This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

# SCoFT: Self-Contrastive Fine-Tuning for Equitable Image Generation

Zhixuan Liu[1] Peter Schaldenbrand[1] Beverley-Claire Okogwu[1] Wenxuan Peng[2]
Youngsik Yun[3] Andrew Hundt[1] Jihie Kim[3] Jean Oh[1*]
[1]Carnegie Mellon University [2]Nanyang Technological University [3] Dongguk University
[*]Corresponding author


(a) *"Photo of a traditional building, in [Culture]"*

(b) *"Two people wearing traditional clothing, in [Culture]"*

Figure 1. Comparison between Stable Diffusion with and without our proposed fine-tuning approach, SCoFT, on our proposed CCUB dataset. Stable Diffusion perpetuates harmful stereotypes that assume dirty buildings are representative of some nations, and often generates regionally irrelevant designs. By contrast, our approach decreases stereotypes and improves cultural relevance of generated images.

## Abstract

*Accurate representation in media is known to improve the well-being of the people who consume it. Generative image models trained on large web-crawled datasets such as LAION are known to produce images with harmful stereotypes and misrepresentations of cultures. We improve inclusive representation in generated images by (1) engaging with communities to collect a culturally representative dataset that we call the Cross-Cultural Understanding Benchmark (CCUB) and (2) proposing a novel Self-Contrastive Fine-Tuning (SCoFT, pronounced /sôft/) method that leverages the model's known biases to self-improve. SCoFT is designed to prevent overfitting on small datasets, encode only high-level information from the data, and shift the generated distribution away from misrepresentations encoded in a pretrained model. Our user study conducted on 51 participants from 5 different countries based on their self-selected national cultural affiliation shows that fine-tuning on CCUB consistently generates images with higher cultural relevance and fewer stereotypes when compared to the Stable Diffusion baseline, which is further improved with our SCoFT technique. Resources and code are at* `https://ariannaliu.github.io/SCoFT`.

## 1. Introduction

Representation matters. In media, studies repeatedly show that representation affects the well-being of its viewers [8, 12, 53]. Representation can positively affect viewers by providing them with role models that they identify with, but it can also negatively affect viewers by creating harmful, stereotypical understandings of people and culture [7]. When people are accurately represented in media, it allows people to properly understand cultures without harmful stereotypes forming [11, 37]. To date, unfortunately, many media-generating AI models show poor representation in their results [32, 39] and have been deployed for cases with negative impacts on various groups, such as non-consensual images [34]. Such issues stem from every stage of an ML system's lifecycle [56], with key steps being algorithms and the large training datasets gathered by crawling the Internet with inadequate filtering supervision. They contain or amplify malignant stereotypes and ethnic slurs, among other problematic content [4], and impact a range of applications [22]. Researchers have shown that large datasets such as LAION-400M [51] used to train many text-to-image synthesis models, including Stable Diffusion [46], center the Global North [4, 5, 32] and struggle to accurately depict cultures from the Global South as shown in Figure 1.

To ensure models better represent culture and more accurately represent the world, we introduce a new task of *culturally-aware* image synthesis with the aim of addressing representation in image generation: generating visual content that is perceived to be more representative of national cultural contexts. Our overarching goal is to improve the well-being of viewers of the AI-generated images with particular attention to those who are from a selection of groups marginalized by existing methods.

Our research question is, how can effective, existing text-to-image models be improved to become more culturally representative and thus less offensive? Since it may be infeasible to vet billions of training examples for accurate cultural content, we hypothesize that a small dataset that is veritably representative of a culture can be used to prime pre-trained text-to-image models to guide the model toward more culturally accurate content creation. To verify the hypothesis, we collected a dataset of image and caption pairs for 5 cultures. For each culture, data was collected by people who self-selectedly affiliated with that culture as they are the people who properly understand it and are most affected by its misrepresentations. We call this the Cross-Cultural Understanding Benchmark (CCUB) dataset which comprises 150 - 200 images for each culture each with a manually written caption as shown in Figure 2.

To encode the culturally representative information in CCUB into a pre-trained model, we propose to fine-tune the model with the new dataset. Existing fine-tuning techniques work well for low-level adaptations such as style changes or introducing new characters to models [21], but we show that these methods struggle to encode high-level, complex concepts such as culture. Additionally, fine-tuning on small datasets, such as CCUB, can lead to overfitting.

Unlike concept editing tasks [15, 16] with specific image editing directions, depicting cultural accuracy remains more abstract and challenging. We propose a novel fine-tuning approach, Self-Contrastive Fine-Tuning (SCoFT, pronounced /sôft/), to address these issues. SCoFT leverages the pre-trained model's cultural misrepresentations against itself. We harness the intrinsic biases of large pre-trained models as a rich source of counterexamples; shifting away from these biases gives the model direction towards more accurate cultural concepts. Image samples from the pre-trained model are used as negative examples, and CCUB images are used as positive examples, to train the model to discern subtle differences. We de-noise latent codes in several iterations, project them into the pixel space, and then compute the contrastive loss. The loss is backpropagated to the diffusion model UNet and optimized to push generated samples towards the positive distribution and away from the negative. This is all done in a perceptual feature space so that the model learns more high-level, complex features from the images.

To evaluate our results we recruited participants who identify as being a member of the cultural communities in the CCUB dataset to rank images generated by Stable Diffusion with or without the proposed self-contrastive fine-tuning on CCUB. Fine-tuning on CCUB was found to decrease offensiveness and increase the cultural relevance of generated results based on 51 participants across five cultures and 766 image comparisons. Our proposed SCoFT approach further improved these results. We share the findings from our experiments to provide a basis for an important aspect of AI-generated imagery: that cultural information should be accurately presented and celebrated equitably. Our contributions are as follows:

1. The introduction of culturally-aware text-to-image synthesis as a valuable task within text-to-image synthesis;
2. The Cross-Cultural Understanding Benchmark (CCUB) dataset of culturally representative image-text pairs across 5 countries; and
3. Self-Contrastive Fine-Tuning (SCoFT), a novel technique for encoding high-level information into a pre-trained model using small datasets.

