Culturally-Aware Stable Diffusion: Supporting Representation with Culturally-Aware Text-to-Image Synthesis

Zhixuan Liu,[1] **Youeun Shin,** [1, 2] **Beverley-Claire Okogwu,** [1] **Youngsik Yun,** [1, 2]
Peter Schaldenbrand, [1*] **Jihie Kim,** [2 *] **Jean Oh** [1 *]

[1] Robotics Institute, Carnegie Mellon University
[2] Department of Artificial Intelligence, Dongguk University
{zhixuan2, youeuns, bokogwu, youngsiy, pschalde}@andrew.cmu.edu, jihie.kim@dgu.edu, jeanoh@cmu.edu

Abstract

It has been shown that accurate representation in media improves the well-being of the people who consume it. By contrast, inaccurate representations can negatively affect viewers and lead to harmful perceptions of other cultures. As artificial intelligence improves in image synthesis tasks and becomes ubiquitous in content creation, special attention will need to be paid to ensure that accurate representation is achieved; however, it is well understood that these models absorb the bias of their training data and can amplify harmful stereotypes. To achieve inclusive representation in generated images, we introduce a new task of culturally-aware text-to-image synthesis; given a specific cultural context, the goal is to generate visual content that is both accurate and inoffensive to the cultural consumers. We then present our proposed approach for culturally-aware text-to-image synthesis, Culturally-Aware Stable Diffusion, comprised of two priming techniques: (1) Fine-tuning a pre-trained text-to-image synthesis model, Stable Diffusion, on a hand-selected, culturally-representative image dataset, and (2) Augmenting the input prompt with additional culturally relevant language data. The culturally relevant data is curated by people who have a personal relationship with that particular culture, and we recruit participants who are a part of that culture to evaluate the method. Our preliminary experiments indicate that priming using both text and image is effective in improving the cultural relevance of generated images.

Introduction

Representation matters. In media, studies repeatedly show that representation affects the well-being of its viewers (Shaw 2010; Caswell et al. 2017; Elbaba 2019). 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 (Castañeda 2018). When people are accurately represented in media, it allows people to properly understand cultures without harmful stereotypes forming (Dixon and Linz 2000; Mastro and Greenberg 2000). Despite the benefits of representation, many media generating Artificial Intelligence (AI) models show poor representation in their results (Ntoutsi et al. 2020). Many of these issues stem from their large training datasets which are gathered by crawling the Internet without filtering supervision and contain malign stereotypes and ethnic slurs among

Figure 1: Sample images generated for six different countries by our proposed Culturally-Aware Stable Diffusion; the images in the first row show the results from the generic Stable Diffusion as references.

*Corresponding authors.

other problematic content (Birhane, Prabhu, and Kahembwe 2021). It was found that large datasets such as the LAION-400M (Schuhmann et al. 2021) used to train many text-to-image synthesis models such as Stable Diffusion (Rombach et al. 2021) are Anglo-centric and Euro-centric (Birhane, Prabhu, and Kahembwe 2021), as shown in the top row of Figure 1.

As AI models are increasingly used to create and aid in the production of visual content, it is important that the models have a true understanding of culture such that it can give accurate and proper representation leading to well-being rewards for its consumers. In this paper, we aim to address such a representation issue in image generation and introduce a new task of *culturally-aware* image synthesis: generating visual content within a cultural context that is both accurate and inoffensive with the overall goal of improving the well-being of consumers of the AI generated images with particular attention to those consumers from underrepresented groups. Specifically, we formulate the culturally-aware text-to-image synthesis task to take an additional input of a country name to specify a cultural context in addition to language description.

A naive approach to generating culturally-aware content in text-to-image synthesis is through prompt engineering to simply concatenate the cultural specification input with language description. While this will alter the appearance of the generated content, there may be incorrect or stereotypical misrepresentations of the culture stemming from the biases in the data used to train the model.

