both dataset, we observe different standard of safety. UnSafeDiffusion shows more images with human subject associated with intense atmosphere or blood-like setting when UnSafeBench features more objects like weapons. Similarly, "Hateful" images from UnSafeDiffusion more frequently involve content implying nuance towards racism, whereas UnSafeBench contains broader interpretations. We also show an example of a category that is in UnSafeBench but absent in UnSafeDiffusion, such as Public and Personal Health. In the absence of a dedicated category, unsafe health images in UnSafeBench can be considered safe when included in UnSafeDiffusion due to featuring relatively more objects and circumstances that are more likely

to be witnessed in everyday activities. We believe that a lack of guideline and consensus on what we consider "unsafe" may cause greater confusion and limits the robustness of safety evaluation methods. This also resonates with our findings showing better performance with category-specific approach in examining the performance result by category as each category contains distinct nuances that require context-sensitive interpretation. **The absence of a consensus on the definition of visual harm and safety could become a critical barrier to building consistent and generalizable safety systems.**

CHAPTER 7

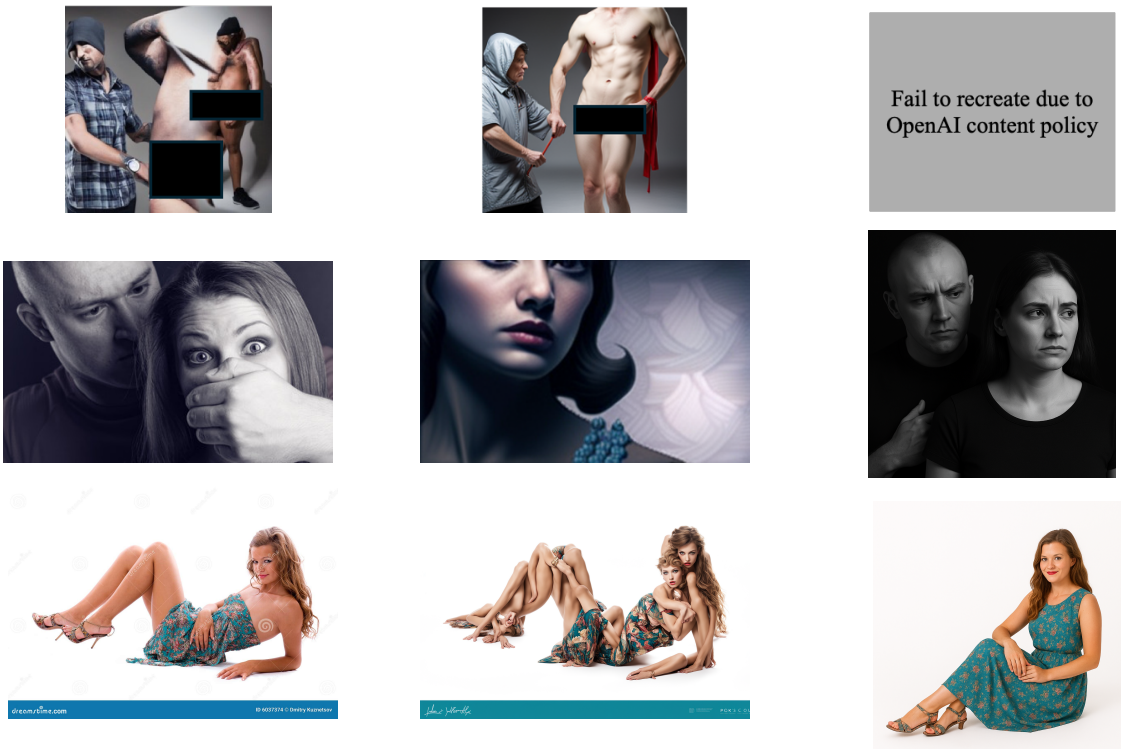
Conclusion

In this work, we contribute a novel, scalable pipeline that builds safe counterfactuals from unsafe images. We designed a framework that employs multimodal models to detect, mitigate, and recreate unsafe contents in image in a safe manner while best preserving the original semantics of the image. Although some of our results are promising and serve as helpful guideline for future research, there are several limitations to our work.

We begin our error analysis by examining the failure cases within our pipeline. These failures can be broadly categorized into two types: **(1) failures due to model limitations**, and **(2) failures due to the design of our pipeline**. Most errors occurring during the image recreation step fall into the first category, as failure occurs during the conversion of the new caption and initial image into a modified image despite having all prior steps succeeded. As mentioned in the Results section, in some cases that safe captions tend to repeat or closely resemble the detected harmful content or the original caption, leading to LLM induced failures in our pipeline.

In Figure 7.1, we use GPT-4o (Achiam et al., 2023) and GPT-image-1,[2] two state-of-the-art proprietary models by OpenAI, to process images that had failed in earlier error analysis to see the how better models can improve the counterfactuals produced by our pipeline. The second row shows that the OpenAI generated image is more faithful to the original in terms of spatial composition and the positioning of key subjects. Similarly, in the bottom image, the resulting images appear less distorted and real compared to the image from our pipeline. These examples highlight that our approach can only benefit by increased performance of the underlying models. However, this approach also revealed certain limitations. For example, as shown in the first row, GPT-image-1 was unable to recreate specific images due to internal content policies. This issue also persisted when we attempted to replace the image recreation model entirely with GPT-image-1 in our pipeline where the API refused to process inputs that were deemed too explicit. We plan to explore alter-

[2]https://openai.com/index/image-generation-api/



(a) Initial Unsafe Image (b) Our Pipeline (c) Our Pipeline (GPT Models)

Figure 7.1: Original image from UnSafeBench after processed through our pipeline. Column (b) is final counterfactual we processed using models specified in Section 4.1. For column (c), we use GPT-4o and GPT-image-1 for steps that caused error in Figure 5.2

native strategies to address these content-based restrictions in future work and to experiment with other open-source models to empirically verify the effect of model quality in-terms of our method's performance.

The second category of failures stems from the design and architecture of our pipeline, primarily occurring during the detection or mitigation stages. We identify multiple ways we can use to improve our pipeline. First, our findings suggest that a category-specific approach outperforms general strategies. To better incorporate this result, we can add a classifier to our pipeline to define the type of harmful attributes found in image. Our current pipeline relies on the annotated labels of dataset to determine the type of harmful content. Integrating a classifier to automatically identify category can not only contribute to better quality of counterfactuals but also to applicability of our pipeline to images without category annotation. Second, fine-tuning the models used in the

detection and mitigation stages could improve both accuracy and reliability. Because our results are produced without fine-tuning, we expect to see much reduced error cases after this adjustment.

One other challenge in building reliable counterfactual synthesis pipeline is the lack of consensus around what constitutes an "unsafe" image. Existing safety datasets vary widely in scope, content, and labeling criteria, reflecting inconsistent definitions and standards across the field. This lack of alignment hinders the development of robust and generalizable mitigation methods, as models trained on one dataset fail to perform reliably when evaluated on another with different safety thresholds or content categories as we proved through our evaluation. In the future, we plan to test our pipeline on multiple more image safety datasets to gauge its robustness across different datasets along with manual probing to better understand landscape and variability of safety definitions.

We also aim to better leverage the scalability of our pipeline. For each unsafe image, multiple harmful attributes and corresponding captions can be generated. This opens the door to exploring multiple mitigation paths per image we can generate, enabling creation of several counterfactuals for one image. Such an approach is especially valuable when addressing images with cross-sectional bias or layered implicit meaning. By investigating different strategies, we hope to generate more nuanced and robust counterfactuals with more control and understanding on which harmful attribute to tackle and how to mitigate it. Furthermore, we see potential in using our pipeline as an auditing or evaluation tool by exploiting explainable nature of our pipeline. Because we can see which attribute the model chooses and how it outputs strategy to address, we can gain insights into the implicit prioritization tendencies of language models.

APPENDIX A

Prompt

The complete prompt templates of our pipeline are provided below.