## 2. Related Work

**Cultural Datasets** Various efforts have been made to evaluate and document the impacts of datasets[23, 50, 60]. Dollar Street [45] aimed to capture accurate demographic information based on socioeconomic features, such as everyday household items and monthly income of 63 countries worldwide. However, this dataset offers less diverse scenarios, as most of its images are indoor views with limited cultural features. Likewise, the WIT [55] and MLM [1] strive for cultural representation but use Wikipedia/WikiData sources for images that are not representative of all aspects of culture and are over-saturated with old, streotypical images. Other works, capturing the idea of a diverse dataset, aim to reduce stereotypical bias through self-curation [10] or cultural-driven methods [17, 36, 38], inspiring our data collection methodology.

The MaRVL dataset [30], for example, was curated by people who identify as affiliated with one of several particular cultures, MaRVL was developed to mitigate biases for reasoning tasks covering popular concepts and is unsuitable for text-to-image synthesis. Our dataset was also designed to cover a diverse sample of cultures and engage with people who are native, but it is specifically curated for vision-language tasks and diverse cultural categories.

**Fine-Tuning Text-to-Image Models** Fine-tuning pre-trained text-to-image synthesis models with a new dataset is an approach to encode additional new simple concepts and content into the model [21, 27, 31, 41, 48]. But culture is a complex, high-level concept that poses many challenges when attempting to fine-tune a model to understand it.

**Fine-Tuning on Cultural Data.** Prior work in adapting pre-trained models to be culturally relevant found some success using millions of culturally relevant text-image pairs, as in ERNIE-ViLG 2.0. [13] and Japanese Stable Diffusion [54]. The size of these training datasets leads to better cultural representations of Japan and China, but they are not easy to deploy universally, as these approaches require millions of training examples which cannot possibly be met for cultures with less internet presence. Besides, these datasets are so large that it is infeasible to 100% vet them for harmful and stereotypical information [43]. We propose a fine-tuning technique that adapts pre-trained models to learn complex, elusive concepts, namely culture, from small datasets.

**Fine-Tuning Stable Diffusion in the Pixel Space.** Latent diffusion models are customarily trained in the latent space, however, the latent codes can be decoded into images differentiably. Multiple works compute losses on the decoded images to optimize over the input latent code [59] or the decoder weights [2]. DiffusionCLIP [25] optimizes UNet parameters using losses computed on image outputs, however, this is performed on a non-latent diffusion model. High parameter count in latent diffusion models complicates gradient recording through multiple UNet passes, hindering decoding into pixel space. We propose a novel method to reduce the computation graph size, facilitating tractable back-propagation of loss through Stable Diffusion.

**Perceptual Similarity.** Perceptual metrics, such as LPIPS [62], and recent foundation models for perceptual similarity [14, 29, 40, 42], have been shown to align more closely with human perception than pixel space Euclidean distance [49]. Perceptual similarity ignores low-level differences and captures high-level details, which are more important for complex concepts such as cultures. To our knowledge, no other work has trained a latent diffusion model using perceptual loss, likely due to the technical challenges of back-propagating loss through the diffusion process and latent decoder, which we address in this paper.

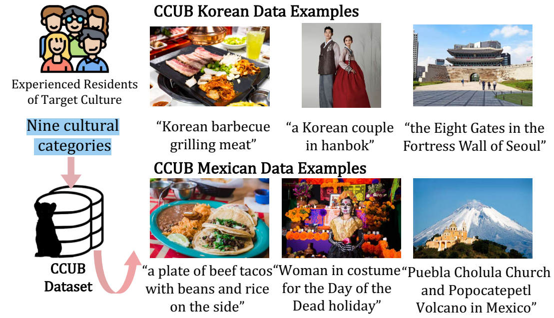

Figure 2. CCUB dataset cultural image and description samples.

## 3. CCUB Dataset

*Can a CCUB tame a LAION?* As opposed to the LAION [52] dataset which scraped images and captions from the internet with minimal supervision leading to a prominence in harmful content, our CCUB dataset was collected by hand by the people most affected by cultural misrepresentations in text-to-image synthesis.

Following the definition of culture in [18], [30], and [24], nine categories are used to represent cultural elements in our dataset: architecture (interior and exterior), city, clothing, dance music and visual arts, food and drink, nature, people and action, religion and festival, utensils and tools. The categories are further divided into traditional and modern to reflect how cultural characteristics change over time.

For each culture, we recruited at least 5 people with at least 5 years of experience living in one of 5 countries (China, Korea, India, Mexico, and Nigeria) to each provide 20-30 images and captions. The images were collected either by collecting image links from Google searches or the collectors' own photographs. Fig. 2 has selected samples of our CCUB dataset. More details are in supplement Sec. 9.