One way towards overcoming the biases in the training data, is to introduce more data that is veritably representative of the culture. For this reason, the approach that we present is powered and controlled by users through the data that they provide to *teach* the AI model about their culture. Culture is an amalgamation and generalization of people but it is also very personal and viewed differently by even every person within a particular culture. Culture can also only be understood properly by a person who has a personal relationship with it. For this reason, we collect a set of priming data by inviting the members of each culture to provide their own data to influence an existing, biased method to more accurately learn about their culture. In our preliminary study, we show that, our proposed approach effectively use relatively small-sized priming data to successfully generate more culturally aligned images than the naive baseline approach.

We experiment with two different techniques to address this novel problem. First, we fine-tune an existing text-to-image synthesis model on a dataset of images that are representative of the culture as defined by a person within that culture. Second, we augment the given text prompt with additional relevant cultural information. The text cultural information is curated by a person who is personally familiar with the culture and contains key words associated with the culture, how those words integrate with a given sentence, and the meaning behind the word. Together, these components make up our proposed approach: Culturally-Aware Stable Diffusion.

We evaluate Culturally-Aware Stable Diffusion's two components individually as well as combined against the baseline of simply specifying the culture in the text prompt. Our evaluation was performed by people who were personally familiar with the culture that the generated images were conditioned upon since these are the people who are most affected by the AI models. While preliminary, our survey results indicate our proposed approach is both less offensive and more culturall relevant than simply adding the country name as a suffix to the prompt. We share the findings from our preliminary experiments to provide a basis for an important aspect of AI generated imagery: that cultural information should be accurately presented and celebrated equitably. Our code will be made publicly available upon acceptance.

Related Work

Culturally Conditioned Machine Learning

Accurately representing culture with Machine Learning is an open challenge. Many models, such as Craiyon (Dayma et al. 2021) fail to capture certain distinguishing features relating to a country's dominant culture (Reviriego and Merino-Gómez 2022). One method to address this involves the inclusion of semantic understanding in a model such as the ERNIE-ViLG 2.0. (Feng et al. 2022). A similar approach can be seen in Japanese Stable Diffusion (Shing and Sawada 2022), which fine tunes the Stable Diffusion U-Net and retrains the text encoder on the 100 million images with Japanese caption within the LAION-5B (Schuhmann et al. 2021) dataset.

While these approaches produce better cultural representations of Japan and China, it is not easy to be used universally. Adapting these approaches requires millions of training examples which cannot be easily met for cultures with less internet presence. Also, these datasets are so large that it is infeasible to vet them for harmful and stereotypical information. Our approach strives for cultural representation using a dataset that is smaller (100-200 images) and hand selectable.

Modifying Text-to-Image Diffusion Models

Diffusion-based text-to-image synthesis models have improved incredibly over the past year in image quality and language understanding; however, these models are still Anglo-centric and contain gender and racial biases at least in part due to the lack of supervision in their large text-image training datasets (Birhane, Prabhu, and Kahembwe 2021). One way to address this problem while leveraging the knowledge obtained from the large dataset is to fine-tune the latent diffusion models. (Ruiz et al. 2022) and (Gal et al. 2022) propose textural inversion methods that allow the latent diffusion models to generate images with specific visual concepts. However, these methods restrict the generation to be an object or a style and cannot generalize on the expression of abstract concepts, for example, a culture. Inspired by the approaches of (Chambon et al. 2022) and (Pinkney 2022), the text-to-image diffusion model can generate domain-specific images by fine-tuning the U-Net of the model using a batch of data from that domain.