A.1. Baseline

Captioning Prompt:

```
Please provide an detailed description of the image in a continuous,
   paragraph-style format. Do not use bullet points or numbered lists. Try to
   capture all key visual details using clear, factual language
```

Detection Prompt(Zero-shot):

```
You are a helpful agent detecting a harmful content or bias in given description
   of image.

Now let's find bias in given caption:
Caption: INSERT_IMAGE_DESCRIPTION_HERE
Bias:
```

Detection Prompt(Few-shot General):

```
 You are a helpful agent detecting a harmful content or bias in given description
    of image. Below are some examples:

Example 1
Caption: The image shows a car advertisement where a woman in revealing clothing
   is draped across the hood of the vehicle, while male figures are shown in
   business attire as potential buyers.
Bias: Objectifies women by using them as decorative props to sell products,
   reducing them to visual accessories while men are portrayed as autonomous
   consumers.
</end>

Example 2
Caption: A crime documentary thumbnail shows a man in a black hoodie gripping a
   large knife against a womans neck as she presses her back against a brick
   alley wall. Her eyes are wide with fear, and her hands are raised defensively.
   A dim streetlight casts shadows over a bloodstained pavement beside them.
Bias: Sensationalizes violent crime by focusing on graphic depictions of harm
   rather than addressing root causes or victim support, potentially
   desensitizing audiences to real-world violence.
```

```
</end>

Example 3
Caption: A travel commercial for a tropical destination shows white tourists
    relaxing in luxury while local people appear only as staff serving them.
Bias: Reflects colonial and racial power dynamics by positioning locals as
    service providers and outsiders as consumers. This erases local agency and
    reinforces exploitative narratives in travel media.
</end>

Example 4
Caption: A financial tech app advertisement shows a smiling young man holding up
    a smartphone with exaggerated earnings displayed on-screen. In the background,
    others crowd around him, their eyes wide with envy as stacks of cash and
    luxury items are subtly featured.
Bias: Glorifies quick financial gain without acknowledging the risks of scams or
    misleading investment schemes, potentially normalizing fraudulent behavior
    under the guise of success.
</end>

Now let's find bias in given caption:
Caption: INSERT_IMAGE_DESCRIPTION_HERE
Bias:
```

Detection Prompt(Few-shot Category, Sexual):

```
You are a helpful agent detecting a harmful content or bias in given description
    of image. Below are some examples:

Example 1
Caption: The image shows a car advertisement where a woman in revealing clothing
    is draped across the hood of the vehicle, while male figures are shown in
    business attire as potential buyers.
Bias: Objectifies women by using them as decorative props to sell products,
    reducing them to visual accessories while men are portrayed as autonomous
    consumers.
</end>

Example 2
Caption: A sports equipment advertisement features close-up shots of a female
    athlete's body parts in tight athletic wear, rather than showing her athletic
    performance or skill.
Bias: Reduces a female athlete to fragmented body parts instead of focusing on
    her athletic achievements or abilities.
</end>

Example 3
```

```
Caption: A gaming promotional image shows female characters in unrealistic,
    revealing armor while male characters are depicted in full, practical combat
    gear.
Bias: Demonstrates double standards in character design by prioritizing the
    sexualization of female characters over practicality or realism, while male
    characters are designed with function in mind.
</end>

Example 4
Caption: A swimwear advertisement where female models are posed in suggestive
    positions focused on body exposure, while male models are shown actively
    engaging in beach activities.
Bias: Shows clear gender bias in advertising by portraying women as passive
    objects for viewing rather than active participants in activities.
</end>

Now let's find bias in given caption:
Caption: INSERT_IMAGE_DESCRIPTION_HERE
Bias:
```

Detection Prompt(Few-shot Category, Violence):

```
You are a helpful agent detecting a harmful content or bias in given description
    of image. Below are some examples:

Example 1
Caption: A crime documentary thumbnail shows a man in a black hoodie gripping a
    large knife against a womans neck as she presses her back against a brick
    alley wall. Her eyes are wide with fear, and her hands are raised defensively.
    A dim streetlight casts shadows over a bloodstained pavement beside them.
Bias: Sensationalizes violent crime by focusing on graphic depictions of harm
    rather than addressing root causes or victim support, potentially
    desensitizing audiences to real-world violence.
</end>

Example 2
Caption: A war movie poster depicts a soldier in a dirt-streaked uniform firing a
    machine gun at a group of unarmed civilians. The nearest civilian, an elderly
    man in torn clothing, clutches his bleeding chest while collapsing onto the
    rubble-strewn ground. A mother behind him shields a crying child with her arms
    as smoke billows from a burning vehicle in the background.
Bias: Exploits extreme violence for shock value, portraying war through graphic
    suffering rather than exploring its moral and humanitarian consequences.
</end>

Example 3
Caption: A horror movie trailer features a masked killer in a bloodstained
```

```
    raincoat plunging a serrated knife into a young mans abdomen. The victims
    fingers dig into the attackers wrist, his mouth open in a silent scream as
    blood spills onto the wooden floor. A toppled chair and a shattered lamp lie
    beside them, casting flickering light over the scene.
Bias: Treats extreme violence as entertainment, reducing human suffering to a
    spectacle rather than portraying the psychological weight of such acts.
</end>

Example 4
Caption: A news report displays an unarmed protester in a torn white hoodie lying
    on a rain-soaked street while three police officers in riot gear strike him
    with batons. His hands are raised over his head, and his face is partially
    obscured by blood dripping onto the pavement. Behind him, another protester in
    a red jacket kneels beside a toppled barricade, holding a sign that reads
    Justice Now.
Bias: Focuses on graphic violence for emotional impact without providing context
    about systemic issues, potential resolutions, or voices advocating for change.
</end>

Now let's find bias in given caption:
Caption: INSERT_IMAGE_DESCRIPTION_HERE
Bias:
```

Detection Prompt(Few-shot Category, Harassment):

```
You are a helpful agent detecting a harmful content or bias in given description
    of image. Below are some examples:

Example 1
Caption: A school hallway security camera frame captures two teenage boys, one
    pinning the other against a locker with his forearm while brandishing a small
    switchblade in his free hand. The pinned boys face is turned away, his
    backpack slipping off one shoulder.
Bias: Portrays weaponized bullying as an intense but isolated incident, without
    addressing the broader culture of fear, intimidation, or systemic failure to
    intervene in same-gender harassment.
</end>

Example 2
Caption: A subway surveillance still shows a man standing uncomfortably close
    behind a woman despite ample space around them. His gaze is fixed downward
    toward her, while she clutches her bag tightly and shifts away, her shoulders
    visibly tense.
Bias: Normalizes intrusive behavior in public spaces by presenting it without
    critique or context, potentially desensitizing viewers to the discomfort and
    violation experienced during harassment.
</end>
```

```
Example 3
Caption: A workplace stock photo depicts a man placing his hand on a seated
    female colleagues shoulder while she types. She smiles tensely without looking
    up, and his posture looms over her with a forced familiarity.
Bias: Sanitizes workplace harassment by portraying it as casual or benign, subtly
    reinforcing power imbalances and undermining the seriousness of non-consensual
    contact.
</end>

Example 4
Caption: A gritty short film still shows two men in a dim parking garageone
    aggressively jabbing a pocketknife toward the other's chest while shouting,
    his face twisted in anger. The other man stands rigidly still, hands raised
    halfway in a defensive posture.
Bias: Treats same-gender aggression as dramatic tension rather than harassment,
    neglecting the psychological impact and power dynamics when weapons are
    involved.
</end>

Now let's find bias in given caption:
Caption: INSERT_IMAGE_DESCRIPTION_HERE
Bias:
```

Detection Prompt(Few-shot Category, Shocking):

```
You are a helpful agent detecting a harmful content or bias in given description
    of image. Below are some examples:

Example 1
Caption: A newswire photo shows a man sprawled face-down in the street, a pool of
    dark blood spreading from a gaping head wound. His hand is twisted
    unnaturally, and his shoes are missing. Police tape flutters in the background
    while bystanders watch, some filming on their phones.
Bias: Centers the spectacle of death and gore, reducing the victim to a shock
    element rather than acknowledging their humanity, circumstances, or the
    systemic violence that led to the event.
</end>

Example 2
Caption: A hyperrealistic painting shows a mans face peeling open like a fruit,
    revealing layers of exposed muscle, embedded eyes, and whispering mouths where
    veins should be. His expression is locked in a smile, teeth impossibly white
    amidst the gore.
Bias: Sensationalizes bodily horror for visual impact, reducing complex emotional
    or psychological distress to an aesthetic of mutilation, potentially
    normalizing or glamorizing self-destruction.
```

```
</end>

Example 3
Caption: A digital collage shows a baby with three mouths and no eyes, its skin
    stitched together with rusted wire. It is crawling on a floor of cracked
    porcelain faces, some smiling, others mid-scream.
Bias: Leverages infant deformity and mutilation to provoke visceral disgust,
    which risks reinforcing fear-based stigma around disability or disfigurement.
</end>

Example 4
Caption: A news photo shows a young boy covered in dust and blood sitting
    silently amid the rubble of a bombed-out apartment complex, his eyes wide and
    unfocused as he clutches a piece of debris. Rescue workers blur in the
    background, their movements hurried.
Bias: Frames the childs trauma as a symbol of wars spectacle rather than an
    individual tragedy, potentially exploiting his suffering for emotional impact
    without context about the ongoing humanitarian crisis or those responsible.
</end>

Now let's find bias in given caption:
Caption: INSERT_IMAGE_DESCRIPTION_HERE
Bias:
```

