# Partiality and Misconception: Investigating Cultural Representativeness in Text-To-Image Models


Lili Zhang
lilian@hainanu.edu.cn
Hainan University
Haikou, China

Xi Liao
liaoxi@hainanu.edu.cn
Hainan University
Haikou, China

Zaijia Yang
sunshineyang@hainanu.edu.cn
Hainan University
Haikou, China

Baihang Gao
gaobh@hainanu.edu.cn
Hainan University
Haikou, China

Chunjie Wang
wangcj18@hainanu.edu.cn
Hainan University
Haikou, China

Qiuling Yang
qlyang@hainanu.edu.cn
Hainan University, China Innovation Platform for Academicians of Hainan Province
Haikou, China

Deshun Li*
lideshun@hainanu.edu.cn
Hainan University, China Innovation Platform for Academicians of Hainan Province
Haikou, China



## ABSTRACT

Text-to-image (T2I) models enable users worldwide to create high-definition and realistic images through text prompts, where the underrepresentation and potential misinformation of images have raised growing concerns. However, few existing works examine cultural representativeness, especially involving whether the generated content can fairly and accurately reflect global cultures. Combining automated and human methods, we investigate this issue in multiple dimensions quantificationally and conduct a set of evaluations on three prevailing T2I models (DALL-E v2, Stable Diffusion v1.5 and v2.1). Introducing attributes of cultural cluster and subject, we provide a fresh interdisciplinary perspective to bias analysis. The benchmark dataset UCOGC is presented, which encompasses authentic images of unique cultural objects from global clusters. Our results reveal that the culture of a disadvantaged country is prone to be neglected, some specified subjects often present a stereotype or a simple patchwork of elements, and over half of cultural objects are mispresented.



## CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; • **Human-centered computing**; • **Social and professional topics** → **Cultural characteristics**;

## KEYWORDS

text-to-image generation, cultural representativeness, cultural cluster, bias, stereotype

**ACM Reference Format:**
Lili Zhang, Xi Liao, Zaijia Yang, Baihang Gao, Chunjie Wang, Qiuling Yang, and Deshun Li. 2024. Partiality and Misconception: Investigating Cultural Representativeness in Text-To-Image Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 25 pages. https://doi.org/10.1145/3613904.3642877


## 1 INTRODUCTION

Text-to-image (T2I) generative models are progressively being embraced as a valuable tool for creative expression. Through the utilization of simple text prompts, millions of high-definition and realistic images[5] are generated and subsequently disseminated widely via various media. Some models unveiled in the recent period, such as Dall-E v2[57] and Stable Diffusion[59], have achieved consecutive breakthroughs in fidelity and alignment performance[16, 25, 59], sparking significant intrigue among both scholarly practitioners and the general populace. Nevertheless, the T2I models also confront a problem common to ML models: the biased reflection of the world's realities[11, 61]. Machine programs undergo training through data, and if the data samples carry biases, the machine consequently assimilates the inherent biases in the learning


*Corresponding author




*CHI '24, May 11–16, 2024, Honolulu, HI, USA*



ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642877

process[11, 65]. Similarly, T2I models learn and generate new visual content from abundant internet-sourced images, where latent biases can also be transmitted and intensified.

In response to several studies' appeal to acknowledge biases in text-to-image generation[5, 7, 13, 38, 81], a few research initiatives have focused on race, gender, and age perspectives. However, an integral aspect of this-cultural representativeness-largely remains unexplored. Limited understanding of culture leads to impoverished discussions and imagination, and inadequate discourse exacerbates cultural biases that undermine comprehension[12, 34, 78]. Hence the underlying biases and potential misinformation of disadvantaged countries also shape cultural misrepresentations in T2I models, where inequitable distribution and erroneous interpretation manifest subtly in generated images. These misrepresentations covertly spread and are assimilated within society, further disempowering and marginalizing disadvantaged cultures.

For example, nations with substantial populations, including India and China, witness limited inclusion of their cultural characteristics within most text-image datasets[65]. When generating content related to cultures such as food and architecture, the T2I models will automatically add details if the region is not specified. As a result of not specifying the country, there are very few images that contain the cultural features of the two countries, which is quite unfair for them. Even if we specify to generate a particular cultural object from the countries, it still may not be accurately portrayed. When asking models to generate images of Gopuram, a well-known architectural form in India, a lot of generated images had low similarity to reality. That is, models may not have encountered and learned the relevant data, or even if learned, it might have been partial or incorrect, leading to the possibility of an erroneous portrayal due to a lack of comprehensive understanding.