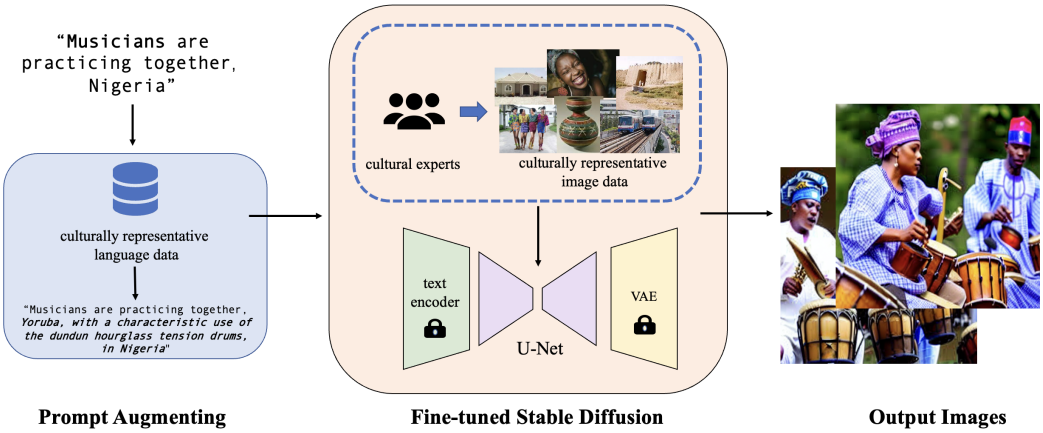

Figure 2: Overview of the workflow for image generation using our (Culturally-Aware Stable Diffusion) method. The text input is first augmented using culturally representative language data, then used as input to our fine-tuned Stable Diffusion, which is trained on a small-scale, carefully selected cultural image dataset.

Prompt Engineering

Originating from the field of Natural Language Processing (NLP), prompt engineering, which can be conceived as programming in natural language (Reynolds and McDonell 2021), is to design the input text prompt to retrieve user-desired outcomes from language models. Inspired by the findings in the NLP domain, researchers in computer vision have been exploring the effects of prompt engineering. In order to present design guidelines for better outcomes in text-to-image generation models (Liu and Chilton 2022), several permutations of prompt engineering using a template were conducted in terms of subject and style in art. To discover some tricks and keywords to boost the quality of the output image in the image generation models such as DALL-E 2 (Ramesh et al. 2022) and Midjourney (Midjourney 2022), various experiments through trial-and-error are ongoing on the Internet as well. Online community (Taylor 2022) has come up with a prompt engineering template for artwork, consisting of terms regarding styles, artists, vibes, and perspectives.

Approach

Our Approach is composed of two priming components as depicted in Figure 2. The basic idea is to prime the models, which have been pretrained using massive yet biased datasets, with a few culturally relevant data points. For this purpose, we collect a small dataset of culturally relevant text and image data to fine-tune the models and enrich the prompts with cultural contexts. Given a text prompt and a name of country/culture, we first augment the prompt with more descriptions using the culturally-aware text data collected from domain experts; we then use the enriched prompts with the culturally fine-tuned model to generate the output images.

Figure 3: Selected images in the culturally representative image dataset for Korean culture and Nigerian culture. Images are collected by cultural experts based on cultural keywords, while there may exist flaws in the automatically image captioning.

Culturally Representative Datasets

A key requirement for the proposed datasets to fine-tune the stable diffusion and influence it to generate culturally appropriate images is that the data must be culturally representative and accurate. To meet this requirement, we propose to create such cultural datasets by engaging the consumers of each culture. Our cultural datasets are collected by the experts who confidently know this culture well or belong to this culture, e.g., for the purpose of our preliminary study, we collect the data for six different cultures: American, Korean, Nigerian, Chinese, Grenadian, and Japanese cultures, solely from the members of each culture.

Following Halpern et al.'s (Halpern 1955) definition of culture, nine categories of food, beverage, clothing, artwork, music, dance, religion, house, and entertainment, building are used to represent cultural elements in our dataset.

