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

Supplementary Material

Overview. In Sec. 8, we present the pseudocode for SCoFT and report the training details for the experiments. Sec. 9 provides statistics about the CCUB dataset. Sec. 10 showcases the results of automatic metrics per culture, followed by the human evaluation survey question and additional analysis on human evaluation feedback in Sec. 11. Finally, Sec. 12 includes more examples for our models, and Sec. 13 discusses the ethics and limitations.

8. SCoFT Method

8.1. SCoFT Pseudocode

We show here the pseudocode of our SCoFT method.

Algorithm 1 SCoFT for Stable Diffusion

dataset: CCUB dataset $\mathcal{D}_{ccub} = \{(\mathbf{x}_{ccub}, \mathbf{c}_{ccub}, \mathbf{c}_{blip})\}$
inputs: Pre-trained Stable Diffusion 1.4 model ϵ_θ, LoRA weights $\hat{\theta}$
for $(\mathbf{x}^i_{ccub}, \mathbf{c}^i_{ccub}, \mathbf{c}^i_{blip}) \in \mathcal{D}_{ccub}$ **do**
  $\mathbf{z}_0 = ENCODE(\mathbf{x}^i_{ccub})$
  $\epsilon_t \sim \mathcal{N}(0, \mathbf{I})$
  $\mathbf{z}_t = \sqrt{\alpha_t}\mathbf{z}_0 + \sqrt{1-\alpha_t}\epsilon_t$
  $\mathcal{L}_{LDM} = MSE(\epsilon_t, \epsilon_{\hat{\theta}}(\mathbf{z}_0, \mathbf{c}_{ccub}, t))$ ▷ $\mathcal{L}_{LDM}$
  $\mathcal{L}_M = MSE(\epsilon_{\hat{\theta}}(\mathbf{z}_0, \mathbf{c}_{ccub}, t), \epsilon_{\hat{\theta}}(\mathbf{z}_0, \mathbf{c}_{blip}, t))$ ▷ $\mathcal{L}_M$
  if every 10 timesteps **then**
    $\mathbf{x}^+ = \mathbf{x}_{ccub}$
    $\mathbf{x}^- = \{\Theta(\mathcal{D}_{depth}(\mathbf{x}^+), \mathbf{c}_{blip})\}$
    $t_{record} = \begin{cases} t & \text{record first gradient} \\ t_{rand} & \text{record random gradient} \\ 1 & \text{record last gradient} \end{cases}$
    $\hat{\mathbf{z}}_t = \mathbf{z}_t$
    for $u = t, ..., 1$ **do**
      if $u \neq t_{record}$ **then**
        $\hat{\mathbf{z}}_u = \text{stop\_grad}(\hat{\mathbf{z}}_u)$
      end if
      $\hat{\mathbf{z}}_{u-1} = \epsilon_{\hat{\theta}}(\hat{\mathbf{z}}_u, \mathbf{c}_{ccub}, u)$ // de-noise
    end for
    $\hat{\mathbf{x}} = DECODE(\hat{\mathbf{z}}_0)$
    $\mathcal{L}_C(\hat{x}, x^+, x^-) = \mathbb{E}_{\hat{x},x^+,x^-}[\max(\mathcal{S}(\hat{x}, x^+; f_\theta)$
    $-\lambda\mathcal{S}(\hat{x}, x^-; f_\theta) + m, 0)]$ ▷ $\mathcal{L}_P + \mathcal{L}_C$
  end if
  $\mathcal{L} = \lambda_l\mathcal{L}_{LDM} + \lambda_m\mathcal{L}_M + \lambda_c\mathcal{L}_C$
  $g = \nabla\mathcal{L}(\theta^{LoRA})$
  $\theta^{LoRA} \leftarrow \theta^{LoRA} - \eta g$
end for

	food & drink	people & actions	clothing	architecture	city	dance music art	nature	utensil & tool	religion & festival	**total**
Korea	34	22	20	20	21	20	8	6	11	162
China	33	31	24	23	27	23	5	15	8	189
Nigeria	21	16	18	17	13	21	11	8	15	140
Mexico	22	19	14	18	15	10	7	10	19	134
India	19	26	24	21	16	18	9	7	8	148
United States	23	22	10	17	29	15	16	12	7	151
Total	152	136	110	116	121	107	56	58	68	**924**

Table 3. This table shows the scale of our CCUB dataset, detailing the number of hand-selected images and their corresponding captions across nine cultural categories for six different cultures.

8.2. Training Details

8.2.1 Stable Diffusion version

We utilize Stable Diffusion 1.4. While we conducted preliminary studies on versions 1.5/2.0/XL, only SDXL showed minor improvements. The similarities in training datasets among versions likely explain the minimal variance. Stable Diffusion 1.4 was specifically chosen due to its widespread adoption in concept editing and fine-tuning research [26].

8.2.2 Training specifics and hyperparameters

We maintain uniform settings for each model to ensure a fair comparison. Throughout training, we employ the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for 3000 iterations, utilizing a learning rate of $1e-4$. The batch size is set to 1, and LoRA is exclusively applied to UNet parameters with a rank of 64. We select CLIPConv using the fifth convolutional layer as the backbone and record the first gradient during backpropagation through sampling. For each positive example, we generate 5 negative examples, employing DreamSim to filter out false negatives which are similar to the positive samples. The training weights we set for different losses are $\lambda_l = 0.7$ and $\lambda_m = 0.3$. To manage time costs, we compute $\mathcal{L}_c$ every 10 iterations, denoising the latent $\mathbf{z}_t$ using 20 steps of DDIM sampling, with $\lambda_c = 0.1$. The training process runs for approximately 2.5 hours on a single NVIDIA V100 GPU for each CCUB cultural dataset. Note: We utilize both blip1 [28] and blip2 [29] for different parts of our system.