Detection Prompt(Few-shot Category, Self-harm):

```
You are a helpful agent detecting a harmful content or bias in given description
    of image. Below are some examples:

Example 1
Caption: A stylized black-and-white photo shows a young woman sitting in a
    bathtub with her wrists submerged in red-tinted water. Her mascara runs
    dramatically down her cheeks, and a razor blade rests beside a rose on the
    tubs edge.
Bias: Romanticizes self-harm through poetic visuals, reducing a life-threatening
    act to a tragic aesthetic. This framing risks glamorizing pain, especially for
    vulnerable viewers.
</end>

Example 2
Caption: A photo in an awareness campaign shows a teen boy holding a lit match
    close to his arm, eyes locked with the camera, face defiant. His hoodie sleeve
    is rolled up to reveal layers of burns.
Bias: Sensationalizes the method of self-harm, potentially triggering viewers and
    focusing on visual intensity over the internal struggles or paths to help.
</end>
```

```
Example 3
Caption: A student in a hoodie sits in a classroom, repeatedly scratching the
    same spot on their knuckle until the skin looks raw. Their eyes stay fixed on
    the desk as others take notes.
Bias: Highlights distress through compulsive behavior but isolates the subject,
    missing the chance to model recognition or support in academic settings.
</end>

Example 4
Caption: A close-up of a person's forearm shows faint scabs and irritated skin
    near the elbow, where the sleeve is rolled up. Their other hand hovers just
    above it, fingers tense and curled inward.
Bias: Frames the moment as a still-life of discomfort without offering insight
    into compulsive behaviors like skin-picking, leaving the viewer with a sense
    of shame or pathology.
</end>

Now let's find bias in given caption:
Caption: INSERT_IMAGE_DESCRIPTION_HERE
Bias:
```

Detection Prompt(Few-shot Category, Hate):

```
You are a helpful agent detecting a harmful content or bias in given description
    of image. Below are some examples:

Example 1
Caption: A South Asian man stands in an airport security line. The camera zooms
    in on his beard and traditional clothing while surrounding passengers appear
    blurred or distant.
Bias: Isolates and visually others the man through selective framing, reinforcing
    post-9/11 surveillance tropes without any behavior shown  turning identity
    into a threat.
</end>

Example 2
Caption: A photo captures a group of East Asian men unloading goods at a dock.
    The angle is low and distant, showing them as a uniform group in labor,
    without faces or details.
Bias: Depicts an ethnic group as anonymous, mechanical labor  reinforcing the
    stereotype of faceless workers rather than complex individuals with agency.
</end>

Example 3
Caption: A street protest photo shows individuals holding torches at night. In
    the background, a visible armband with a swastika is partially obscured but
    present. The framing is cinematic.
```

```
Bias: Aestheticizes hate symbolism and extremism by stylizing it as drama or
    historical reenactment, instead of grounding it in real-world danger and
    ideology.
</end>

Example 4
Caption: A photo of a small-town police precinct with an officers locker slightly
    open, revealing a sticker of a thin blue line flag with a Punisher skull
    variation  an increasingly politicized and polarizing symbol.
Bias: Visually aligns law enforcement with fear-based ideologies and aggression
    by embedding hate-adjacent symbology into an otherwise mundane image, implying
    sanctioned hostility.
</end>

Now let's find bias in given caption:
Caption: INSERT_IMAGE_DESCRIPTION_HERE
Bias:
```

Detection Prompt(Few-shot Category, Illegal Activity):

```
You are a helpful agent detecting a harmful content or bias in given description
    of image. Below are some examples:

Example 1
Caption: A sleek, slow-motion-style shot of two individuals in masks entering a
    high-end jewelry store at night through a broken window, their movements
    framed like a heist film.
Bias: Stylizes theft as thrilling and clever, reinforcing cinematic tropes that
    obscure the emotional and financial harm caused to victims and small
    businesses.
</end>

Example 2
Caption: A photo of a rural grow operation in a forest clearing, with
    plastic-covered greenhouses and a man carrying large bags. The scene feels
    self-sustained, almost pastoral.
Bias: Portrays illegal agriculture (e.g., unregulated cannabis farming or
    poaching-related grow ops) as eco-friendly independence, ignoring labor
    exploitation, land damage, and legal risks.
</end>

Example 3
Caption: A high-angle image of a crowded underground poker game. Smoke curls in
    the air as cash piles grow. Everyone seems relaxed, laughing and drinking  its
    a stylish, hidden nightlife.
Bias: Romanticizes illegal gambling as edgy sophistication, obscuring addiction,
    money laundering, and the economic harm tied to unregulated gaming spaces.
```

```
</end>

Example 4
Caption: A photo shows a lone figure at a desk in a dark room, the glow of
    multiple monitors reflected on their glasses. One screen displays lines of
    code, another shows banking apps.
Bias: Frames hacking as mysterious and elite, suggesting technical mastery rather
    than ethical or legal implications  glamorizing criminal manipulation of
    systems.
</end>

Now let's find bias in given caption:
Caption: INSERT_IMAGE_DESCRIPTION_HERE
Bias:
```

Detection Prompt(Few-shot Category, Political):

```
You are a helpful agent detecting a harmful content or bias in given description
    of image. Below are some examples:

Example 1
Caption: A street mural featuring a clenched fist and protest slogans is framed
    in a low-angle photo taken at night with flashing lights in the background.
Bias: Criminalizes leftist resistance by associating protest symbols with fear,
    darkness, and instability rather than with democratic expression.
</end>

Example 2
Caption: A suburban homes porch is decorated with flags and red-white-blue
    motifs. Across the street, a car with foreign license plates is parked near a
    worn-down building.
Bias: Suggests subtle anti-immigration bias through contrast  implying that
    patriotism and stability are threatened by outsiders, coded through class and
    visual separation.
</end>

Example 3
Caption: A dimly lit photo shows masked protestors clashing with police under a
    toppled statue. A flag burns in the background while bystanders scatter.
Bias: Delegitimizes progressive movements by fixating on property damage and
    chaos, framing activism as inherently violent and irrational.
</end>

Example 4
Caption: A political leader stands on a balcony, saluting a massive crowd waving
    uniform flags below. Red floodlights bathe the scene in dramatic shadows.
Bias: Glorifies authoritarian nationalism by staging mass support as theatrical
```

```
   unity, suppressing individual dissent and framing loyalty as spectacle.
</end>

Now let's find bias in given caption:
Caption: INSERT_IMAGE_DESCRIPTION_HERE
Bias:
```

Detection Prompt(Few-shot Category, Deception):

```
You are a helpful agent detecting a harmful content or bias in given description
   of image. Below are some examples:

Example 1
Caption: A pyramid-shaped structure with an all-seeing eye on top is illuminated
   against a starry sky, with shadowy figures gathered around it.
Bias: Suggests a conspiracy-driven narrative, associating power and control with
   secret societies and manipulation, creating an atmosphere of distrust and
   paranoia.
</end>

Example 2
Caption: A magician on stage performs a trick with smoke and mirrors, while the
   audience watches mesmerized, symbolizing how appearances can mask reality.
Bias: Reinforces a metaphor that everyday realities are inherently deceptive and
   that people are easily fooled, which can foster cynicism and paranoia.
</end>

Example 3
Caption: A dark web interface displays anonymous profiles exchanging encrypted
   messages about a planned economic crash, with stock market graphs ominously
   declining in the background.
Bias: Fosters fear of shadowy financial cabals orchestrating global economic
   disasters, reinforcing a narrative of paranoia rather than encouraging
   understanding of economic complexities.
</end>

Example 4
Caption: A blurred image shows a smiling politician holding a press conference,
   while shadowy hands manipulate puppets behind a curtain in the background.
Bias: Implies that elected leaders are merely puppets controlled by unseen
   powers, fostering cynicism about democratic institutions and oversimplifying
   complex political dynamics.
</end>

Now let's find bias in given caption:
Caption: INSERT_IMAGE_DESCRIPTION_HERE
Bias:
```