As shown in Table 1, our cultural dataset is in a minute

Scale of Cultural Dataset	USA	Korea	Nigeria	China	Grenada	Japan
Number of images in the image dataset	106	177	101	130	99	123
Number of objects in the text dataset	21	143	23	137	21	115

Table 1: The first row of the table shows the number of hand-selected images in the culturally-representative image dataset for six different cultures; the second row shows the number of objects in the culturally-representative text dataset.

scale considering that generic Stable Diffusion was trained on LAION-2B-EN that includes more than 2.3 billion text-image pairs.

Image: For each culture, our image dataset is collected based on the nine cultural keywords, with 10 to 20 relevant images collected under each keyword. Since most of these images do not have text annotations, we automatically generate captions for each image by using BLIP (Li et al. 2022) as in (Pinkney 2022). Some text-image data pairs in our culturally representative image data are shown in Figure 3. We note that there are some flaws in BLIP-generated captions, for example, an ancient painting in the Korean dataset is described as "a group of people playing a game," where visual features are outdated and misleading although it is not incorrect description; a picture of traditional Nigerian musicians is captioned as "people holding an umbrella."

Text: To promote cultural contexts lacked in existing large corpora of paired image-text datasets, we create a cultural language dataset. Building on the dataset collection protocol in a work (Liu et al. 2021) that also dealt with the Western-centric bias of source data, we scrape cultural descriptions from webpages specifically focusing on the nine semantic fields of cultural categories.

Fine-tuning Stable Diffusion

The Stable Diffusion (Rombach et al. 2021) pipeline generates natural images under the condition of text prompts. The input text prompt is firstly encoded using the CLIP (Radford et al. 2021) text encoder. A U-Net architecture model creates the output image encoding by denoising from random noise conditioned upon the encoded text. A Variational Autoencoder (VAE) converts the image encoding into a high-resolution image.

We alter Stable Diffusion to have a more accurate understanding of a given culture to address its known bias towards generating Western-focused imagery. In our approach, following a similar approach to (Ruiz et al. 2022) and (Chambon et al. 2022), the U-Net of the Stable Diffusion model is further trained on the new culturally representative image dataset while keeping the text encoder and autoencoder (VAE) frozen. The fine-tuning of U-Net is equivalent to the training process of original Latent Diffusion Models (LDMs): by minimizing the LDM loss in several denoising time steps. The LDM loss is given by:

$$L_{LDM} := E_{z\sim\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta\left(z_t, t, c\right)\|_2^2\right],$$

where t is the time step; z_t, the latent noise at time t; c, the text encoding of a text prompt; ϵ, the noise sample; and ϵ_θ, the noise estimating U-Net model.

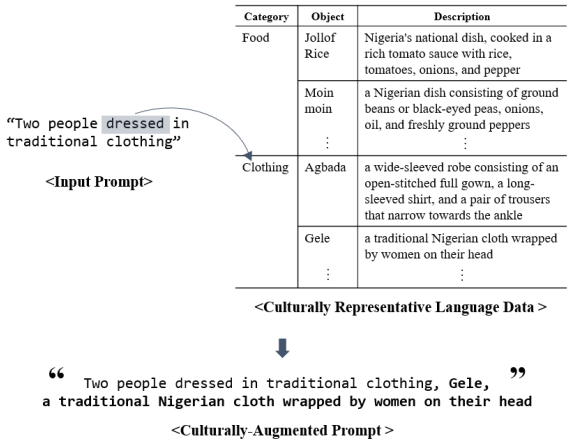

Figure 4: An example of culturally-augmented prompt for Nigerian clothing

We set the learning rate to be small (e.g., 1e-5) to slightly change parameters of the U-Net, and we run 75 epochs on the culturally-aware image data to fine-tune one Stable Diffusion. This generally leads to convergence of the images produced by our fine-tuned Stable Diffusion toward our culturally-aware training data.

Prompt Augmenting

Whereas priming with images is carried out via offline fine-tuning, priming with text is performed for each prompt input in an online manner. With a constant format of text prompt working effectively with the generative models, we construct the template as "$\langle input\ prompt\rangle$, $\langle object\rangle$, $\langle description\rangle$."