## 4. Method

Given our CCUB dataset, our objective is to alter Stable Diffusion to have a more accurate understanding of a given culture. We next offer a brief background on training latent diffusion models (Sec. 4.1), a modification to a regularization loss to prevent over-fitting (Sec. 4.2), a novel approach to computing perceptual loss on decoded images (Sec. 4.3), and a method to contrastively use Stable Diffusion's misrepresentations of culture to refine itself (Sec. 4.4). **SCoFT.** Our full fine-tuning approach, SCoFT, is a weighted sum of all of the following loss functions:

### 4.1. Latent Diffusion Model Loss

Diffusion models are latent variable models that sample from a desired distribution by reversing a forward noising process. The latent noise at any timestep $t$ is given by $\mathbf{z}_t = \sqrt{\alpha_t}\mathbf{z}_0 + \sqrt{1-\alpha_t}\epsilon$, where $\mathbf{z}_0$ is the latent variable

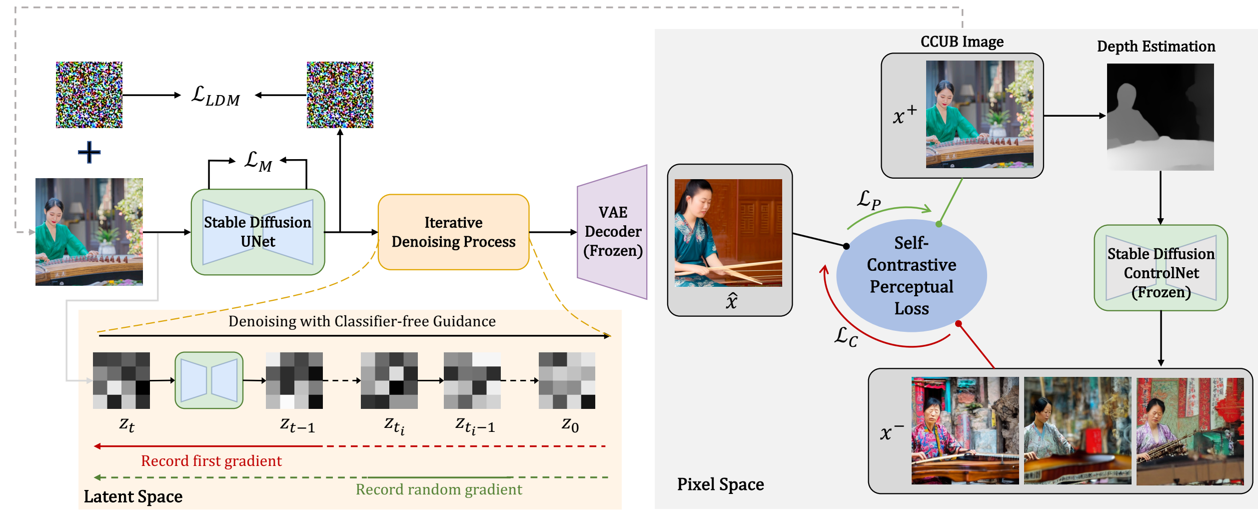


Figure 3. **SCoFT Overview.** A conventional fine-tuning loss, $\mathcal{L}_{LDM}$, and memorization penalty loss, $\mathcal{L}_M$, are computed in the Stable Diffusion latent space using images and captions from our CCUB dataset. After 20 denoising steps, the latent space is decoded. Perceptual features are extracted from the generated image and compared contrastively to CCUB images as positive and non-fine-tuned Stable Diffusion images as negative examples to form our Self-Contrastive Perceptual Loss, $\mathcal{L}_C$.

encoded from the real image and $\alpha_t$ is the strength of the gaussian noise $\epsilon$. We are interested in the pretrained denoising network $\epsilon_\theta(\mathbf{z}_t, \mathbf{c}, t)$, which denoises the noisy latent to get $\mathbf{z}_{t-1}$ conditioned on the text $\mathbf{c}$. The training objective for the denoiser $\epsilon_\theta$ is minimizing the noise prediction:

$$\mathcal{L}_{LDM}(\mathbf{z}, \mathbf{c}) = \mathbb{E}_{\epsilon, \mathbf{z}, \mathbf{c}, t}\left[w_t \left\|\epsilon - \epsilon_\theta\left(\mathbf{z}_t, \mathbf{c}, t\right)\right\|\right] \tag{1}$$

where $w_t$ is the weight to control the noise schedules. A pretrained model can be further trained, or fine-tuned, using the original objective with a new dataset. We use Low-Rank Adaptation (LoRA) [21] to reduce memory demands and over-fitting by training new low-rank weight matrices.

### 4.2. Memorization Loss

Despite CCUB's rich cultural content, its size still remains small compared to LAION. The challenges of language drift and decreased output diversity are prevalent issues encountered during this few-shot fine-tuning [47]. Moreover, CCUB's text captions can be highly specific, as shown in Fig. 2. Fine-tuning solely on the CCUB dataset using Equation 1 may lead to the undesirable outcome of reproducing training data as shown in Fig. 6. Our approach is inspired by [26], which proposed a model-based concept ablation by letting the model memorize the mapping between newly generated anchor images $\mathbf{x}^{\varnothing}$ and $\mathbf{c}^*$, except we focus on preventing memorization during fine-tuning on a small dataset (e.g., CCUB). We harness the property of BLIP automatic caption $\mathbf{c}_{blip}$ on $\mathbf{x}_{ccub}$, which is the naive version of our cultural text prompt $\mathbf{c}_{ccub}$, to regularize the model outputs conditioned on $\mathbf{c}_{ccub}$, as shown in Fig. 6. To achieve this, we introduce a memorization penalty loss leveraging BLIP [28] generated captions of CCUB images. We utilize multiple BLIP captions $\{\mathbf{c}_{blip}\}$ to regularize the one-on-one mapping between CCUB images and cultural captions:

$$\begin{aligned}\mathcal{L}_M(\mathbf{x}_{ccub}, \mathbf{c}_{ccub}, \mathbf{c}_{blip}) = \mathbb{E}_{\epsilon,t}[||\epsilon(\mathbf{x}_{ccub}, \mathbf{c}_{ccub}, t) \\ -\mathbb{E}_i[\epsilon(\mathbf{x}_{ccub}, \mathbf{c}^i_{blip}, t).\,\mathrm{sg}()]||],\end{aligned} \tag{2}$$

where $.\,\mathrm{sg}()$ stands for a stop-gradient operation of the current network to reduce memory cost.