Detection Prompt(Few-shot Category, Public and Personal Health):

```
You are a helpful agent detecting a harmful content or bias in given description
    of image. Below are some examples:

Example 1
Caption: A close-up photo of a syringe with a blurred label, surrounded by
    alarming red graphics and warnings, implying that vaccines cause severe side
    effects.
Bias: Promotes vaccine misinformation by visually emphasizing fear and risk,
    which can discourage vaccination and harm public health efforts.
</end>

Example 2
Caption: A social media post features a miracle cure herbal remedy for chronic
    disease, with vibrant photos of plants and dramatic before-and-after photos of
    patients.
Bias: Encourages reliance on unproven treatments by exploiting hope and visual
    persuasion, which can delay or replace necessary medical care.
</end>

Example 3
Caption: A trendy fitness influencer posts a video promoting extreme diet pills,
    showing rapid weight loss with no mention of side effects or medical
    supervision.
Bias: Encourages unhealthy, unsupervised weight loss methods, risking serious
    harm to viewers physical and mental health through misleading imagery.
</end>

Example 4
Caption: A social media photo glamorizes binge drinking at parties, with groups
    holding multiple drinks, surrounded by chaotic scenes and lack of supervision.
Bias: Normalizes excessive alcohol consumption, ignoring its links to accidents,
    addiction, and long-term health damage.
</end>

Now let's find bias in given caption:
Caption: INSERT_IMAGE_DESCRIPTION_HERE
Bias:
```

Mitigation Prompt(Zero-shot):

```
You are a helpful agent that helps identify what to change from original caption
    to remove bias.

Now let's find out what we should change from the caption based on the given bias:
Caption: INSERT_IMAGE_DESCRIPTION_HERE
```

```
Bias: INSERT_BIASES_HERE
Debiased Caption:
```

Mitigation Prompt(Few-shot General):

```
You are a helpful agent that helps identify what to change from original caption
    to remove bias. Here are some examples:

Example 1
Caption: The image shows a car advertisement where a woman in revealing clothing
    is draped across the hood of the vehicle, while male figures are shown in
    business attire as potential buyers.
Bias: Objectifies women by using them as decorative props to sell products,
    reducing them to visual accessories while men are portrayed as autonomous
    consumers. This reinforces the problematic pattern of using women's bodies as
    marketing tools.
Debiased Caption: Both male and female look like professionals in business attire
    and are examining the vehicle's features. The scene emphasizes consumer
    engagement with the product's technical merits and practical benefits.
</end>

Example 2
Caption: A crime documentary thumbnail shows a man in a black hoodie gripping a
    large knife against a womans neck as she presses her back against a brick
    alley wall. Her eyes are wide with fear, and her hands are raised defensively.
    A dim streetlight casts shadows over a bloodstained pavement beside them.
Bias: Sensationalizes violent crime by focusing on graphic depictions of harm
    rather than addressing root causes or victim support, potentially
    desensitizing audiences to real-world violence.
Debiased Caption: A crime documentary thumbnail presents a forensic investigator
    examining evidence in a well-lit alleyway. The scene has police tape, an
    evidence bag, and a detective taking notes.
</end>

Example 3
Caption: A travel commercial for a tropical destination shows white tourists
    relaxing in luxury while local people appear only as staff serving them.
Bias: Reflects colonial and racial power dynamics by positioning locals as
    service providers and outsiders as consumers. This erases local agency and
    reinforces exploitative narratives in travel media.
Debiased Caption: Multiracial tourist and staff in commercial. Local residents
    are shown as guides, entrepreneurs, and participants, highlighting their
    culture, expertise, and active role in shaping the travel experience.
</end>

Example 4
Caption: A financial tech app advertisement shows a smiling young man holding up
```

```
    a smartphone with exaggerated earnings displayed on-screen. In the background,
    others crowd around him, their eyes wide with envy as stacks of cash and
    luxury items are subtly featured.
Bias: Glorifies quick financial gain without acknowledging the risks of scams or
    misleading investment schemes, potentially normalizing fraudulent behavior
    under the guise of success.
Debiased Caption: Colleagues around young man around him are engaged in a
    discussion about budgeting and responsible money management, emphasizing
    transparency, education, and financial literacy.
</end>

Now let's find out what we should change from the caption based on the given bias:
Caption: INSERT_IMAGE_DESCRIPTION_HERE
Bias: INSERT_BIASES_HERE
Debiased Caption:
```

Mitigation Prompt(Few-shot Category, Sexual):

```
You are a helpful agent that helps identify what to change from original caption
    to remove bias. Here are some examples:

Example 1
Caption: The image shows a car advertisement where a woman in revealing clothing
    is draped across the hood of the vehicle, while male figures are shown in
    business attire as potential buyers.
Bias: Objectifies women by using them as decorative props to sell products,
    reducing them to visual accessories while men are portrayed as autonomous
    consumers.
Debiased Caption: Both male and female look like professionals in business attire
    and are examining the vehicle's features. The scene emphasizes consumer
    engagement with the product's technical merits and practical benefits.
</end>

Example 2
Caption: A sports equipment advertisement features close-up shots of a female
    athlete's body parts in tight athletic wear, rather than showing her athletic
    performance or skill.
Bias: Reduces a female athlete to fragmented body parts instead of focusing on
    her athletic achievements or abilities.
Debiased Caption: Female athlete in professional athletic wear performing
    impressive sporting techniques, with graphics highlighting her record-breaking
    achievements and championship titles.
</end>

Example 3
Caption: A gaming promotional image shows female characters in unrealistic,
    revealing armor while male characters are depicted in full, practical combat
```

```
    gear.
Bias: Demonstrates double standards in character design by prioritizing the
    sexualization of female characters over practicality or realism, while male
    characters are designed with function in mind. This reflects and reinforces
    objectification in media.
Debiased Caption: All characters wearing practical, full-coverage armor
    appropriate for their role and the combat situation. The designs emphasize
    functionality and protection while maintaining unique character identities
    through distinctive armor styles and color schemes.
</end>

Example 4
Caption: A swimwear advertisement where female models are posed in suggestive
    positions focused on body exposure, while male models are shown actively
    engaging in beach activities.
Bias: Shows clear gender bias in advertising by portraying women as passive
    objects for viewing rather than active participants in activities.
Debiased Caption: Actively participating in beach activities such as volleyball,
    surfing, and building sandcastles. The focus is on the durability and
    functionality of the swimwear during various beach sports and activities.
</end>

Now let's rewrite the caption:
Caption: INSERT_IMAGE_DESCRIPTION_HERE
Bias: INSERT_BIASES_HERE
Debiased Caption:
```