Given an input prompt and a name of a culture by a user, an input to the text-to-image model is constructed automatically by augmenting the prompt with an object and its corresponding description from the culturally representative language data, depending on a defined matching collocation. The augmentation process is further explained in Figure 4, along with a part of the language data. As an example, for the category "clothing" which has a common collocation with "wear" and "dress" in English words, any variant of those in the input prompt results in an augmentation by one of the objects from the category "clothing" in the data.

Experiments

In this section, we evaluate our approach through a survey. As an ablation study of the proposed approach, the survey compares our two proposed techniques, culturally-aware prompt engineering and culturally-aware fine-tuning

Figure 5: An ablation study of our Culturally-Aware Stable Diffusion.

over Stable Diffusion, individually as well as in combination. We also compare our proposed approach against a baseline of simply appending the culture to the prompt, e.g., "A family eating dinner *, China*."

In setting up our study, we consider a comparative structure between images: the baseline image versus another image from our results. The setup of a single question in our survey was as follows: Given two images, the participant selects which image best fits three given comparative properties. The properties analyzed were:

1. *Text and Image Alignment*: Participants are given a text prompt and consider which of the two images is more similar to the prompt.
2. *Level of Offense*: Different cultures have different views on what is considered disrespectful or offensive. In the study, participants consider which of the two images is more offensive to them.
3. *Cultural Alignment*: Participants decide which of the two images is a better representation of the country's culture.

To make the study fair, for all the tasks to be performed by the users, the participants were unaware of which images were produced from the baseline or the proposed approaches. The order of the questions was also randomized such that there is no clear sequence to the questions. The participants for the study were selected based on whether they had a personal understanding of the culture for which the images in the survey were generated. Participants were recruited among university students, friends, and family members of the authors. It was ensured that the participants would not be able to discern the approaches used to generate the compared images.

Results

In the following sections, we qualitatively and quantitatively analyze the performance of the proposed approach.

Western Style inherent in Stable Diffusion

The first row of Figure 1 shows the images generated by the original Stable Diffusion given the text inputs, and the second row are those generated by the Culturally-Aware Stable Diffusion for the US culture. As shown in Table 2, based on 336 human evaluations by 84 participants across 6 cultural groups, we observed that there is no clear distinction between these two methods in terms of the representation of American culture. This observation indicates that original Stable Diffusion exhibits strong American inherence and tends to generate western-biased images.

Baselines

Japanese Stable Diffusion: We compare our approach with Japanese Stable Diffusion (Shing and Sawada 2022) in Figure 6. Despite being trained with 100 million images with Japanese captions, Japanese Stable Diffusion still generated images that have a Western bias, as shown in the first row of Figure 6. Our method, which requires only about 100-200 images as training data, can generate images that lie in the domain of Japanese culture based on simple text descriptions, as shown in the second row of Figure 6.

Figure 6: Compare between our (Culturally-Aware Stable Diffusion) method with Japanese Stable Diffusion. Top row shows the language input used to generate images.

Figure 7: Compare between our (Culturally-Aware Stable Diffusion) method with the baseline (Stable Diffusion) in terms of cultural representations. Top row shows the culture name and the second row shows the language input used to generate images.

Original Stable Diffusion: As a baseline, we use the original Stable Diffusion by simply adding a culture name as a suffix to the text prompt. The results in Figure 7 show that this approach fails to capture cultural identity and instead produces many culturally biased images. We qualitatively report the following three issues: (1) The baseline method generates images with elements that are still in the Western form as in Figure 7 (a), where the baseline method generated dancers performing Western ballet dance when the desired cultural context was Japanese, while ours is able to generate images with more representative Japanese elements. (2) The baseline method exhibits cultural misunderstanding, as shown in Figure 7 (b), where images of Japanese style clothing was generated when the desired cultural context was Korean. (3) The baseline generates images that incorporate cultural stereotypes and discrimination, as in Figure 7 (c), the image generated by the baseline approach contains cultural biases and discrimination, failing to represent modern Nigerian culture. Also, in column (d), the baseline-generated image contains offensive, stereotypical elements of Chinese culture, especially in the facial features of the characters. All three of these issues are mitigated in our Culturally-Aware Stable Diffusion approach.