9. CCUB Dataset

While is generally infeasible to entirely eliminate bias from data, we designed the CCUB dataset collection protocol to reduce cultural biases. It aims to provide a more comprehensive representation of culture, a facet not adequately addressed by the LAION dataset. Our experimental results, validated by resident participants, demonstrate a significant reduction (75%-80% less) in perceived bias when using our dataset, especially in comparison to the

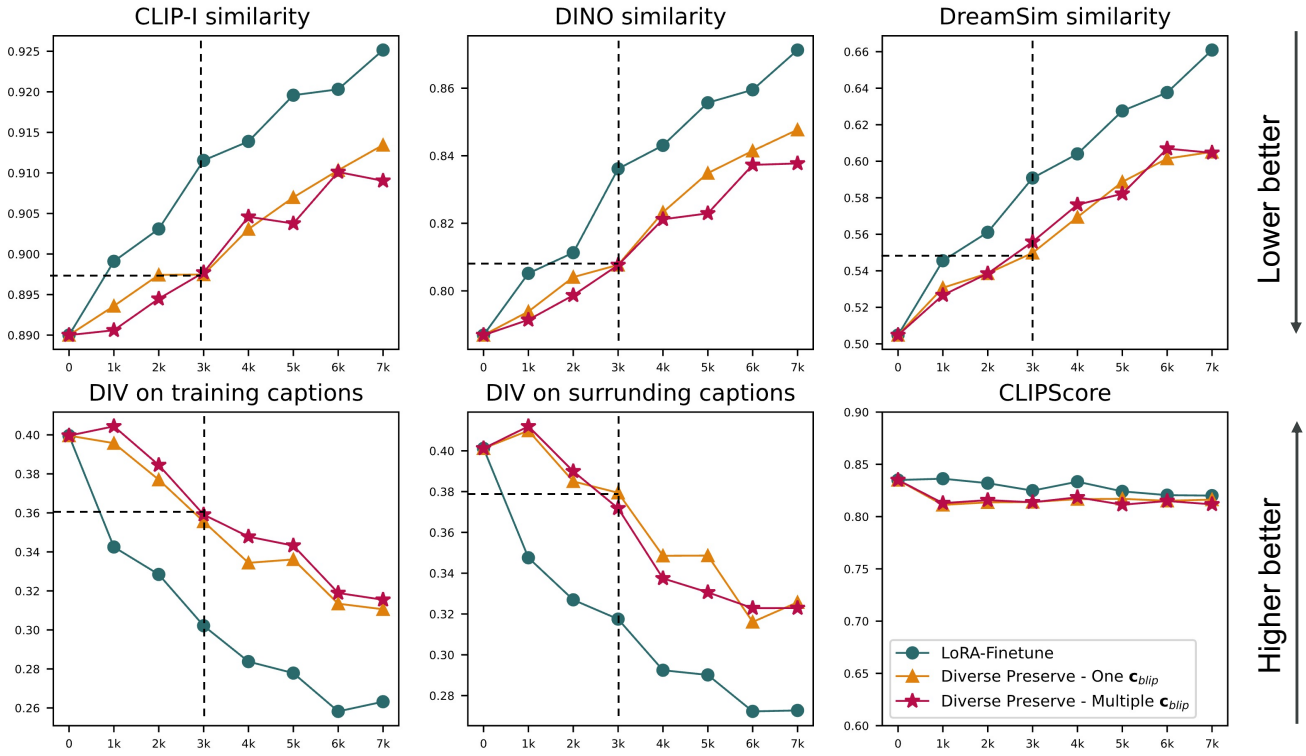


Figure 9. **Preventing Memorizing during training**. We compare training with ("LoRA-Finetune") and without $\mathcal{L}_M$ ("Diverse Preserve"), analyzing the effects of using one or multiple $\mathbf{c}_{blip}$. The scores are averaged across three CCUB culture datasets. Fine-tuning with $\mathcal{L}_M$ for 3000 steps achieves even better performance than without it for 1000 steps. Adding $\mathcal{L}_M$ effectively prevents overfitting, and we notice that using more than one $\mathbf{c}_{blip}$ has little effect.

Stable Diffusion baseline. This necessarily requires, and thus also demonstrates, that a reduction in bias is achieved by our dataset when compared to LAION.

Table 3 illustrates the scale of our CCUB dataset. The number of hand-selected images and their corresponding captions in nine cultural categories for six different cultures are listed. To ensure cultural authenticity and diversity, a minimum of 5 university-educated annotators per culture, with verified cultural knowledge and balanced gender representation, were recruited. Data were gathered across 9 key cultural categories. Voluntary survey annotators contributed 2-3 culturally rich image URLs per category, plus modern and traditional facets. They provided concise, culturally nuanced captions in English, averaging 10.16 words. Additional reviewers conducted a final validation. CCUB contains the image web links and the manual captions. We do not own the copyright of the images.