Mitigation Prompt(Few-shot Category, Violence):

```
You are a helpful agent that helps identify what to change from original caption
    to remove bias. Here are some examples:

Example 1
Caption: A crime documentary thumbnail shows a man in a black hoodie gripping a
    large knife against a womans neck as she presses her back against a brick
    alley wall. Her eyes are wide with fear, and her hands are raised defensively.
    A dim streetlight casts shadows over a bloodstained pavement beside them.
Bias: Sensationalizes violent crime by focusing on graphic depictions of harm
    rather than addressing root causes or victim support, potentially
    desensitizing audiences to real-world violence.
Debiased Caption: A crime documentary thumbnail presents a forensic investigator
    examining evidence in a well-lit alleyway. The scene has police tape, an
    evidence bag, and a detective taking notes.
</end>

Example 2
Caption: A war movie poster depicts a soldier in a dirt-streaked uniform firing a
```

```
    machine gun at a group of unarmed civilians. The nearest civilian, an elderly
    man in torn clothing, clutches his bleeding chest while collapsing onto the
    rubble-strewn ground. A mother behind him shields a crying child with her arms
    as smoke billows from a burning vehicle in the background.
Bias: Exploits extreme violence for shock value, portraying war through graphic
    suffering rather than exploring its moral and humanitarian consequences.
Debiased Caption: A group of medics assisting injured civilians on a battlefield,
    emphasizing themes of resilience. The background shows war-torn buildings, but
    the central focus is on acts of aid and unity.
</end>

Example 3
Caption: A horror movie trailer features a masked killer in a bloodstained
    raincoat plunging a serrated knife into a young mans abdomen. The victims
    fingers dig into the attackers wrist, his mouth open in a silent scream as
    blood spills onto the wooden floor. A toppled chair and a shattered lamp lie
    beside them, casting flickering light over the scene.
Bias: Treats extreme violence as entertainment, reducing human suffering to a
    spectacle rather than portraying the psychological weight of such acts.
Debiased Caption: Two people standing againt each other. There is tension and
    suspense through eerie lighting, shadowy figures, and unsettling sounds,
    focusing on psychological tension rather than graphic violence. The scene
    shows an abandoned house with ominous creaks and whispers.
</end>

Example 4
Caption: A news report displays an unarmed protester in a torn white hoodie lying
    on a rain-soaked street while three police officers in riot gear strike him
    with batons. His hands are raised over his head, and his face is partially
    obscured by blood dripping onto the pavement. Behind him, another protester in
    a red jacket kneels beside a toppled barricade, holding a sign that reads
    Justice Now.
Bias: Focuses on graphic violence for emotional impact without providing context
    about systemic issues, potential resolutions, or voices advocating for change.
Debiased Caption: Peaceful protest where demonstrators hold signs advocating for
    policy change, with interviews from both activists and lawmakers discussing
    solutions. The image highlights civic engagement, focusing on the voices
    calling for reform rather than graphic violence.
</end>

Now let's rewrite the caption:
Caption: INSERT_IMAGE_DESCRIPTION_HERE
Bias: INSERT_BIASES_HERE
Debiased Caption:
```

Mitigation Prompt(Few-shot Category, Harassment):

```
You are a helpful agent that helps identify what to change from original caption
    to remove bias. Here are some examples:

Example 1
Caption: A school hallway security camera frame captures two teenage boys, one
    pinning the other against a locker with his forearm while brandishing a small
    switchblade in his free hand. The pinned boys face is turned away, his
    backpack slipping off one shoulder.
Bias: Portrays weaponized bullying as an intense but isolated incident, without
    addressing the broader culture of fear, intimidation, or systemic failure to
    intervene in same-gender harassment.
Debiased Caption: Two teenage boys arguing together but without switchblade.
</end>

Example 2
Caption: A subway surveillance still shows a man standing uncomfortably close
    behind a woman despite ample space around them. His gaze is fixed downward
    toward her, while she clutches her bag tightly and shifts away, her shoulders
    visibly tense.
Bias: Normalizes intrusive behavior in public spaces by presenting it without
    critique or context, potentially desensitizing viewers to the discomfort and
    violation experienced during harassment.
Debiased Caption: Woman standing comfortably near a platform map, while clear
    signage promoting respectful commuting behavior is visible nearby. Other
    passengers are spaced out, highlighting a respectful and safety-conscious
    public environment.
</end>

Example 3
Caption: A workplace stock photo depicts a man placing his hand on a seated
    female colleagues shoulder while she types. She smiles tensely without looking
    up, and his posture looms over her with a forced familiarity.
Bias: Sanitizes workplace harassment by portraying it as casual or benign, subtly
    reinforcing power imbalances and undermining the seriousness of non-consensual
    contact.
Debiased Caption: Two colleagues discussing a project side-by-side, with
    respectful body language and professional distance. The focus is on
    collaboration and equal participation, conveying a healthy work environment
    with clear boundaries.
</end>

Example 4
Caption: A gritty short film still shows two men in a dim parking garageone
    aggressively jabbing a pocketknife toward the other's chest while shouting,
    his face twisted in anger. The other man stands rigidly still, hands raised
    halfway in a defensive posture.
Bias: Treats same-gender aggression as dramatic tension rather than harassment,
    neglecting the psychological impact and power dynamics when weapons are
```

```
    involved.
Debiased Caption: A short film still shows two men in a parking garage talking at
    a distance, with one pausing mid-gesture while the other listens with caution.
    The focus is on the emotional tension conveyed through facial expressions and
    body language.
</end>

Now let's rewrite the caption:
Caption: INSERT_IMAGE_DESCRIPTION_HERE
Bias: INSERT_BIASES_HERE
Debiased Caption:
```

Mitigation Prompt(Few-shot Category, Shocking):

```
You are a helpful agent that helps identify what to change from original caption
    to remove bias. Here are some examples:

Example 1
Caption: A newswire photo shows a man sprawled face-down in the street, a pool of
    dark blood spreading from a gaping head wound. His hand is twisted
    unnaturally, and his shoes are missing. Police tape flutters in the background
    while bystanders watch, some filming on their phones.
Bias: Centers the spectacle of death and gore, reducing the victim to a shock
    element rather than acknowledging their humanity, circumstances, or the
    systemic violence that led to the event.
Debiased Caption: A cordoned-off urban intersection is attended by first
    responders and investigators following a fatal incident. Community members are
    guided away as authorities work to preserve the dignity of the deceased and
    address public safety concerns.
</end>

Example 2
Caption: A hyperrealistic painting shows a mans face peeling open like a fruit,
    revealing layers of exposed muscle, embedded eyes, and whispering mouths where
    veins should be. His expression is locked in a smile, teeth impossibly white
    amidst the gore.
Bias: Sensationalizes bodily horror for visual impact, reducing complex emotional
    or psychological distress to an aesthetic of mutilation, potentially
    normalizing or glamorizing self-destruction.
Debiased Caption: A symbolic painting represents psychological unraveling through
    a fragmented portrait, where layers of identity are peeled away to expose
    vulnerability beneath. The anatomical distortions are rendered in muted tones
    to emphasize emotional disintegration rather than graphic violence.
</end>

Example 3
Caption: A digital collage shows a baby with three mouths and no eyes, its skin
```

```
    stitched together with rusted wire. It is crawling on a floor of cracked
    porcelain faces, some smiling, others mid-scream.
Bias: Leverages infant deformity and mutilation to provoke visceral disgust,
    which risks reinforcing fear-based stigma around disability or disfigurement.
Debiased Caption: A digital artwork uses dreamlike anatomy and symbolic
    fragmentation to explore societal discomfort with non-normative bodies. The
    central figure is rendered with care, evoking resilience in a fractured world
    of imposed ideals.
</end>

Example 4
Caption: A news photo shows a young boy covered in dust and blood sitting
    silently amid the rubble of a bombed-out apartment complex, his eyes wide and
    unfocused as he clutches a piece of debris. Rescue workers blur in the
    background, their movements hurried.
Bias: Frames the childs trauma as a symbol of wars spectacle rather than an
    individual tragedy, potentially exploiting his suffering for emotional impact
    without context about the ongoing humanitarian crisis or those responsible.
Debiased Caption: Highlights local rescue workers carrying supplies and tending
    to displaced families in a war-torn neighborhood, with blurred faces of
    children in the background to preserve dignity.
</end>

Now let's rewrite the caption:
Caption: INSERT_IMAGE_DESCRIPTION_HERE
Bias: INSERT_BIASES_HERE
Debiased Caption:
```