In this work, we proposed a set of quantitative methods to scrutinize whether the generated content can fairly and accurately reflect global cultures. The complete experimental procedure and methodology are illustrated in Figure 1. Given the multidimensional nature of culture, our study is closely aligned with previous global cross-cultural studies. We selected three countries as representatives from each of the ten cultural clusters defined in the previous GLOBE study and nine cultural subjects, including both material and non-material culture. By using three prevailing T2I models: DALL-E v2, Stable Diffusion v1.5 and v2.1, a large number of images were generated for cultural representativeness analysis via inputting three types of prompts.

We conducted feature extraction from generated images using prompts specifying cultural subjects and country names (expanded prompts). By employing these features to categorize images whose prompts have no regional constraints(neutral prompts), we can analyze their distribution by default. Observing these images generated with neutral prompts, we find that there are problems of uneven diversity and the patchwork of elements, thus further assessing them in quantitative manners. To measure whether each specific cultural object can be accurately portrayed, the Unique Cultural Objects from Global Clusters (UCOGC) dataset was established, serving as a benchmark to test the similarity between generated images and authentic images, namely fidelity. In the quantitative evaluation procedure, we particularly noted the problems that previous works investigating T2I models exist, where

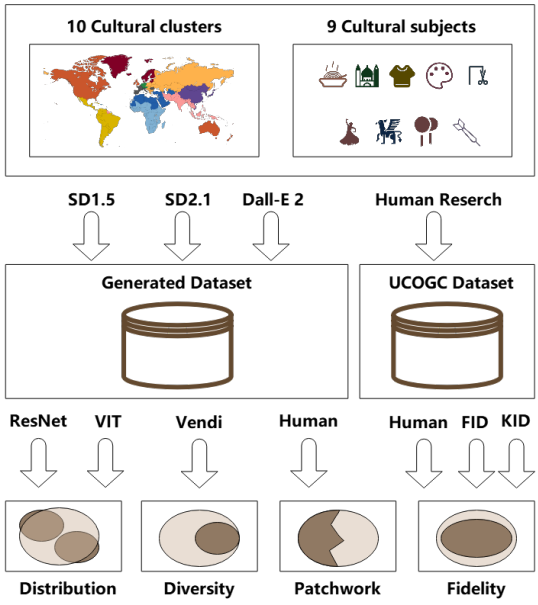



**Figure 1: Our approach to evaluating cultural representativeness in T2I Models. 10 cultural clusters and 9 subjects are introduced. A set of hybrid machine-human evaluation methods is used to analyze the generated dataset and UCOGC dataset.**

the metrics were inconsistent[48, 68] with human perceptions and lack of interpretation[48, 68] and the human evaluation was unreliable and unrepeatable[48], thus devising a hybrid machine-human evaluation methodology to ensure the reliability of our results.

The contributions of our research can be summarized as follows:

(1) Introducing Attributes of cultural clusters and subjects for analyzing cultural representativeness, providing a fresh perspective and interdisciplinary approach to multimodal bias analysis in cultural representativeness.
(2) Analyzing 193,700 generated images and 37,600 authentic images, our research is the first comprehensive exploration of T2I models' performance in cultural representativeness, uncovering biases in distribution, diversity, patchwork, and fidelity.
(3) The UCOGC dataset, proposed by us, encompasses authentic images of diverse cultural subjects from 30 countries. Serving as a benchmark for T2I models, it also supports training visual models for accurate global cultural object recognition.

## 2 RELATED WORK

### 2.1 Cultural Cluster and Categorization

The cross-national research on cultural differences has a rich history, starting from Hofstede's initial study involving 53 countries[26, 27] to the later research conducted by Inglehart and Welzel[4], covering 81 nations. In 2004, the renowned GLOBE study[28] categorized 62 countries into 10 clusters based on their proposed six dimensions. Due to the high complexity of cultural studies, it was difficult to

conduct these studies in every country of the world, but in 2013, extension research on the groundwork of the GLOBE study investigated most of the countries of the world and categorized all of them into the 10 cultural clusters defined by the GLOBE study. Based on the results of this extension study's delineation of most countries in the world into cultural clusters, we selected representative countries from each cluster.

There are many theories on the categorization of culture, and there is no uniform criterion for the division of cultural subcategories as subjects, each of which often overlaps or duplicates the other. Building upon E. B. Tylor's classic definition[74] of culture, culture has ceased to be a mere conceptual category and it is not the dichotomous term of material culture. Instead, culture can be categorized into two types: material culture and nonmaterial culture. Reynolds formally defined the concept of material culture[58] in 1987. In sociology, Material culture refers to the physical objects, artifacts, and tangible elements that are part of a society's cultural expression, while nonmaterial culture encompasses intangible aspects such as beliefs, values, norms, and language[32]. Compared to other non-uniform ways of categorizing cultures, a direct division of culture into the two types is undoubtedly more succinct and without omission. To carry out a comprehensive investigation of cultural representativeness, it is necessary to consider both the material culture and the nonmaterial culture.