10. Automatic Metrics

Figure 9 demonstrates that $\mathcal{L}_M$ serves to prevent memorization and enhance diverse expression. The metrics are calculated during training and averaged across three cultures. Additionally, Table 4 offers supplementary details, abalating SCoFT for automatic metrics performance across each culture.

11. Human Evaluation

11.1. Survey Description

Our goal of improving the cultural perception of generated images is a subjective metric largely determined by members of a given identity group. Fundamentally, all generated images and depictions of cultural elements are subjective. We work towards a more quantitative understanding of subjectivity as per our IRB-approved study grounded in collective insights from residents of each culture, rather than individual opinions. To evaluate our performance on this criteria, 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 each of the four generators using a common random seed, for four total images. Each survey page has a total of four survey items (rows that participants respond to, see Table 5) to rank relative to (a) Description and Image Alignment, (b) Cultural Representation, (c) Stereotypes, and (d) Offensiveness.

Our survey items are conceptually intertwined to create a scale suitable for our cultural measurement goals. Related items compliment one another so no one typo or answer disrupts our overall findings. Our aim is to maximize the scale quality, while considering survey fatigue. Participants respond by numerically ranking

Model Name	KID-CCUB $\times 10^3 \downarrow$					KID-COCO $\times 10^3 \downarrow$					CLIP Score $\uparrow$				
	CN	KR	NG	IN	MX	CN	KR	NG	IN	MX	CN	KR	NG	IN	MX
Stable Diffusion	25.244	28.061	32.687	39.988	25.796	**4.396**	4.396	**4.396**	**4.396**	4.396	**0.827**	**0.844**	**0.799**	**0.773**	0.819
SCoFT+M	18.798	24.118	24.539	20.566	25.196	4.890	4.497	4.461	5.342	4.364	0.804	0.832	0.792	0.757	0.825
SCoFT+MP	15.895	23.007	25.638	19.799	**22.462**	5.090	4.403	5.135	5.112	4.941	0.798	0.836	0.792	0.753	0.8232
SCoFT+MPC	**14.667**	**17.413**	**23.261**	**18.342**	24.424	5.102	**4.222**	4.987	5.457	**4.325**	0.793	0.825	0.795	0.753	**0.829**

Table 4. We compare our SCoFT ablations to Stable Diffusion using automatic metrics and present the results for five cultures. The KID-CCUB is calculated on the internal CCUB test dataset, containing 150 cultural data for each culture. KID-COCO is calculated on randomly selected 500 text-image pairs from the MS-COCO dataset. CLIPScore is calculated using the CCUB test dataset text prompts.

Item Title	**Item Text**
Description and Image Alignment	Rank the accuracy of the match between the image and its description. Number each image from 1 for the image best described by the description to 4 for the worst match between image and description. **(1=best described, 4=worst described)**
Cultural Representation	Rank the images from 1 for the best representation of $country\_adj$ culture to 4 for the worst cultural representation. Please ignore image artifacts (such as distorted faces, hands, or glitches) when considering Cultural Representation. **(1=most representative, 4=least representative)**
Stereotypes	Rank the images from 1 for being the least stereotypical depiction of $country\_adj$ culture to 4 for the most based on your personal knowledge of $country\_adj$ stereotypes. **(1=least stereotypical, 4=most stereotypical)**
Offensiveness	Rank the images by their offensiveness to you personally, numbering from 1 as the least offensive to 4 for the most offensive. Please ignore image artifacts (such as distorted faces, hands, or glitches) when considering Offensiveness. **(1=least offensive, 4=most offensive)**

Table 5. We shows the survey items for our human evaluation on image generated using different models.

the set of randomly ordered images from best image to worst image once for each item. An image labeled rank 1 would signify both best aligned and least offensive when each case is ranked, while rank four would be least well aligned and most offensive. A sample of a single survey page can be viewed in Figure 25.

11.2. Analysis & Evaluation of Model Performance

We quantitatively estimate the subjective perceived performance of each model using the crowd-kit implementation of the Matrix Mean-Subsequence-Reduced (MMSR) model, an established algorithm 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. MMSR models the varying levels of participant expertise as a vector, which we frame as representing consensus alignment to account for the combination of knowledge and subjective perception we are measuring, and the equivalent term chosen by is "skills". We abstract rankings in survey responses into unique binary pairwise comparison labels asking if the left image is perceived as better than the right image with respect to the given survey item. An example of a binary comparison is when a respondent has indicated that it is true that an image given rank 1 is less offensive than the image given rank 2. MMSR is provided with the participant, survey item, and abstracted response labels, then models the noisy label prediction problem as:

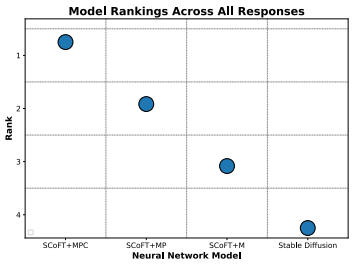
$$\mathbb{E}\left[\frac{M}{M-1}\widetilde{C} - \frac{1}{M-1}\mathbf{1}\mathbf{1}^T\right] = \boldsymbol{s}\boldsymbol{s}^T, \tag{5}$$

where $\widetilde{C}$ is the participant covariance matrix (Figure 12), M is the number of labels (true, false), and $boldsymbol1$ is a matrix filled with the width and height of $\widetilde{C}$. We run 10k iterations of robust rank-one matrix completion with a stopping tolerance of 1e-10 to compute the initial label estimates. We aggregate the binary labels with a weighted majority vote, and then estimate aggregate rankings from labels with a simple majority vote.