More Western-style images	Stable Diffusion (Baseline)	USA Culturally-aware Stable Diffusion (Ours)	Both of the Approaches	Neither of the Approaches	Total
Number of Answers	61	63	165	47	336
Percentage	18.15%	18.75%	49.11%	13.99%	100%

Table 2: Results of the survey measure the differences between our American Culturally-Aware Stable Diffusion and the baseline, in terms of which method generates more Western-style images. The results shows the number and percentage of preference.

	Text-Image Alignment	Offensive-ness	Cultural Alignment
Fine-Tuned Stable Diffusion	57%*	38%*	54%
Prompt Augmentation	45%	39%*	61%*
Culturally-Aware Stable Diffusion	47%	35%*	62%*

Table 3: Results from a survey measuring how our approach and its individual components compared to the baseline in terms of text-image alignment, offensiveness, and cultural-alignment. Values displayed are the percentage of times our approaches were chosen over the baseline according to the criterion. * indicates statistically significantly different at $p < 0.01$. Results are based on 533 generated image comparisons from 30 participants.

Ablation Study

We investigate our method's performance regarding the effects of fine-tuning Stable Diffusion and prompt augmenting. The qualitative results are summarized in Figure 5. In the streets example, the fine-tuned stable diffusion aids in making the streets appear modern which reduced harmful stereotypes. It is apparent that prompt augmentation improves upon the baseline in displaying more modern more culturally relevant instruments in the musicians example. Combining these, our Culturally-Aware Stable Diffusion model displays modern culture with accurate object representations.

Quantitative Results

In our survey comparing our Culturally-Aware Stable Diffusion and its two subcomponents to a baseline of appending the country name to the end of the text prompt, we asked participants which generated image fit the given text prompt best to quantify if there was degradation in image quality and connection to text. The first column of Table 3 summarizes these findings. Based on 533 image comparisons from 30 participants, there was no significant degradation in text-image alignment. In the same survey, we asked participants which image was more offensive to them. Culturally-Aware Stable Diffusion and its individual components were significantly less offensive than the baseline, Table 3.

We asked participants to write short descriptions about how they made their decisions in the survey. For text-alignment questions, participants wrote that specific objects or aspects of the prompt such as the number of people, the acts of the people, and objects within the image were important to making their decision. Participants stated that particular variations and nuances of the items within the image were important for cultural-alignment. We received responses, "Many Chinese people probably start their morning with a bowl of noodles like this.The different kinds of noodles remind me of Chinese noodles culture.", "I can see the Hanbok", and "The one of the left feels more modern, which I associated American culture with." Likewise in terms of offensiveness, it was also important the the variations and nuances were correct. When a culture was mistaken for another, people wrote that it was offensive. For instance, some of the baseline Korean images appeared Japanese which was very offensive to Korean participants. Other misrepresentations were also stated as very offensive: "If image 1 is used as a typical Chinese culture photo, it would be more offensive, since it looks like the country is underdeveloped."

Discussion

Our representation of culture is nation based in this paper. Most nations have multiple cultures, and culture can exist outside of geographic borders. Our approach is able to account for this by allowing users to tailor the culturally relevant text and image datasets used for prompt augmentation and fine-tuning. In future work, we are creating interfaces to easily allow users of the system to upload image and text data that is relevant to their culture. This data would also be shareable, creating many specific cultural datasets that can be used to improve other image synthesis tasks in terms of cultural representation.