### 4.3. Perceptual Loss

$\mathcal{L}_{\text{LDM}}$ and $\mathcal{L}_M$ are loss functions that operate on the latent codes within a diffusion model. Operating directly on these latent codes is ideal for fine-tuning models for adding simple concepts such as adding a character's appearance to the model. For adding more complex, abstract concepts, like culture, we propose to decode the latent space into the pixel space in order to utilize pre-trained perceptual similarity models. We propose to use a perceptual loss, $\mathcal{L}_P$, which is computed as the difference in extracted perceptual features between the decoded, generated image, $\hat{x}$, and an image from a training set, $x$:

$$\mathcal{L}_P(\hat{x}, x) = \mathbb{E}_{\hat{x}, x}[\mathcal{S}(\hat{x}, x; f_\theta)] \tag{3}$$

where $\mathcal{S}$ is some perceptual similarity function and $f_\theta$ is a pretrained feature extractor.

**Backpropagation through sampling.** State-of-the-art perceptual models typically process inputs in the pixel space. In contrast, Stable Diffusion is fine-tuned in a latent space. To fulfill our objective function, an intuitive strategy entails iteratively denoising latent features and then decoding them back into the pixel space for use with perceptual models. Concurrent work [9] denoises the stable diffusion latent from Gaussian noise into the pixel space based solely on text prompts and by optimizing the Stable Diffusion UNet with a reward function computed using the decoded image. Instead, our approach starts from the latent code at timestep $t$: $\mathbf{z}_t = \sqrt{\alpha_t}\mathbf{z}_0 + \sqrt{1-\alpha_t}\epsilon$, making it coupled with the Stable Diffusion fine-tuning process. We utilize classifier-free guidance, iteratively denoising the latent code according to $(1+w(t))\epsilon(\mathbf{z}_t, \mathbf{c}_{ccub}, t) - w(t)\epsilon(\mathbf{x}_t, \varnothing, t)$, where $w(t)$ is the guidance scale and $\varnothing$ is a null text embedding. This enables the model to generate images conditioned on cultural text prompts and unveils its cultural understanding, as illustrated in Fig. 3. In practice, we denoise $\mathbf{z}_t$ for 20 timesteps. Starting from $\mathbf{z}_t$, our method ensures that the denoised image aligns with the same pose and structure as the original training image $\mathbf{x}_0$. This facilitates a more meaningful comparison for perceptual loss and subsequent self-contrastive perceptual loss, revealing cultural differences.

Directly backpropagating through the multiple UNet denoising iterations, the latent space decoder, and the perceptual model incurs significant memory and time costs. To address this, we selectively record the gradient on a single denoising step and employ stopgrad on other denoising steps. Our findings indicate that recording the gradient from the first step has the most significant impact on refining the model's cultural understanding. Further comparisons on gradient recording and perceptual model backbone are detailed in Sec. 6.

### 4.4. Self-Contrastive Perceptual Loss

To further improve Perceptual Loss, we raise an intriguing question: Can Stable Diffusion leverage its intrinsic biases to refine its own? We seek to leverage the model's prior of its cultural understanding and propose a contrastive learning approach. Utilizing our CCUB dataset, we designate the positive examples to be $\{\mathbf{x}^+|\mathbf{x}_{ccub}\}$ representing preferred features. To unveil the cultural biases within Stable Diffusion, we employ images generated by the model itself as negative samples.

It is imperative to ensure that the generated negative samples share a high-level similarity with positive samples, such as pose and structure, thereby emphasizing that the primary distinctions lie in the diffusion model's perception of cultural features. We achieve this by incorporating a pre-trained ControlNet [61] module $\Theta_c$, conditioned on the estimated depth of $\{\mathbf{x}^+\}$, into Stable Diffusion. As depicted in Fig. 3, negative examples are obtained as $\{\mathbf{x}_i^-|\Theta_c(\mathcal{D}(\mathbf{x}^+), \mathbf{c})\}$, where $\mathcal{D}$ is the MiDaS [6, 44] depth estimator, and $\mathbf{c}$ represents $\mathbf{c}_{blip}$ followed by a cultural suffix (e.g., "in Korea"). In practice, we generate 5 negative examples for each positive example, then utilize DreamSim to filter out false negatives similar to the positive instances.

To enhance the cultural fine-tuning process, our objective is to ensure that images generated by the current model, denoted as $\hat{x}^0_{\theta_t}$ have closer perceptual distances to positive examples and farther distances from negative examples. This reinforces the model's alignment with preferred cultural features, distancing itself from undesirable biases in negative examples, indicated by: $\mathcal{S}(\hat{x}^0_{\theta_t}, x^+; f_\theta) > \mathcal{S}(\hat{x}^0_{\theta_t}, x_i^-; f_\theta)$, where $\mathcal{S}$ is some perceptual similarity function and $f_\theta$ is a pre-trained feature extractor. Thus, we formulate this objective, which we call Self-Contrastive Perceptual Loss, using triplet loss:

$$\begin{aligned}\mathcal{L}_C(\hat{x}, x^+, x^-) = \mathbb{E}_{\hat{x}, x^+, x^-}[\max(\mathcal{S}(\hat{x}, x^+; f_\theta) \\ -\lambda\mathcal{S}(\hat{x}, x^-; f_\theta) + m, 0)]\end{aligned} \tag{4}$$

where $\lambda$ denotes the weights on negative examples, $m$ is the constant margin between positive and negative examples, and $f_\theta$ is a feature extractor. We evaluate a variety of state-of-the-art perceptual embeddings and report comparison results in Section 6.