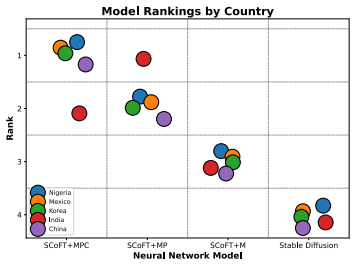
One limitation of MMSR evaluation is that it prioritizes a single consensus response. This means a comparatively small quantity of insightful but marginalized perspectives might be undervalued in a manner undifferentiated from a small quantity of random responses.

We run MMSR under three configurations:

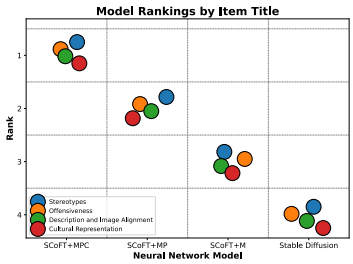
(1) Figure 10a Overall Best Method: where all rankings for all items and all countries are supplied together to estimate the



(a) Overall, participants find our SCoFT+MPC method outperforms all comparable methods when ranked with MMSR+Vote.

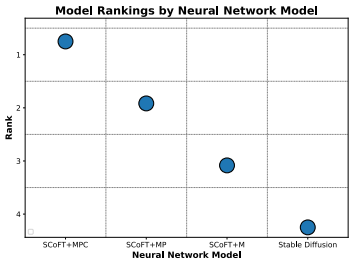


(b) Participants' overall preference for our SCoFT+MPC method is generally consistent across national affiliations when ranked with MMSR+Vote.

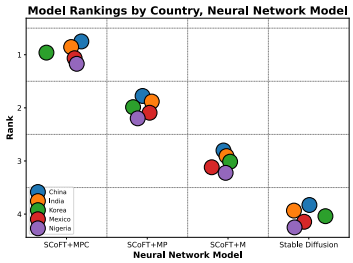


(c) Participants prefer our SCoFT+MPC method with respect to Description and Image Alignment, Cultural Representation, Stereotyping, and Offensiveness when ranked with MMSR+Vote.

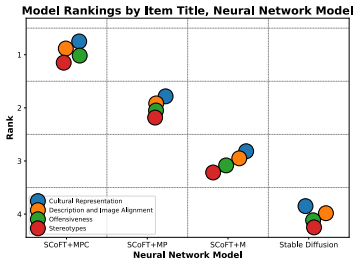
Figure 10. Methods Ranked According to Culturally Experienced Human Participants (MMSR+Vote alg, see Sec. 11.2) Higher ranks (numerically smaller values) are better.



(a) Overall, participants find our SCoFT+MPC method outperforms all comparable methods when ranked with the Noisy Bradley Terry Algorithm.



(b) Participants' overall preference for our SCoFT+MPC method is consistent across national affiliations when ranked with the Noisy Bradley Terry Algorithm.



(c) Participants prefer our SCoFT+MPC method with respect to Description and Image Alignment, Cultural Representation, Stereotyping, and Offensiveness when ranked with the Noisy Bradley Terry Algorithm.

Figure 11. Methods Ranked According to Culturally Experienced Human Participants (Noisy Bradley Terry alg, see Sec. 11.2) Higher ranks (numerically smaller values) are better.

best method across all responses. This model found a participant consensus indicating our SCoFT+MPC method performs best, followed by SCoFT+MP, SCoFT+M, and finally Generic Stable Diffusion as the least preferred.

(2) Figure 10b Overall best method by Country: This model found a participant consensus in complete agreement with configuration 1, with the exception of India where the consensus agreement rated the SCoFT+MP method best, followed by SCoft+MPC, SCoFT+M, and Generic Stable Diffusion.

(3) Figure 10c Overall best method by Survey Item: This model found overall participant ratings of models across each of the survey topics (Prompt to Image Alignment, Cultural Representation, Stereotyping, and Offensiveness) agree with configuration 1.

Figure 12 shows the MMSR Covariance matrix heatmap representing the strength of agreement between different participants. Each row and column is a separate participant. The deep blue areas represent questions with zero overlap, as participants from each country were asked independent groups of questions. The lighter blue through dark red squares are the similarity of answers between ay two participants. The squares from top left to bottom right represent China, India, Mexico, Korea, and Nigeria, respectively. The covariance matrix for China is notably a darker red compared to all other countries, indicating a combination of a stronger agreement in combination with the greater number of questions answered per participant with experiences in China, on average.

Figure 13 displays the counts of survey item responses (Table 5) for each ablation. Our contrastive approach SCoFT+MPC is consistently selected as the top-ranked choice across all survey evaluation items.

Noisy Bradley Terry. We quantitatively estimate the subjective perceived performance of each model using the crowd-kit implementation of the Noisy Bradley Terry (NBT) model, an algorithm that performs noisy binary label aggregation for the purpose of further substantiating our existing results. As Figure 11 shows, the NBT composite score of participant responses consistently ranks SCoFT+MPC method as the best, followed by SCoFT+MP, SCoFT+M, and finally Generic Stable Diffusion as the least preferred in every case, including the Overall rankings in Figure 11a, the rankings by Country in Figure 11b, and the Rankings by Survey Item in Figure 11c.