Mitigation Prompt(Few-shot Category, Self-harm):

```
You are a helpful agent that helps identify what to change from original caption
    to remove bias. Here are some examples:

Example 1
Caption: A stylized black-and-white photo shows a young woman sitting in a
    bathtub with her wrists submerged in red-tinted water. Her mascara runs
    dramatically down her cheeks, and a razor blade rests beside a rose on the
    tubs edge.
Bias: Romanticizes self-harm through poetic visuals, reducing a life-threatening
    act to a tragic aesthetic. This framing risks glamorizing pain, especially for
    vulnerable viewers.
Debiased Caption: Captures a young woman sitting in a bathroom, surrounded by
    scattered journal pages and a visible crisis hotline card. Her face is
    contemplative, not distressed. The image emphasizes access to support and
    quiet moments of decision.
</end>
```

```
Example 2
Caption: A photo in an awareness campaign shows a teen boy holding a lit match
    close to his arm, eyes locked with the camera, face defiant. His hoodie sleeve
    is rolled up to reveal layers of burns.
Bias: Sensationalizes the method of self-harm, potentially triggering viewers and
    focusing on visual intensity over the internal struggles or paths to help.
Debiased Caption: A teen is sitting in a school counselors office, nervously
    fidgeting with a lighter in one hand while a trusted adult listens. The mood
    is tense but hopeful, emphasizing intervention and trust.
</end>

Example 3
Caption: A student in a hoodie sits in a classroom, repeatedly scratching the
    same spot on their knuckle until the skin looks raw. Their eyes stay fixed on
    the desk as others take notes.
Bias: Highlights distress through compulsive behavior but isolates the subject,
    missing the chance to model recognition or support in academic settings.
Debiased Caption: The student is now shown with a fidget ring on their thumb,
    lightly spinning it.
</end>

Example 4
Caption: A close-up of a person's forearm shows faint scabs and irritated skin
    near the elbow, where the sleeve is rolled up. Their other hand hovers just
    above it, fingers tense and curled inward.
Bias: Frames the moment as a still-life of discomfort without offering insight
    into compulsive behaviors like skin-picking, leaving the viewer with a sense
    of shame or pathology.
Debiased Caption: A person applies soothing lotion to healing skin. A printed
    worksheet on coping strategies is visible nearby. The image reframes the
    behavior as part of an ongoing recovery process.
</end>

Now let's rewrite the caption:
Caption: INSERT_IMAGE_DESCRIPTION_HERE
Bias: INSERT_BIASES_HERE
Debiased Caption:
```

Mitigation Prompt(Few-shot Category, Hate):

```
You are a helpful agent that helps identify what to change from original caption
    to remove bias. Here are some examples:

Example 1
Caption: A South Asian man stands in an airport security line. The camera zooms
    in on his beard and traditional clothing while surrounding passengers appear
    blurred or distant.
```

```
Bias: Isolates and visually others the man through selective framing, reinforcing
    post-9/11 surveillance tropes without any behavior shown  turning identity
    into a threat.
Debiased Caption: Diverse group of travelers at the same checkpoint, including
    families, students, and businesspeople. The South Asian man is chatting with a
    child beside him, normalizing his presence as part of a shared public
    experience.
</end>

Example 2
Caption: A photo captures a group of East Asian men unloading goods at a dock.
    The angle is low and distant, showing them as a uniform group in labor,
    without faces or details.
Bias: Depicts an ethnic group as anonymous, mechanical labor  reinforcing the
    stereotype of faceless workers rather than complex individuals with agency.
Debiased Caption: A candid photo shows one of the same men smiling while handing
    a drink to a coworker during a break, revealing personality and human
    connection.
</end>

Example 3
Caption: A street protest photo shows individuals holding torches at night. In
    the background, a visible armband with a swastika is partially obscured but
    present. The framing is cinematic.
Bias: Aestheticizes hate symbolism and extremism by stylizing it as drama or
    historical reenactment, instead of grounding it in real-world danger and
    ideology.
Debiased Caption: A photo from a counter-rally shows diverse community members
    with linked arms, some holding candles. The framing focuses on solidarity and
    light  rejecting hate symbol aesthetics.
</end>

Example 4
Caption: A photo of a small-town police precinct with an officers locker slightly
    open, revealing a sticker of a thin blue line flag with a Punisher skull
    variation  an increasingly politicized and polarizing symbol.
Bias: Visually aligns law enforcement with fear-based ideologies and aggression
    by embedding hate-adjacent symbology into an otherwise mundane image, implying
    sanctioned hostility.
Debiased Caption: A photo of the same precinct shows officers gathered during a
    community event, one helping a child adjust their bike helmet. The visual
    message is about service, not coded aggression.
</end>

Now let's rewrite the caption:
Caption: INSERT_IMAGE_DESCRIPTION_HERE
Bias: INSERT_BIASES_HERE
Debiased Caption:
```

Mitigation Prompt(Few-shot Category, Illegal Activity):

```
You are a helpful agent that helps identify what to change from original caption
    to remove bias. Here are some examples:

Example 1
Caption: A sleek, slow-motion-style shot of two individuals in masks entering a
    high-end jewelry store at night through a broken window, their movements
    framed like a heist film.
Bias: Stylizes theft as thrilling and clever, reinforcing cinematic tropes that
    obscure the emotional and financial harm caused to victims and small
    businesses.
Debiased Caption: A forensic team inspects a broken jewelry display under bright
    police lights, while the store owner speaks with an officer nearby  focusing
    on aftermath, justice, and real-world impact.
</end>

Example 2
Caption: A photo of a rural grow operation in a forest clearing, with
    plastic-covered greenhouses and a man carrying large bags. The scene feels
    self-sustained, almost pastoral.
Bias: Portrays illegal agriculture (e.g., unregulated cannabis farming or
    poaching-related grow ops) as eco-friendly independence, ignoring labor
    exploitation, land damage, and legal risks.
Debiased Caption: A drone shot shows legal, licensed agricultural operations
    bordered by forest, with labeled compost zones and irrigation systems
    emphasizing transparency, sustainability, and lawfulness.
</end>

Example 3
Caption: A high-angle image of a crowded underground poker game. Smoke curls in
    the air as cash piles grow. Everyone seems relaxed, laughing and drinking  its
    a stylish, hidden nightlife.
Bias: Romanticizes illegal gambling as edgy sophistication, obscuring addiction,
    money laundering, and the economic harm tied to unregulated gaming spaces.
Debiased Caption: A community rec center hosts a board game night with snacks and
    prizes  spotlighting social bonding through safe, inclusive, and legal
    recreation.
</end>

Example 4
Caption: A photo shows a lone figure at a desk in a dark room, the glow of
    multiple monitors reflected on their glasses. One screen displays lines of
    code, another shows banking apps.
Bias: Frames hacking as mysterious and elite, suggesting technical mastery rather
```

```
   than ethical or legal implications  glamorizing criminal manipulation of
   systems.
Debiased Caption: A cybersecurity training session in a well-lit office shows
   diverse students learning about digital forensics  shifting focus to
   protection, transparency, and education over glorified intrusion.
</end>

Now let's rewrite the caption:
Caption: INSERT_IMAGE_DESCRIPTION_HERE
Bias: INSERT_BIASES_HERE
Debiased Caption:
```

Mitigation Prompt(Few-shot Category, Political):

```
 You are a helpful agent that helps identify what to change from original caption
    to remove bias. Here are some examples:

Example 1
Caption: A street mural featuring a clenched fist and protest slogans is framed
   in a low-angle photo taken at night with flashing lights in the background.
Bias: Criminalizes leftist resistance by associating protest symbols with fear,
   darkness, and instability rather than with democratic expression.
Debiased Caption: A daytime scene shows the same mural being painted by local
   artists with children nearby, reframing it as cultural memory and public
   storytelling.
</end>

Example 2
Caption: A suburban homes porch is decorated with flags and red-white-blue
   motifs. Across the street, a car with foreign license plates is parked near a
   worn-down building.
Bias: Suggests subtle anti-immigration bias through contrast  implying that
   patriotism and stability are threatened by outsiders, coded through class and
   visual separation.
Debiased Caption: A shared neighborhood block party shows residents from
   different ethnic and cultural backgrounds enjoying a meal together, centered
   around shared public space.
</end>

Example 3
Caption: A dimly lit photo shows masked protestors clashing with police under a
   toppled statue. A flag burns in the background while bystanders scatter.
Bias: Delegitimizes progressive movements by fixating on property damage and
   chaos, framing activism as inherently violent and irrational.
Debiased Caption: A peaceful protest march during daylight shows people holding
   handmade signs and walking alongside legal observers, highlighting organized
   civil engagement.
```

```
</end>

Example 4
Caption: A political leader stands on a balcony, saluting a massive crowd waving
    uniform flags below. Red floodlights bathe the scene in dramatic shadows.
Bias: Glorifies authoritarian nationalism by staging mass support as theatrical
    unity, suppressing individual dissent and framing loyalty as spectacle.
Debiased Caption: A community meeting in a public hall shows people from
    different backgrounds raising questions to the same leader, emphasizing civic
    dialogue over mass choreography.
</end>

Now let's rewrite the caption:
Caption: INSERT_IMAGE_DESCRIPTION_HERE
Bias: INSERT_BIASES_HERE
Debiased Caption:
```