## 5. Experiments

**User Survey.** Our goal of improving the cultural perception of generated images is a subjective metric largely determined by members of a given identity group. To evaluate our performance on this criterion, we recruited people with at least 5 years of cultural experience in each of the 5 countries with survey questions specific to their self-selected national cultural affiliation. A single page of the survey form provides one description (prompt) and one image made by four different generators using a common random seed, for four total images. We compare four image generators: Stable Diffusion against three fine-tuned ablations. All fine-tunings were performed with the CCUB dataset using $\mathcal{L}_{LDM}$ along with one or more proposed loss functions, see Tab. 1. For example, SCoFT+M is Stable Diffusion fine-tuned on CCUB using the sum of $\mathcal{L}_{LDM}$ and $\mathcal{L}_M$ as a loss function. Each survey page has a total of four survey items (rows that participants respond to) to rank images on (a) Description and Image Alignment, (b) Cultural Representation, (c) Stereotypes, and (d) Offensiveness. Participants rank the set of randomly ordered images from best to worst image once for each item. An image labeled rank 1 signifies both best aligned and least offensive, while rank 4 is least aligned and most offensive.

We quantitatively estimate the subjective perceived performance of each method with Matrix Mean-Subsequence-Reduced (MMSR) [33] model in crowd-kit [57], an established algorithm [35] for noisy label aggregation, followed by a weighted majority vote to aggregate labels across workers, and then a simple majority vote aggregating labels into rankings, thus MMSR+Vote (see Supp Sec. 11.2).

**Automatic Metrics.** In addition to the user survey, we use Kernel Inception Distance (KID) [3] and CLIP Score [20] to evaluate the quality of generated images. For automatic evaluation to ablate SCoFT, we use 10 test prompts for each culture, generating 20 images for each prompt.


Figure 4. Qualitative comparison of our SCoFT model ablated and compared to Stable Diffusion without fine-tuning.

| | | | | Automatic Metrics | | | User Survey Results - Average Rank | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model Name | $\mathcal{L}_M$ | $\mathcal{L}_P$ | $\mathcal{L}_C$ | KID-↓ CCUB | KID-↓ COCO | CLIP-↑ Score | Best Described | Most Culturally Representative | Least Stereotypical | Least Offensive |
| Stable Diffusion[46] | | | | 30.355 | **4.396** | **0.813** | 3.09 | 3.07 | 2.98 | 3.07 |
| SCoFT+M | ✓ | | | 22.643 | 4.711 | 0.802 | 2.57 | 2.56 | 2.66 | 2.59 |
| SCoFT+MP | ✓ | ✓ | | 21.360 | 4.936 | 0.800 | 2.33 | 2.35 | 2.34 | 2.30 |
| SCoFT+MPC | ✓ | ✓ | ✓ | **19.621** | 4.819 | 0.799 | **1.83** | **1.84** | **1.91** | **1.78** |

Table 1. We compare our SCoFT ablations to Stable Diffusion using automatic metrics (Sec. 6) and a user survey (Sec. 5). Values in the user survey results report average ranking of images across all five cultures where lower rankings indicate better results (KID is $\times 10^3$)

## 6. Results

**Qualitative Comparison.** We qualitatively compare our SCoFT model versus the original Stable Diffusion in Fig. 1. SCoFT guides Stable Diffusion away from generating stereotypes and misrepresentations of culture. For example, many of the Stable Diffusion results for a "Photo of a traditional building, in ..." depict disheveled structures, which promote a harmful stereotype that some cultures are poor or simple, whereas SCoFT promotes more accurate and less stereotypical buildings for each nation. To investigate the effects of each loss function within SCoFT we also qualitatively compare each ablation in Fig. 4. We tend to see the SCoFT models modernize generated images, which decreases harmful stereotypes.

**User Survey Results.** 51 survey participants from five countries ranked images across four ablations by responding to each of the four survey items in Sec. 5 on freshly generated images. We had 13 Korean, 11 Chinese, 10 Indian, 9 Nigerian, and 7 Mexican participants. The average participant rankings are in Table 1.

We also ran the MMSR [33] noisy data labeling algorithm across all responses (see Sec. 5, Supp. Sec. 11.2), finding a participant consensus ranking of: (1) SCoFT+MPC (2) SCoFT+MP, (3) SCoFT+M, and finally (4) Generic Stable Diffusion. MMSR found that participants reached an identical consensus when separately ranking each ablation with respect each of the four survey items. MMSR also found a participant consensus in which every country individually agreed with the ranking above, with the exception of India, which swapped (1) SCoFT+MP and (2) SCoFT+MPC.

We convert the rankings into binary comparisons by isolating two ablations and comparing their rankings. This way, we can compare the effects of each of the loss functions of SCoFT. SCoFT+M was ranked less offensive than Stable Diffusion 63% of the time, SCoFT+MP was less of-

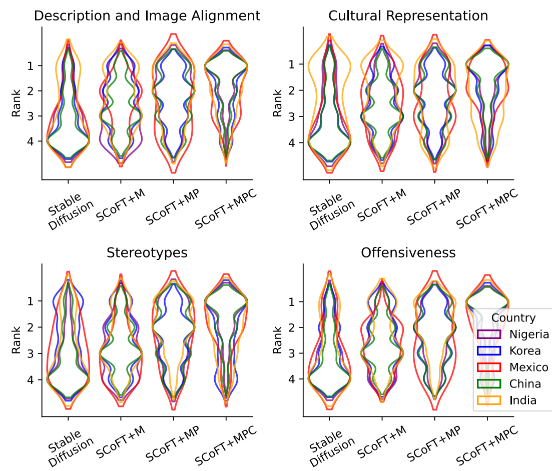


Figure 5. Violin plot of participant rankings across the survey items and countries. A wider strip means more answers with that value. Each new loss in our ablation study improved the rankings, and SCoFT+MPC is best. (Rank 1 is the best; 4, the worst)