Taken together, the reliable consistency of model rankings across multiple different evaluation methods, maximum participant count (Figure 13), and simple averaging, represent very

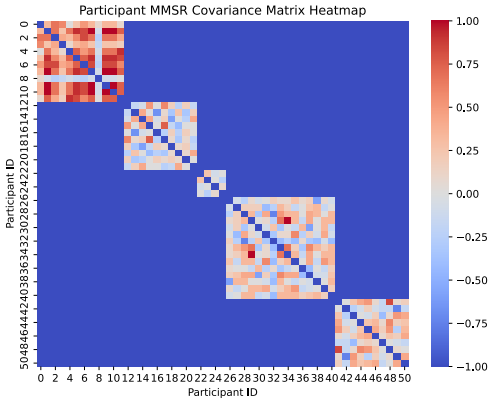


Figure 12. MMSR Covariance Matrix of Participant Response Agreement when evaluating combinations of Neural Network Models (e.g. contrastive), Survey Items (e.g. Offensiveness), and Country (e.g. Nigeria). Each row and column denotes a different person, where the more red the square, the more the participants agree with each other. -1 is maximum disagreement (or no mutual responses), and 1 means maximum agreement. Each of the big grouping rectangles represents data from a different country, ordered from top left to bottom right as China, India, Mexico, Korea, Nigeria.

strong quantitative evidence of the improvements constituted by our SCoFT+MPC method with respect to Description and Image Alignment, Cultural Representation, Stereotyping, and Offensiveness when compared to Stable Diffusion.

12. Additional Qualitative Details and Samples

12.1. Memorization Loss

We present more examples comparing fine-tuning Stable Diffusion with CCUB using $\mathcal{L}_{LDM}$ with and without memorization loss ($\mathcal{L}_M$) in Figures 14, 15, and 16. The addition of $\mathcal{L}_M$ prevents the model from generating images similar to the training images given a training text prompt. This is an important property when fine-tuning on a small dataset, e.g., CCUB.

12.2. Ablation on Self-Contrastive Perceptual Loss

In Figures 17 and 18, we present additional ablations on SCoFT, showcasing the impact of different losses on the CCUB datasets for Chinese and Indian cultures. SCoFT improves generated images compared to conventional fine-tuning by reducing stereotypes such as old, poor looking regions and adhering to culturally important aspects such as art tools, clothing, and architectual styles.

In Figure 19, we validate the efficacy of each loss in our SCoFT framework using an internal prosthetics dataset. Our findings indicate that generic Stable Diffusion tends to introduce inherent biases in representations of the prosthetic using community and often generates inaccurate images. In contrast, our SCoFT+MPC approach consistently produces accurate representations. Moreover, we demonstrate the versatility of SCoFT by applying it to other fine-tuning domains, such as the prosthetics dataset, where it proves effective in generating more accurate images. We hope to perform future work demonstrating SCoFT's abilities to improve generated images for many communities harmed by bias and inaccurate representation.

12.3. Figure 1 Qualitative Details

Figure 1 insights were sourced from experienced residents. For the 'Nigerian Culture' column 2, row 1, Stable Diffusion depicts a dirty, dilapidated structure, while in row 4 SCoFT, our method, correctly generates a Nigerian town hall with a veranda (vernacular, Yoruba architecture). For the 'Korean Culture' column 3 row 3, the prompt was "Two people wearing traditional clothing, in Korea" and Stable Diffusion incorrectly rendered as a Japanese 'Kimono', while row 4 SCoFT, our method, correctly renders a Korean 'Hanbok'. Every Stable Diffusion example highlighted as stereotypical or a misrepresentation also has a run-down appearance and/or a lack of greenery that is addressed by our SCoFT method.

12.4. More Qualitative Examples

In Figures 20 to 24, we present additional qualitative examples illustrating how our SCoFT model outperforms the original Stable Diffusion in cultural understanding and the ability to generate less offensive images for various cultural concepts. Results are showcased across Nigeria, Korea, India, Mexico, and China. Furthermore, our models exhibit good performance for text prompts beyond our nine cultural categories, such as "photo of a bedroom", "students are studying in the classroom".

13. Additional Applications and Limitations

This work has the potential to shift the way that image generators operate at achievable costs to ensure that several categories of harm from 'AI' generated models are mitigated, while the generated images become much more realistic and representative of the AI-generated images that populations want around the world. Our proposed methods have 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 and other criteria.

Additionally, this work carries several risks, for example, the algorithm can easily be inverted to generate the most problematic images possible. While our approach works toward collecting and training datasets in a respectful manner, we do not address the nonconsensual use of images for training the baseline Stable Diffusion models we use as a starting point. Participants also were not asked to provide the copyright details of the images they collected from the internet for inclusion in CCUB.

Finally, we leave integration of SCoFT with models that are already generally effective at representing a given culture to future work, as the very premise remains an open research question, and would be subject to different interpretations amongst different subcultures of any large cultural population.