## 2.2 Cultural Datasets

In recent times, the challenge of inherent biases within datasets has been a persistent concern for researchers[40, 65]. To address this issue, more and more research has shifted towards building datasets that encompass a comprehensive range of cultures and content from around the world[78]. XCOPA, a multilingual dataset[51], was developed to address causal commonsense reasoning in eleven different languages. A dataset from a study on the religion and morality project includes data on demographics, religious beliefs and practices, material security, and intergroup perceptions[52]. Many of the datasets proposed in recent years covering images of food, architecture, and clothing[3, 29, 46, 55, 72] from different countries can reflect the culture of the world to a certain extent. The newly proposed CCUB[37], a cross-cultural understanding benchmark dataset, covers image and caption pairs in food & drink, clothing, artwork, dance & music in eight countries at the same time. However, this dataset selected only eight countries as representative studies with vague criteria, and only about one or two hundred images were collected for the nine cultural elements it introduced.

## 2.3 Bias in Multimodal Models

The social biases present in language models[1, 39, 43, 69, 71] or vision models[14, 54, 56, 73, 82] have been extensively studied. For example, in word embedding, computer programmer is closer to man, while homemaker is closer to women[9]. In facial recognition tasks, dark-skinned women have significantly higher error rates compared to light-skinned women[11]. A series of studies on cultural bias[70] in language models[2, 43, 45, 60] indicate that English language models often contain biases against other cultures around the world.

The latest advances in multimodal artificial intelligence are successfully integrating language and vision models and applying them to various domains of society, as well as increasing the potential vulnerabilities and risks. A growing number of multimodal studies have found that social biases, such as ethnic and occupational biases from the dataset, are disseminated to downstream output tasks[77, 80], for instance, image search[41, 42, 84] and image captioning[6, 23, 83]. In image search, females and some underrepresented ethnic groups perform poorly in the results of occupational queries[30, 41]. In image captioning, image captioning models tend to exaggerate biases present in training data to predict demographic traits[23]. A study advocating the application of multimodal models to digital humanities emphasizes that researchers have to scrutinize the output of multimodal models that may contain cultural bias[67]. In this paper, we focus on the representativeness study of T2I models in the cultural domain, further complementing the bias investigation of multimodal models.

## 2.4 Text-to-Image Models and Their Biases

Riding the wave of burgeoning interest in generative artificial intelligence, text-to-image generative models have been widely embraced and put into practice. A generative model possessing the capability to accurately comprehend natural language and visual concepts and generate high-fidelity images tends to garner greater popularity. Compared to the previously proposed Generative Adversarial Networks (GANs)[38] and Variational Autoencoders[76], as well as their variants, the recently proposed models exhibit superior computational efficiency and the ability to produce high-quality samples. One such example is DALL-E v2[57], which is done by encoding textual descriptions using CLIP[53] to a high-dimensional vector representation, namely CLIP text embeddings, and then converting them via a prior to corresponding image embeddings. Once the image embeddings are generated, an image decoder is applied to generate the final image. Another text-to-image model, Stable Diffusion[59], which is trained on LAION-400 and -5B[63, 64] and is fully open-source, has also gained a lot of attention. Apart from these, a variety of new T2I models, such as Imagen[62], Parti[79], Make-A-Scene[20], and Midjourney, are emerging.

The models mentioned above rely heavily on large datasets crawled from the Internet, where toxic, stereotypical, and biased content[22, 47] is also probably displayed or even amplified by the models. A series of studies reveal that T2I models exhibit different degrees of bias in gender, race, age, geography, etc[5, 38, 44, 75]., while fewer works explore the cultural domain. An investigation of cultural biases in T2I models through homoglyphs shows that the generated content will reflect cultural stereotypes and bias when inserting single non-Latin characters in a textual description[70]. Here, we focus on multiple cultural subjects across the globe, quantitatively analyzing the cultural bias of T2I models.

# 3 METHODOLOGY

Considering the intersectionality of knowledge involved in the topic of cultural representativeness, we are particularly attentive to the conjunction with the humanities regarding the considered dimensions. Because many T2I models cannot be directly probed for their internal structure, we use the black-box method to conduct

experiments. In our evaluation, we chose automated methods that are as valid and interpretable as possible and alternatively replaced or supplemented them with qualified human methods.