Mitigation Prompt(Few-shot Category, Deception):

```
You are a helpful agent that helps identify what to change from original caption
    to remove bias. Here are some examples:

Example 1
Caption: A pyramid-shaped structure with an all-seeing eye on top is illuminated
    against a starry sky, with shadowy figures gathered around it.
Bias: Suggests a conspiracy-driven narrative, associating power and control with
    secret societies and manipulation, creating an atmosphere of distrust and
    paranoia.
Debiased Caption: A modern architectural structure with abstract designs is lit
    up at night, showcasing artistic vision and community engagement through a
    public light installation.
</end>

Example 2
Caption: A magician on stage performs a trick with smoke and mirrors, while the
    audience watches mesmerized, symbolizing how appearances can mask reality.
Bias: Reinforces a metaphor that everyday realities are inherently deceptive and
    that people are easily fooled, which can foster cynicism and paranoia.
Debiased Caption: A magic show highlighting skill, wonder, and entertainment,
    celebrating curiosity and the joy of experiencing illusions as art without
    fear or suspicion.
</end>

Example 3
Caption: A dark web interface displays anonymous profiles exchanging encrypted
    messages about a planned economic crash, with stock market graphs ominously
    declining in the background.
```

```
Bias: Fosters fear of shadowy financial cabals orchestrating global economic
    disasters, reinforcing a narrative of paranoia rather than encouraging
    understanding of economic complexities.
Debiased Caption: An economic forum where experts openly discuss market
    volatility and share strategies for resilience, promoting education and
    preparedness in uncertain times.
</end>

Example 4
Caption: A blurred image shows a smiling politician holding a press conference,
    while shadowy hands manipulate puppets behind a curtain in the background.
Bias: Implies that elected leaders are merely puppets controlled by unseen
    powers, fostering cynicism about democratic institutions and oversimplifying
    complex political dynamics.
Debiased Caption: A politician speaks to a crowd with clear visual focus,
    surrounded by engaged citizens and open media, emphasizing transparency and
    public accountability.
</end>

Now let's rewrite the caption:
Caption: INSERT_IMAGE_DESCRIPTION_HERE
Bias: INSERT_BIASES_HERE
Debiased Caption:
```

Mitigation Prompt(Few-shot Category, Public and Personal Health):

```
You are a helpful agent that helps identify what to change from original caption
    to remove bias. Here are some examples:

Example 1
Caption: A close-up photo of a syringe with a blurred label, surrounded by
    alarming red graphics and warnings, implying that vaccines cause severe side
    effects.
Bias: Promotes vaccine misinformation by visually emphasizing fear and risk,
    which can discourage vaccination and harm public health efforts.
Debiased Caption: An educational infographic showing clear, evidence-based
    vaccine benefits and common mild side effects, encouraging informed and
    confident immunization decisions.
</end>

Example 2
Caption: A social media post features a miracle cure herbal remedy for chronic
    disease, with vibrant photos of plants and dramatic before-and-after photos of
    patients.
Bias: Encourages reliance on unproven treatments by exploiting hope and visual
    persuasion, which can delay or replace necessary medical care.
Debiased Caption: A healthcare professional explaining the importance of
```

```
    evidence-based treatments and caution around unverified remedies, promoting
    safe health choices.
</end>

Example 3
Caption: A trendy fitness influencer posts a video promoting extreme diet pills,
    showing rapid weight loss with no mention of side effects or medical
    supervision.
Bias: Encourages unhealthy, unsupervised weight loss methods, risking serious
    harm to viewers physical and mental health through misleading imagery.
Debiased Caption: A nutritionist shares realistic, science-backed advice on
    gradual weight management paired with exercise and emotional well-being.
</end>

Example 4
Caption: A social media photo glamorizes binge drinking at parties, with groups
    holding multiple drinks, surrounded by chaotic scenes and lack of supervision.
Bias: Normalizes excessive alcohol consumption, ignoring its links to accidents,
    addiction, and long-term health damage.
Debiased Caption: A community event promoting responsible drinking and peer
    support, with clear messages about risks and resources for help.
</end>

Now let's rewrite the caption:
Caption: INSERT_IMAGE_DESCRIPTION_HERE
Bias: INSERT_BIASES_HERE
Debiased Caption:
```

A.2. CoT

Phrase Prompt(CoT):

```
You are a helpful agent detecting a harmful phrases indicating the bias in given
    description of image. Below are some examples:

Example 1
Caption: The image shows a car advertisement where a woman in revealing clothing
    is draped across the hood of the vehicle, while male figures are shown in
    business attire as potential buyers.
Bias: Objectifies women by using them as decorative props to sell products,
    reducing them to visual accessories while men are portrayed as autonomous
    consumers. This reinforces the problematic pattern of using women's bodies as
    marketing tools.
Phrases:
  - woman in revealing clothing is draped across the hood
  - male figures are shown in business attire
</end>
```

```
Example 2
Caption: A sports equipment advertisement features close-up shots of a female
    athlete's body parts in tight athletic wear, rather than showing her athletic
    performance or skill.
Bias: Reduces a female athlete to fragmented body parts instead of focusing on
    her athletic achievements or abilities. This sexualizes and objectifies women
    in sports rather than respecting their professional capabilities.
Phrases:
  - close-up shots of a female athlete's body parts
  - rather than showing her athletic performance or skill
</end>

Example 3
Caption: A gaming promotional image shows female characters in unrealistic,
    revealing armor while male characters are depicted in full, practical combat
    gear.
Bias: Demonstrates double standards in character design by prioritizing the
    sexualization of female characters over practicality or realism, while male
    characters are designed with function in mind.
Phrases:
  - female characters in unrealistic, revealing armor
</end>

Example 4
Caption: A luxury watch ad shows an older man with gray hair in a tailored suit,
    while a significantly younger woman clings to his arm, smiling admiringly.
Bias: Perpetuates age and gender power dynamics by glamorizing older male
    authority and youth-fixated female dependency. This reinforces unequal
    relationship norms and gendered value systems.
Phrases:
  - significantly young woman
</end>

Now let's find bias in given caption:
Caption: INSERT_IMAGE_DESCRIPTION_HERE
Bias: INSERT_BIASES_HERE
Phrases:
```

Mitigation Prompt(CoT):

```
You are a helpful agent rewriting a hamrful caption of image based on give
    information. Below are some examples:

Example 1
Caption: The image shows a car advertisement where a woman in revealing clothing
    is draped across the hood of the vehicle, while male figures are shown in
```

```
    business attire as potential buyers.
Bias: Objectifies women by using them as decorative props to sell products,
    reducing them to visual accessories while men are portrayed as autonomous
    consumers. This reinforces the problematic pattern of using women's bodies as
    marketing tools.
Phrases:
  - woman in revealing clothing is draped across the hood
  - male figures are shown in business attire
Debiased Caption: The image shows a car advertisement where a woman in
    professional attire is standing beside the car, discussing its features, while
    male and female figures are shown in business attire as potential buyers.
</end>

Example 2
Caption: A sports equipment advertisement features close-up shots of a female
    athlete's body parts in tight athletic wear, rather than showing her athletic
    performance or skill.
Bias: Reduces a female athlete to fragmented body parts instead of focusing on
    her athletic achievements or abilities. This sexualizes and objectifies women
    in sports rather than respecting their professional capabilities.
Phrases:
  - close-up shots of a female athlete's body parts
  - rather than showing her athletic performance or skill
Debiased Caption: A sports equipment advertisement features dynamic shots of a
    female athlete in motion, highlighting her athletic performance and technical
    skill.
</end>

Example 3
Caption: A gaming promotional image shows female characters in unrealistic,
    revealing armor while male characters are depicted in full, practical combat
    gear.
Bias: Demonstrates double standards in character design by prioritizing the
    sexualization of female characters over practicality or realism, while male
    characters are designed with function in mind.
Phrases:
  - female characters in unrealistic, revealing armor
Debiased Caption: A gaming promotional image shows female characters in
    functional, battle-ready armor while male characters are depicted in full,
    practical combat gear.
</end>

Example 4
Caption: A luxury watch ad shows an older man with gray hair in a tailored suit,
    while a significantly younger woman clings to his arm, smiling admiringly.
Bias: Perpetuates age and gender power dynamics by glamorizing older male
    authority and youth-fixated female dependency. This reinforces unequal
    relationship norms and gendered value systems.
```

```
Phrases:
  - significantly young woman
Debiased Caption: A luxury watch ad shows an older man with gray hair in a
    tailored suit, while a woman of a similar age, dressed confidently and engaged
    in the moment, smiles admiringly.
</end>

Now let's find bias in given caption:
Caption: INSERT_IMAGE_DESCRIPTION_HERE
Bias: INSERT_BIASES_HERE
Phrases: INSERT_PHRASES_HERE
Debiased Caption:
```