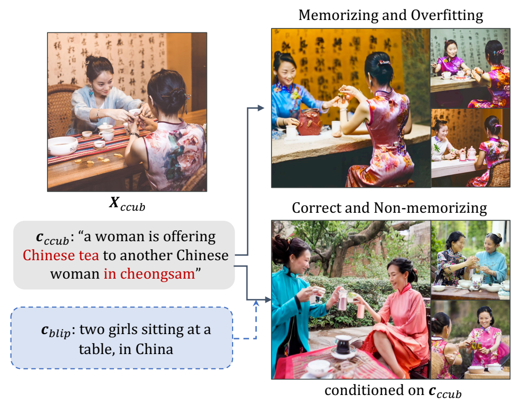


Figure 6. Top-right: Fine-tuning Stable Diffusion on CCUB data using only a conventional loss ($\mathcal{L}_{LDM}$, Sec. 4.1) leads to overfitting on CCUB captions. Bottom-right: Adding memorization loss ($\mathcal{L}_M$, Sec. 4.2) prevents overfitting with small datasets by ensuring images generated by general captions ($\mathbf{c}_{blip}$) are similar to those generated using CCUB's cultural captions.

fensive than SCoFT+M 56% of the time, and SCoFT+MPC was less offensive than SCoFT+MP 62% of the time. We see that each loss function contributed significantly to decreasing the offensiveness of generated results, and this trend continued for the other three survey items. We note that the initial addition of fine-tuning and the contrastive loss produced more dramatic improvements in SCoFT compared to adding perceptual loss.

We compare whole distributions of the rankings in Fig. 5. Across survey items, we see very similar distributions. For example, Stable Diffusion images were very commonly ranked fourth for both Stereotypes and Cultural Representation. Participants in the Chinese and Korean surveys ranked images with less variance than participants in the Indian and Mexican surveys. This is potentially due to a difference in the number of participants for each survey.

**Automatic Metric Results.** The automated evaluation results in Table 1 show that our proposed approach achieves the highest KID score on the CCUB test set, indicating that the fine-tuned model is able to generate images with a similar quality to the culturally curated data. In contrast, the original Stable Diffusion model scores highest in the CLIP Score and KID score on the MS COCO dataset. This result is not surprising as the CLIP model is itself known to be biased [4] in ways shared with Stable Diffusion [32], where the number of measurable outputs people perceive as hatred scales with the training set [5]. CLIP biases have also been quantitatively shown to be passed on to downstream applications [22]. Human evaluators favor our SCoFT+MPC method over Stable Diffusion, suggesting CLIP-Score's inadequacy in assessing cultural competence.

**Perceptual Backbones and Gradient Recording.** SCoFT uses a perceptual backbone to compare images in feature spaces rather than pixel space to avoid overfitting to training images. We test several backbones to extract image features, including the output of CLIP [42] convolutional layers [58], and the last layer of DreamSim [14], DINOv2 [40], BLIP2 [29]. For each comparison, we generate 200 images and calculate the KID with CCUB test set with the Korean, Chinese, and Indian models and see that the CLIP convolutional layers and DreamSim output provide the best generalization. For all other experiments, we use the CLIP convolutional layers as the SCoFT backbone.

We also compare the effect of recording the gradient during the first, last, and random iterations of denoising during fine-tuning, as reported in Fig. 7. We see that the best generalization comes when recording the gradient during the first iteration of denoising. This is in contrast to concurrent work [9] which recorded the last gradient, indicating that culture is a high-level concept where important information is created early in the diffusion process, as opposed to aesthetics which are low-level and correspond more strongly to later stages of diffusion.

**Effectiveness of the Memorization Loss.** We fine-tune Stable Diffusion using the original $\mathcal{L}_{LDM}$ with and without Memorization Loss, $\mathcal{L}_M$, on the CCUB dataset. To evaluate the consequence of over-fitting and reproducing training images during few-shot fine-tuning, we randomly select 10 text-image pairs for each culture from the CCUB training set. For each training text prompt, we generate 20 images. We evaluate the process using three metrics: CLIP-Image (CLIP-I), DINO, and DreamSim. All metrics measure the average pairwise cosine similarity between the embeddings of generated and real training images that share the same text prompt. For both metrics, lower values indicate more effective prevention of overfitting and reproduction of training images. Fig. 6 depicts qualitative results of generating creative images. Quantitative results in Table 2 show that

| | Prompts from Training set | | | | Prompts from Test set | |
|---|---|---|---|---|---|---|
| Method | CLIP-I ↓ | DINO ↓ | DreamSim ↓ | DIV train ↑ | DIV test ↑ | CLIPScore ↑ |
| Finetune | 0.912 | 0.836 | 0.591 | 0.302 | 0.317 | **0.824** |
| Finetune w/ $\mathcal{L}_M$ | **0.897** | **0.808** | **0.550** | **0.356** | **0.379** | 0.814 |

Table 2. Images generated from Stable Diffusion fine-tuned with and without Memorization Loss, $\mathcal{L}_M$, are compared in the feature space from various feature extractors. We see that $\mathcal{L}_M$ encourages the model to produce images with more diversity (larger feature difference).