## 3.1 Dimensions of Cultural Representativeness

To improve the statistical analysis of cultural representation in the T2I model, we introduced two key attributes: cultural cluster and cultural subject. Cultural cluster emphasizes the spatial property of culture, while cultural subject covers the two general types of cultural content, material culture and nonmaterial culture, more comprehensively.

*3.1.1 Cultural Clusters.* In contrast to most of the previous work, which simply selected representatives by geography, for example, selecting three countries on each of the four continents, we specifically set up the attribute of cultural clusters. Simply dividing representative countries based on geographic location does not fully represent global cultures. In the study of cultural differences, different countries can be classified into the same cultural cluster, which means these cultures have similarities in some aspects. By selecting a certain sample of countries from the defined cultural clusters, the general characteristics of the clusters can be reflected.

In our investigation, global culture was divided into ten cultural clusters (AFRICAN, ANGLO-SAXON, CONFUCIAN, EASTERN EUROPEAN, GERMAN, LATIN-AMERICAN, LATIN-EUROPEAN, MIDDLE EASTERN, NORDIC, SOUTH-EAST ASIAN) based on the GLOBE study. Almost all countries of the world were classified into corresponding cultural clusters. We selected three countries from each cluster and used them as expanded prompts to facilitate checking the cultural representation of T2I models around the world.

We chose three countries that have the largest populations to represent their cultural clusters, e.g., China, Japan, and Vietnam are chosen to represent the CONFUCIAN cluster, and the USA, Canada, and the UK are chosen to represent the ANGLO-SAXON cluster[1] (Listed are all the selected countries in each cluster, Table 5 in Appendix A).

*3.1.2 Cultural Subjects.* Most of the proposed datasets that can reflect the cultures of different regions tended to focus on food, architecture, and other single aspects, which cannot reflect the culture of the region in its entirety. Even if multiple aspects of culture were selected simultaneously, only the separate selection of cultural elements was considered, ignoring the representativeness of cultural elements in cultural categorization, which may easily lead to an imbalance in the representation of the relevant results in material or immaterial culture. To comprehensively assess cultural representativeness, the selection of cultural subjects should include both material and nonmaterial cultures. On the material side, we chose five subjects (apparel, food, architecture, artwork, handicraft). On the nonmaterial side, we chose four subjects to represent (mythical figure, performance, game activity, festival activity).

The above subjects were chosen for two reasons: firstly, to ensure the universality of the subjects in each country because some subjects. Knowledge and practices concerning nature and the universe, like local fauna and flora or traditional healing systems, are only relevant in some countries but not in others, so they cannot be used to compare horizontally the model's representation of culture in different regions. Secondly, to ensure that these cultural objects are easy to visualize and easy to recognize from T2I models, some complex subjects, such as ritual, are portrayed with low quality and therefore omitted.

## 3.2 Data Collecting

*3.2.1 Prompt Creation.*

*Neutral and Expanded Prompts.* We use two types of prompts to generate various types of cultural subjects, country-specific and country-neutral. For neutral prompts without country specification, we set them to "photograph of a special [subject] from a country", and the text generation model will automatically add their geographical features. For expanded prompts, we set it to "photograph of a special [subject] from [country]" Before this, we tried prefixes such as "high definition image of ", "a kind of", etc., but found that the images generated with "photograph of" were more realistic. We chose the adjective "special" as a qualifier, which emphasizes the unique cultural content of different regions more than "typical" and "common". We also experimented with the word "traditional", and found that countries with longer histories were more likely to appear even if they were not specified.

*Specific Cultural Object Prompts.* To assess the model's capacity to describe specific cultural objects,three categories of specific cultural objects are collected for each of the nine cultural subjects in 30 countries, for instance, under the category of food in Egypt (MID-DLE EASTERN cluster), we select Hamaam, Kushari, and Falafel three categories of specialized food. The generated images of these specific categories are produced through the format of "photograph of [object], [subject]". To ensure that our selection of cultural objects is representative, we first conduct a manual search, and then we examine the selected objects with experts to ensure that they are suitable. The selection of cultural objects is also an important part of our work to build the UCOGC dataset, which will be described in detail in Section 3.2.3.

*3.2.2 Image Generation.*

*Size of the Generated Images.* With neutral prompts, we generated 1,500 images (40,500 images in total) for each cultural subject on the three models. With expanded prompts, we generated 50 images per cultural subject for 30 countries (40,500 images in total). With cultural object prompts, we generated 50 images for each of the selected 752 cultural objects for different countries and subjects (112,700 images in total[2] ). Figure 2 shows an example of generating corresponding images with neutral and expanded prompts.