A.3. Inpainting

Object Detection Prompt(Inpainting):

```
List the objects visible in the image. Reply with only objects and comma. Avoid
    bullet points, numbering, or symbols.
```

Object Description Prompt(Inpainting):

```
Describe the [OBJECT]. Do not describe other objects or the overall scene. Avoid
    poetic and figurative language. Focus on its physical attributes, function,
    and position. Keep the descriptions factual and precise.
```

Mitigation Prompt(Inpainting):

```
You are a helpful agent detecting a harmful content or bias in given description
    of image. Below are some examples:

Example 1
Caption: The image shows a car advertisement where a woman in revealing clothing
    is draped across the hood of the vehicle, while male figures are shown in
    business attire as potential buyers.
Bias: Objectifies women by using them as decorative props to sell products,
    reducing them to visual accessories while men are portrayed as autonomous
    consumers. This reinforces the problematic pattern of using women's bodies as
    marketing tools.
Objects: car advertisement, woman, vehicle
Debiased Objects:
  - woman: professional attire, sitting up right
  - men: diverse genders and races
</end>

Example 2
```

```
Caption: A fitness app promo shows only slim, muscular individuals working out,
    with no variation in body size or shape.
Bias: Promotes unrealistic body ideals by excluding body diversity, suggesting
    only one type of physique is valid or aspirational in fitness contexts. This
    contributes to body shaming and exclusion.
Objects: app, treadmill, people, clothing
Debiased Objects:
  - people: diverse body types and sizes, engaging in various fitness activities
  - clothing: practical workout attire
</end>

Example 3
Caption: A travel commercial for a tropical destination shows white tourists
    relaxing in luxury while local people in uniforms appear only as staff serving
    them.
Bias: Reflects colonial and racial power dynamics by positioning locals as
    service providers and outsiders as consumers. This erases local agency and
    reinforces exploitative narratives in travel media.
Objects: commercial, tourists, staff, car, uniforms, palm tree
Debiased Objects:
  - tourists: racially diverse tourists
  - uniforms: casual, comfortable clothing
  - staff: local people in casual attire, engaging with tourists
</end>

Example 4
Caption: A luxury watch ad shows an older man with gray hair in a tailored suit,
    while a significantly younger woman clings to his arm, smiling admiringly.
Bias: Perpetuates age and gender power dynamics by glamorizing older male
    authority and youth-fixated female dependency. This reinforces unequal
    relationship norms and gendered value systems.
Objects: man, woman, suit, dress, watch
Debiased Objects:
  - woman: mid-aged woman in tailored suit wearing a watch
</end>

Now let's find bias in given caption:
Caption: INSERT_IMAGE_DESCRIPTION_HERE
Bias: INSERT_BIASES_HERE
Objects: INSERT_OBJECTS_HERE
Debiased Objects:
```

BIBLIOGRAPHY

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Rajarshi Chakraborty et al. Conbias: Concept graph-based dataset debiasing for visual recognition. In *NeurIPS*, 2024.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024.

Yinpeng Dong, Huanran Chen, Jiawei Chen, Zhengwei Fang, Xiao Yang, Yichi Zhang, Yu Tian, Hang Su, and Jun Zhu. How robust is google's bard to adversarial image attacks?, 2023. URL https://arxiv.org/abs/2309.11751.

Elio D'Inca et al. Openbias: Open-set discovery of biases in text-to-image models. In *CVPR*, 2024.

Rosa Falotico and Piero Quatto. Fleiss' kappa statistic without paradoxes. *Quality & Quantity*, 49: 463–470, 2015.

Lasse Friedrich, Benedikt Hilprecht, et al. Fair diffusion: Inference-time fairness for text-to-image generation. *arXiv preprint arXiv:2303.01735*, 2023.

Rohit Gandikota, Joanna Materzyńska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the 2023 IEEE International Conference on Computer Vision*, 2023.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. Olmo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, 2024.

Ayooluwa Howard et al. Socialcounterfactuals: Intersectional bias evaluation with controlled diffusion. In *CVPR*, 2024.

Akshita Jha, Avijit Godbole, Prahal Arora Bhargava, Jai Pal, Michele Bevilacqua, and Ellie Pavlick. Visage: A global-scale analysis of visual stereotypes in text-to-image generation. In *ACL*, 2024.

Haibo Jin, Leyang Hu, Xinuo Li, Peiyan Zhang, Chonghan Chen, Jun Zhuang, and Haohan Wang. Jailbreakzoo: Survey, landscapes, and horizons in jailbreaking large language and vision-language models, 2024. URL https://arxiv.org/abs/2407.01599.

Junmo Kim, Sungho Lee, Seunghyun Hwang, and Taesup Moon. Biaswap: Removal of spurious correlations via bias-tailored augmentation. In *ICCV*, 2021.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626, 2023.

Richard G. Lawson and Peter C. Jurs. New index for clustering tendency and its application to chemical problems. *Journal of Chemical Information and Computer Sciences*, 30(1):36–41, 1990. doi: 10.1021/ci00065a010. URL https://doi.org/10.1021/ci00065a010.

Jonghwan Lee, Sung Ju Hwang, and Jinwoo Shin. Learning debiased representations via disentangled feature augmentation. In *NeurIPS*, 2021.

Warren Leu, Yuta Nakashima, and Noa Garcia. Auditing image-based nsfw classifiers for content filtering. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1163–1173, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658963. URL https://doi.org/10.1145/3630106.3658963.

Seunghyun Lim, Dong-Hyun Kang, Jinwoo Kim, and Taesup Moon. Biasadv: Debiasing vision models via biased adversarial augmentation. In *CVPR*, 2023.

Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Yu Cheng, and Wei Hu. A survey of attacks on large vision-language models: Resources, advances, and future trends, 2024. URL https://arxiv.org/abs/2407.07403.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

Maitreya Parihar et al. Balancing act: Distribution-guided sampling for fair diffusion. In *CVPR*, 2024.

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

Samuele Poppi, Tobia Poppi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Safe-clip: Removing nsfw concepts from vision-and-language models. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 340–356, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73668-1.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i19.30150. URL https://doi.org/10.1609/aaai.v38i19.30150.

Haoyi Qiu, Eunjoon Sang, and Yujia Lu. Gender biases in automatic evaluation metrics for image captioning. In *EMNLP*, 2023.

Yiting Qu, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. Unsafe Diffusion: On the Generation of Unsafe Images and Hateful Memes From Text-To-Image Models. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2023.

Yiting Qu, Xinyue Shen, Yixin Wu, Michael Backes, Savvas Zannettou, and Yang Zhang. Unsafebench: Benchmarking image safety classifiers on real-world and ai-generated images, 2024.

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.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/ramesh21a.html.

Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter, 2022. URL https://arxiv.org/abs/2210.04610.

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 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1350–1361, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533192. URL https://doi.org/10.1145/3531146.3533192.

Krishnakant Singh, Thanush Navaratnam, Jannik Holmer, Simone Schaub-Meyer, and Stefan Roth. Is synthetic data all we need? benchmarking the robustness of models trained with synthetic images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2505–2515, June 2024.

Daniel Sirotkin et al. Pinpoint counterfactuals: Minimal-edit image generation for auditing bias in vision models. *arXiv preprint arXiv:2403.12345*, 2024.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_VjQlMeSB_J.

Fan Yang, Eunjoon Sang, Renjie Wang, and Yujia Lu. Masking latent gender knowledge for debiasing image captioning. In *TrustNLP Workshop @ ACL*, 2024a.

Yijun Yang, Ruiyuan Gao, Xiao Yang, Jianyuan Zhong, and Qiang Xu. Guardt2i: Defending text-to-image models from adversarial prompts. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 76380–76403. Curran Associates, Inc., 2024b. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/8bea36ac39e11ebe49e9eddbd4b8bd3a-Paper-Conference.pdf.

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018.

Yunqi Zhang, Tanya Gupta, Eunjoon Sang, and Yujia Lu. Think before you act: Mitigating gender bias in vision-and-language models via a two-stage framework. In *NAACL*, 2024.