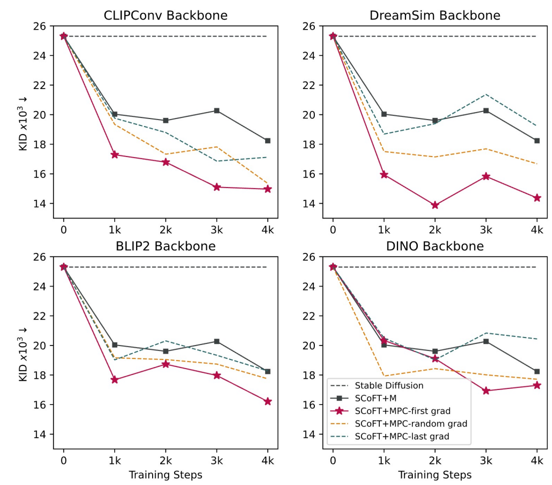


Figure 7. Comparison of perceptual models used for feature extraction in SCoFT. KID between held out CCUB images and generated images is plotted versus training iterations representing generalization to a validation set, score averaged across three cultures.

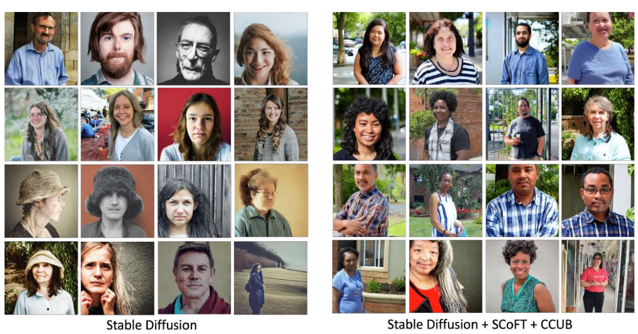


Figure 8. Generated images for text prompt: "a photo of a person." The proposed approach is able to generate more diverse images given a generic prompt without a specific cultural context.

the memorization loss effectively reduces overfitting.

To quantify output diversity, we randomly select 10 training text prompts and 10 CCUB testing text prompts. For each text prompt, we generate 20 images. We introduce the diversity metric (DIV), which calculates the average pair-wise DreamSim cosine distance between generated images with the same text prompt. Higher values indicate enhanced diversity in the generated outputs, reflecting a more varied and expressive synthesis of images. We also report a comparable CLIP Score on the generated image using CCUB testing text prompts with baseline fine-tuning.

**Limitations.** To tackle the bias in the data, we aim for two goals: 1) to generate accurate images given a specific cultural context and 2) to generate diverse images given a generic text prompt without any specific cultural context. Our current approach is focused on achieving the first goal. Our current model can generate promisingly diverse images for some generic prompts as shown in Figure 8 when compared to the baseline model that generates biased images.

Our CCUB dataset was collected by experienced residents; however more vigorous verification will be needed to improve the quality of the dataset. (See Supp. Sec. 13)

# 7. Conclusion

The biases of generative AI have already led to substantial and very public impacts [19, 34], so it is essential that we ensure models generate images that more accurately represent the diversity of the world. We propose Self-Contrastive Fine-Tuning (SCoFT), which is specifically designed to fine-tune the model for high-level concepts using a small dataset, for instance, to pay attention to subtle cultural elements. SCoFT has potential to generalize to applications in other domains, such as, reducing the risk of copyright infringement, better respecting cultural and community-defined boundaries, and addressing offensiveness across a broader range of identity characteristics, amongst other criteria. For example, supplement Fig. 19 shows initial results improving images for individuals with disabilities who use mobility aids. We have also confirmed positive associations with some automated metrics while demonstrating others are not a good fit for this task.

Our extensive user survey and metric evaluation quantitatively demonstrate improvements in subjective metrics with respect to image and description alignment, more culturally representative image contents, as well as reductions in stereotyping and offensiveness.

**Acknowledgment.** We thank Youeun Shin, Lia Coleman, and all contributors for their contributions to our dataset. This work was funded in part by NSF IIS-2112633, the Technology Innovation Program (20018295, Meta-human: a virtual cooperation platform for specialized industrial services) of the Ministry of Trade, Industry & Energy (MOTIE, Korea), the Ministry of Science and ICT, Korea, under the Information Technology Research Center support program (IITP-2024-2020-0-01789), and the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2024-RS-2023-00254592) supervised by the Institute for Information & Communications Technology Planning & Evaluation, and the NSF under Grant # 2030859 to the Computing Research Association CIFellows Project subaward #2021CIF-CarnegieMellon-72.

## References


[1] Jason Armitage, Endri Kacupaj, Golsa Tahmasebzadeh, Swati, Maria Maleshkova, Ralph Ewerth, and Jens Lehmann. Mlm: a benchmark dataset for multitask learning with multiple languages and modalities. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2967–2974, 2020. 2

[2] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Transactions on Graphics (TOG)*, 42 (4):1–11, 2023. 3

[3] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 5

[4] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021. 2, 7

[5] Abeba Birhane, Vinay Prabhu, Sang Han, and Vishnu Naresh Boddeti. On hate scaling laws for data-swamps, 2023. 2, 7

[6] Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3.1 – a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023. 5

[7] Mari Castañeda. The power of (mis) representation: Why racial and ethnic stereotypes in the media matter. *Challenging inequalities: Readings in race, ethnicity, and immigration*, 2018. 2

[8] Michelle Caswell, Alda Allina Migoni, Noah Geraci, and Marika Cifor. 'to be able to imagine otherwise': community archives and the importance of representation. *Archives and Records*, 38(1):5–26, 2017. 1

[9] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023. 4, 7

[10] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431*, 2021. 3

[11] Travis L Dixon and Daniel Linz. Overrepresentation and underrepresentation of african americans and latinos as lawbreakers on television news. *Journal of communication*, 50 (2):131–154, 2000. 2