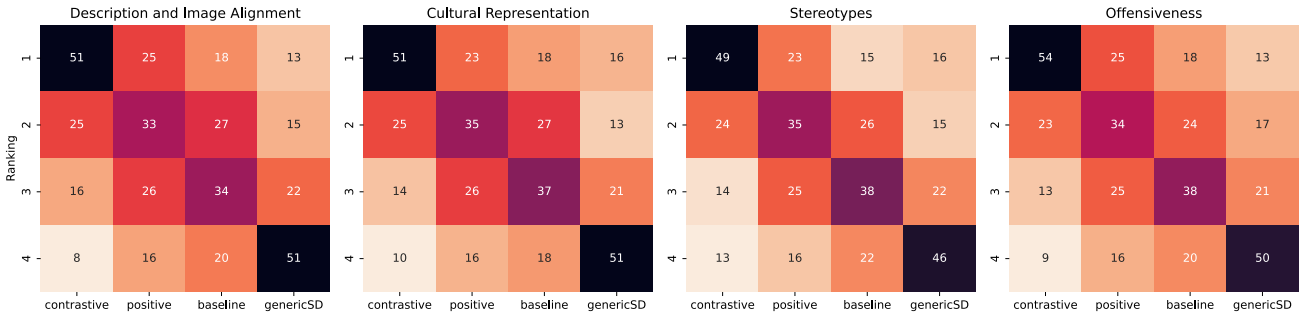


Figure 13. Counts of Participant-selected Model ranking for each survey item across all responses, more participants choosing a better rank value (lower number) is better. Our contrastive approach is selected for rank 1 most frequently across all survey evaluation items.

Figure 14. Additional qualitative example comparing training of 6000 steps with and without $\mathcal{L}_M$. The generated images are conditioned on the training text prompt: "women dressed in Korean traditional Hanbok are walking down a street".

Figure 15. Additional qualitative example comparing training of 6000 steps with and without $\mathcal{L}_M$. The generated images are conditioned on the training text prompt: "two women in Chinese cheongsam standing in front of a table with one holding a Chinese hand fan".

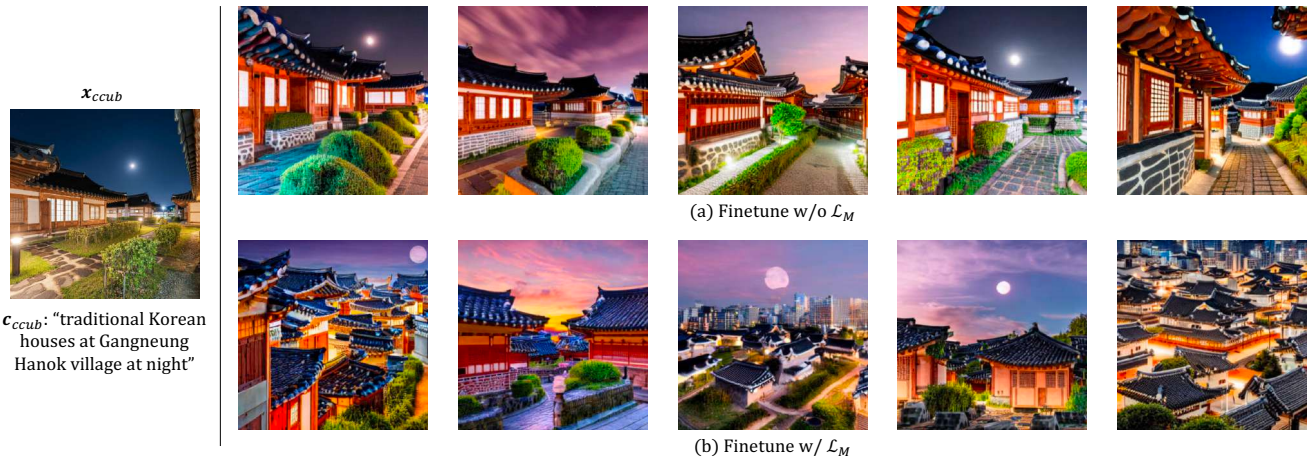


Figure 16. Additional qualitative example comparing training of 6000 steps with and without $\mathcal{L}_M$. The generated images are conditioned on the training text prompt: "traditional Korean houses at Gangneung Hanok village at night".

Figure 17. **Ablation on SCoFT for Chinese culture.** We present additional qualitative examples for fine-tuning on the CCUB Chinese dataset using different losses. Generic Stable Diffusion often results in stereotypes and misrepresentations of Chinese culture. In contrast, our SCoFT+MPC approach achieves superior results, generating accurate and less offensive images.

Figure 18. **Ablation on SCoFT for Indian culture.** We present additional qualitative examples for fine-tuning on the CCUB India dataset using different losses. Generic Stable Diffusion often results in stereotypes and misrepresentations of Indian culture. In contrast, our SCoFT+MPC approach achieves superior results, generating accurate and less offensive images.