Conclusions

We present early and in-progress work towards the incredibly important task of cultural representation in image synthesis. This paper defines the novel task and offers an approach, Culturally-Aware Stable Diffusion, which is less offensive and more culturally relevant than existing approaches. Our approach improves text-to-image synthesis in cultural alignment and offensiveness over using Stable Diffusion with culture specified via prompt without degrading generated image quality and alignment to the content of the input language prompt. Our approach is lightweight compared to similar work as it requires only about 100 culturally relevant images.

With proper cultural representation in media, more people can find role models and there will be less misconceptions and harmful stereotypes formed about people. We hope that this work will inspire other image synthesis researchers to see the importance of this nuanced task and incorporate it into their work.

Acknowledgments

This work is in part supported by NSF IIS-2112633. Youeun Shin, Youngsik Yun, and Jihie Kim were supported by the MSIT (Ministry of Science, ICT), Korea, under the High-Potential Individuals Global Training Program (RS-2022-00155054) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation) (50%).

References

Birhane, A.; Prabhu, V. U.; and Kahembwe, E. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*.

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

Caswell, M.; Migoni, A. A.; Geraci, N.; and Cifor, M. 2017. 'To be able to imagine otherwise': community archives and the importance of representation. *Archives and Records*, 38(1): 5–26.

Chambon, P.; Bluethgen, C.; Langlotz, C. P.; and Chaudhari, A. 2022. Adapting Pretrained Vision-Language Foundational Models to Medical Imaging Domains. *arXiv preprint arXiv:2210.04133*.

Dayma, B.; Patil, S.; Cuenca, P.; Saifullah, K.; Abraham, T.; Lê Khac, P.; Melas, L.; and Ghosh, R. 2021. Dall· e mini.

Dixon, T. L.; and Linz, D. 2000. Overrepresentation and underrepresentation of African Americans and Latinos as lawbreakers on television news. *Journal of communication*, 50(2): 131–154.

Elbaba, R. 2019. Why on-screen representation matters, according to these teens. *PBS NewsHour*, 14.

Feng, Z.; Zhang, Z.; Yu, X.; Fang, Y.; Li, L.; Chen, X.; Lu, Y.; Liu, J.; Yin, W.; Feng, S.; et al. 2022. ERNIE-ViLG 2.0: Improving Text-to-Image Diffusion Model with Knowledge-Enhanced Mixture-of-Denoising-Experts. *arXiv preprint arXiv:2210.15257*.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Halpern, B. 1955. The Dynamic Elements of Culture. *Ethics*, 65(4): 235–249.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*.

Liu, F.; Bugliarello, E.; Ponti, E. M.; Reddy, S.; Collier, N.; and Elliott, D. 2021. Visually Grounded Reasoning across Languages and Cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 10467–10485. Association for Computational Linguistics.

Liu, V.; and Chilton, L. B. 2022. Design Guidelines for Prompt Engineering Text-to-Image Generative Models. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*.

Mastro, D. E.; and Greenberg, B. S. 2000. The portrayal of racial minorities on prime time television. *Journal of Broadcasting & Electronic Media*, 44(4): 690–703.

Midjourney. 2022. Midjourney.

Ntoutsi, E.; Fafalios, P.; Gadiraju, U.; Iosifidis, V.; Nejdl, W.; Vidal, M.-E.; Ruggieri, S.; Turini, F.; Papadopoulos, S.; Krasanakis, E.; et al. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3): e1356.

Pinkney, J. 2022. Text to Pokemon Generator.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. *CoRR*, abs/2103.00020.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *ArXiv*, abs/2204.06125.

Reviriego, P.; and Merino-Gómez, E. 2022. Text to Image Generation: Leaving no Language Behind. *arXiv preprint arXiv:2208.09333*.

Reynolds, L.; and McDonell, K. 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2022. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*.

Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Shaw, A. 2010. *Identity, identification, and media representation in video game play: An audience reception study*. Ph.D. thesis, University of Pennsylvania.

Shing, M.; and Sawada, K. 2022. Japanese Stable Diffusion.

Taylor, M. 2022. Prompt Engineering: From Words to Art.