[12] Rawan Elbaba. Why on-screen representation matters, according to these teens. *PBS NewsHour*, 14, 2019. 1

[13] Zhida Feng, Zhenyu Zhang, Xintong Yu, Yewei Fang, Lanxin Li, Xuyi Chen, Yuxiang Lu, Jiaxiang Liu, Weichong Yin, Shikun Feng, et al. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. *arXiv preprint arXiv:2210.15257*, 2022. 3

[14] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023. 3, 7

[15] Rohit Gandikota, Joanna Materzynska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models. *arXiv preprint arXiv:2311.12092*, 2023. 2

[16] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5111–5120, 2024. 2

[17] Noa Garcia, Yusuke Hirota, Yankun Wu, and Yuta Nakashima. Uncurated image-text datasets: Shedding light on demographic bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6957–6966, 2023. 3

[18] Ben Halpern. The dynamic elements of culture. *Ethics*, 65 (4):235–249, 1955. 3

[19] Melissa Heikkilä. The viral ai avatar app lensa undressed me without my consent, 2022. Accessed on December 16, 2023. 8

[20] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 5

[21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 3, 4

[22] Andrew Hundt, William Agnew, Vicky Zeng, Severin Kacianka, and Matthew Gombolay. Robots enact malignant stereotypes. *FAccT*, page 743–756, 2022. 2, 7

[23] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, page 560–575, New York, NY, USA, 2021. Association for Computing Machinery. 2

[24] Mary Ritchie Key and Bernard Comrie, editors. *IDS*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2021. 3

[25] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022. 3

[26] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *ICCV*, 2023. 4, 1

[27] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1931–1941, 2023. 3

[28] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 4, 1

[29] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with

frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 3, 7, 1

[30] Fangyu Liu, Emanuele Bugliarello, E. Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. Visually grounded reasoning across languages and cultures. *ArXiv*, abs/2109.13238, 2021. 3

[31] Haoming Lu, Hazarapet Tunanyan, Kai Wang, Shant Navasardyan, Zhangyang Wang, and Humphrey Shi. Specialist diffusion: Plug-and-play sample-efficient fine-tuning of text-to-image diffusion models to learn any unseen style. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14267–14276, 2023. 3

[32] Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Evaluating societal representations in diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 2, 7

[33] Qianqian Ma and Alex Olshevsky. Adversarial crowdsourcing through robust rank-one matrix completion. In *Advances in Neural Information Processing Systems*, pages 21841–21852. Curran Associates, Inc., 2020. 5, 6

[34] Emanuel Maiberg. Inside the ai porn marketplace where everything and everyone is for sale. 2, 8

[35] Mohammad S. Majdi and Jeffrey J. Rodriguez. Crowdcertain: Label aggregation in crowdsourced and ensemble learning classification, 2023. 5

[36] Abhishek Mandal, Susan Leavy, and Suzanne Little. Dataset diversity: measuring and mitigating geographical bias in image search and retrieval. In *Proceedings of the 1st International Workshop on Trustworthy AI for Multimedia Computing*, pages 19–25, 2021. 3

[37] Dana E Mastro and Bradley S Greenberg. The portrayal of racial minorities on prime time television. *Journal of Broadcasting & Electronic Media*, 44(4):690–703, 2000. 2

[38] Marc Miquel-Ribé and David Laniado. Wikipedia cultural diversity dataset: A complete cartography for 300 language editions. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 620–629, 2019. 3

[39] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. Bias in data-driven artificial intelligence systems—an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3): e1356, 2020. 2

[40] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 3, 7

[41] Justin Pinkney. Text to pokemon generator, 2022. 3

[42] 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. 3, 7

[43] Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. Ai and the everything in the whole wide world benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. Curran, 2021. 3

[44] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022. 5

[45] William Gaviria Rojas, Sudnya Diamos, Keertan Ranjan Kini, David Kanter, and Vijay Janapa Reddi. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. 2022. 2

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2, 6

[47] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022. 4

[48] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 3

[49] Peter Schaldenbrand, James McCann, and Jean Oh. Frida: A collaborative robot painter with a differentiable, real2sim2real planning environment. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11712–11718. IEEE, 2023. 3

[50] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. Do datasets have politics? disciplinary values in computer vision dataset development. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), 2021. 2

[51] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 2

[52] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 3

[53] Adrienne Shaw. *Identity, identification, and media representation in video game play: An audience reception study*. PhD thesis, University of Pennsylvania, 2010. 1

[54] Makoto Shing and Kei Sawada. Japanese stable diffusion. 2022. 3

[55] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image

text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449, 2021. 2

[56] Harini Suresh and John Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, New York, NY, USA, 2021. Association for Computing Machinery. 2

[57] Dmitry Ustalov, Nikita Pavlichenko, and Boris Tseitlin. Learning from crowds with crowd-kit, 2023. 5

[58] Yael Vinker, Ehsan Pajouheshgar, Jessica Y Bo, Roman Christian Bachmann, Amit Haim Bermano, Daniel Cohen-Or, Amir Zamir, and Ariel Shamir. Clipasso: Semantically-aware object sketching. *ACM Transactions on Graphics (TOG)*, 41(4):1–11, 2022. 7

[59] Bram Wallace, Akash Gokul, Stefano Ermon, and Nikhil Naik. End-to-end diffusion latent optimization improves classifier guidance. *arXiv preprint arXiv:2303.13703*, 2023. 3

[60] Angelina Wang, Solon Barocas, Kristen Laird, and Hanna Wallach. Measuring representational harms in image captioning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, page 324–335, New York, NY, USA, 2022. Association for Computing Machinery. 2

[61] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 5

[62] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 3