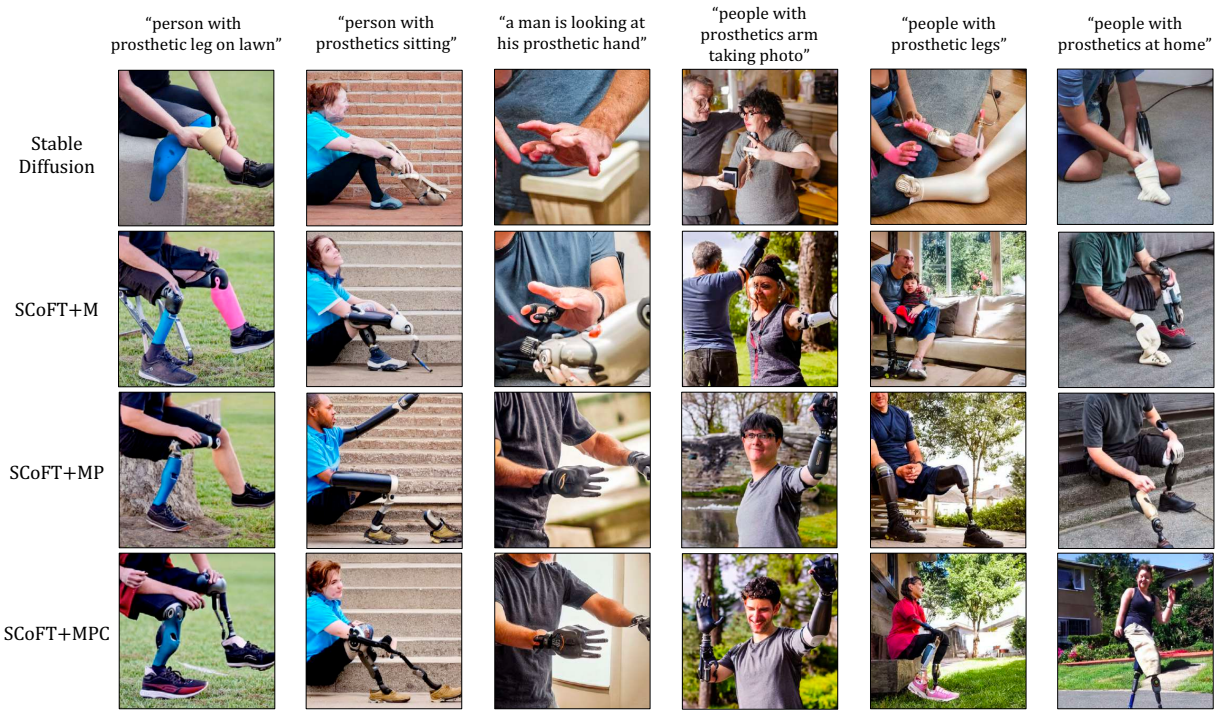


Figure 19. **Ablation on SCoFT for prosthetic dataset.** We provide additional qualitative examples for fine-tuning on the prosthetic dataset using different losses. Generic Stable Diffusion struggles to generate accurate representations for people with prosthetics. In contrast, our SCoFT+MPC approach achieves more accurate representations.

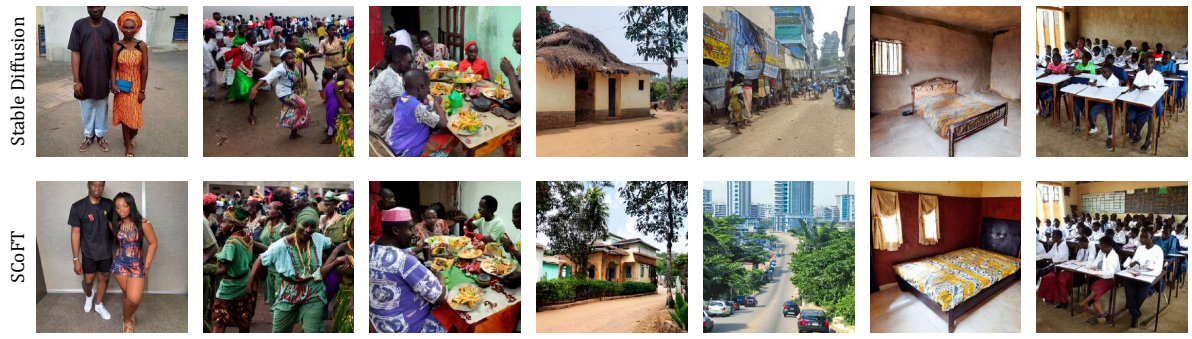


Figure 20. **Additional qualitative examples comparing Stable Diffusion with our SCoFT model for Nigerian culture.** The text prompts are "Nigerian people in casual clothing nowadays", "dancers are performing for a crowd, in Nigeria", "family is eating together, in Nigeria", "photo of a house, in Nigeria", "photo of a street, in Nigeria", "photo of a bedroom, in Nigeria", "student studying in the classroom, in Nigeria".

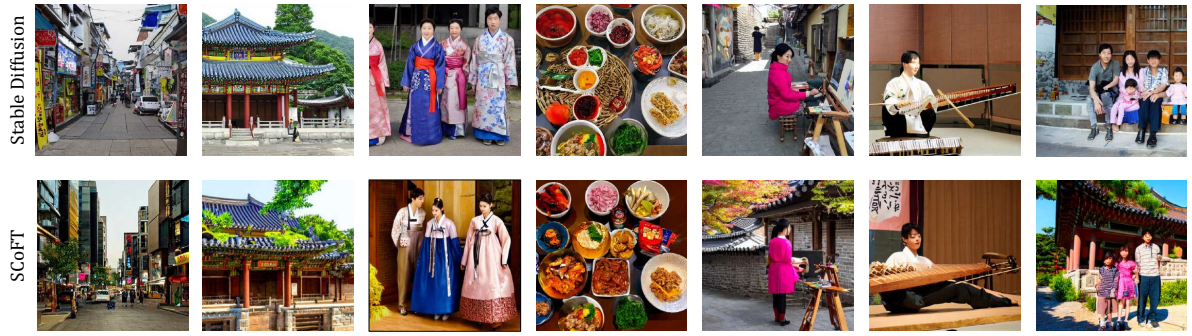


Figure 21. **Additional qualitative examples comparing Stable Diffusion with our SCoFT model for Korean culture.** The text prompts are "photo of a street, in Korea", "photo of a traditional building, in Korea", "people wearing traditional clothing, in Korea", "a table of food in Korea", "a woman is painting in a traditional style, in Korea", "musician performing Korean traditional instrument", "photo of a family, in Korea".

Figure 22. **Additional qualitative examples comparing Stable Diffusion with our SCoFT model for Indian culture.** The text prompts are "photo of children in India", "photo of a house, in India", "people wearing traditional clothing, in India", "family is eating together, in India", "photo of a street, in India", "people walking on the street, in India", "people inside their house, in India".

Figure 23. **Additional qualitative examples comparing Stable Diffusion with our SCoFT model for Mexican culture.** The text prompts are "photo of a building, in Mexico", "people wearing traditional clothing, in Mexico", "photo of a family, in Mexico", "photo of a school, in Mexico", "university student studying, in Mexico", "people performing traditional music instrument, in Mexico", "family is eating together, in Mexico".

Figure 24. **Additional qualitative examples comparing Stable Diffusion with our SCoFT model for Chinese culture.** The text prompts are "people are performing traditional instrument, in China", "photo of a school, in China", "photo of a street, in China", "family is eating together, in China", "two girls wearing Chinese traditional Han dress", "a man and a woman, in China", "woman is painting in a traditional style, in China".

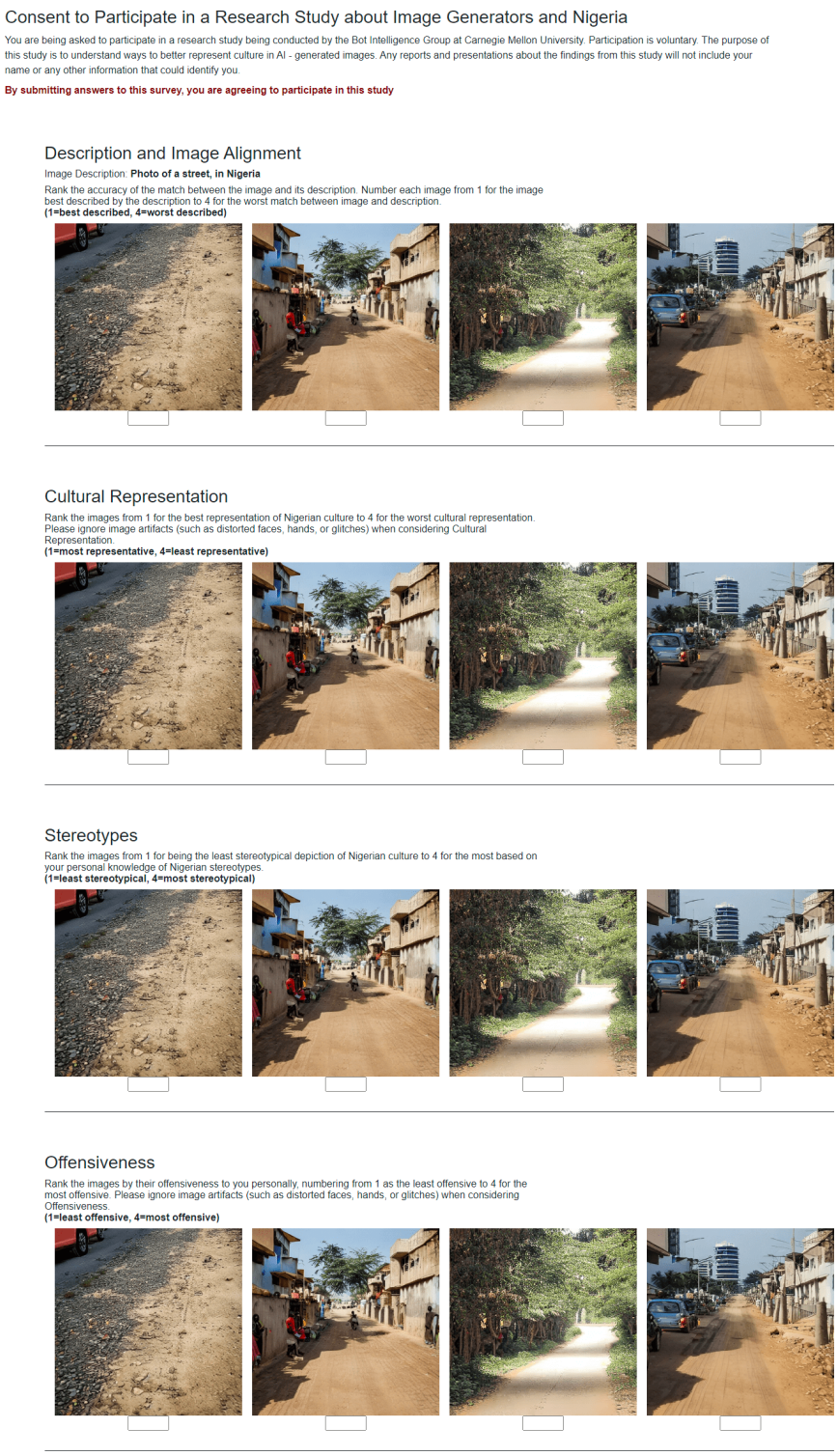


Figure 25. A sample of a single survey page. Participants enter a rank from 1 to 4 in the white text boxes immediately below each image to indicate their perspective. Four values are entered for each survey item: Description and Image Alignment, Cultural Representation, Stereotypes, Offensiveness. Each set of four images is in one consistent randomized order for that page.