*Configuration of T2I Models.* For Stable Diffusion v1.5[3] and v2.1[4], we employed the standard parameters provided by Huggingface's

[1]Since the cultural characteristics of the top three populous countries in LATIN-EUROPEAN and EASTERN EUROPEAN are too alike and would not facilitate the subsequent selection of unique cultural objects, we replace those countries having overlapping characteristics with Spain and Poland, respectively.

[2]DALL-E v2 generates only 750 cultural objects because the terms "Isiagu" from Nigerian apparel and "gypsy jazz performance" from Belgium were unable to be generated properly. The images seem to undergo censorship and interception, possibly due to the model mistakenly categorizing these terms as related to prohibited language.

[3]https://huggingface.co/runwayml/stable-diffusion-v1-5

[4]https://huggingface.co/stabilityai/stable-diffusion-2-1

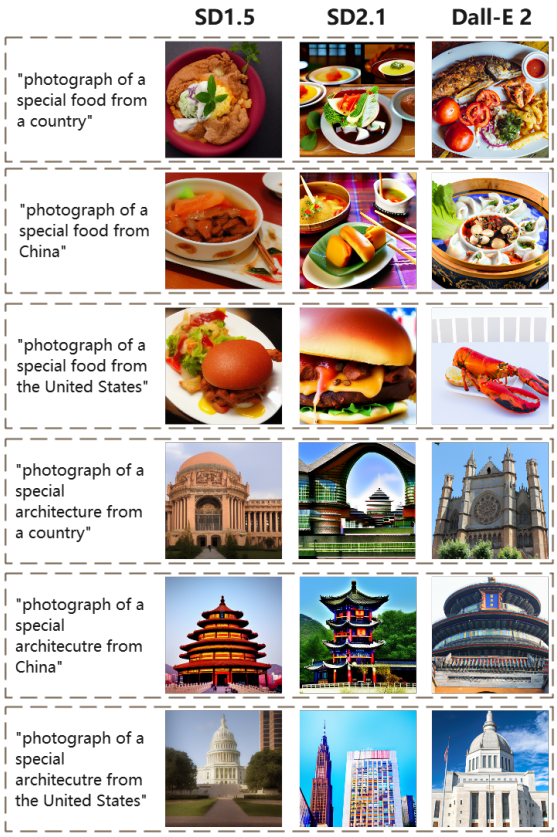



**Figure 2: Images generated by SD1.5, SD2.1, Dall-E 2 under different neutral and expanded prompts on the subjects of food and architecture.**

API to generate images, with an image size of 512×512 pixels and sampling steps of 20. For DALL-E v2[5], we created the images using OpenAI's corresponding API, also with an image size of 512×512 pixels. While the generated images from DALL-E v2 and Stable Diffusion v1.5 exhibit vivid colors and realism, it has been observed that Stable Diffusion v2.1 is prone to including a substantial number of black-and-white photos. This issue could potentially impact our feature extraction process. To mitigate this, we utilize a negative prompt function to filter out "Gray-scale image" from Stable Diffusion v2.1, thereby preventing the generation of black and white pictures.

#### 3.2.3 *UCOGC Benchmark Dataset.*

*Item selection.* As mentioned earlier, our survey spans 30 countries across all ten cultural clusters, and three representative unique cultural objects were selected for each cultural subject. The process of selecting cultural objects is as follows: we first searched through Google Chrome to find articles about the cultural objects of a country's related subject and crawl the names of cultural objects that

[5]https://openai.com/dall-e-2

**Table 1: Scale of the Generated Dataset and UCOGC Dataset Used in the Experiments**

| | Neutral Prompts | Expanded Prompts | Cultural Object | Total |
|---|---|---|---|---|
| SD1.5 | 13500 | 13500 | 37600 | 64600 |
| SD2.1 | 13500 | 13500 | 37600 | 64600 |
| Dall-E 2 | 13500 | 13500 | 37500 | 64500 |
| UCOGC | - | - | 37600 | 37600 |
| Total | 40500 | 40500 | 150300 | 231300 |

appear more frequently. To ensure the representation of these cultural objects, we invited two professors who are experienced in the field of global cross-culture to assist our research team in selecting the most suitable three cultural objects for each country based on the principle of their reputation and recognizability. For example, for China, we choose Mantou, Baozi, and Shao Bing for the food subject; Hanfu, Tang Suit, and Cheongsam for the apparel subject; and Peking Opera, Dragon Boat Racing Huagu Opera for the performance subject. (Listed are the cultural objects we identified for each region and subject, Table 6 in Appendix B).

Because different regions have different histories and different cultural characteristics, we took special consideration of the universality of cultural objects in each country and subject. With the simultaneous consideration of 30 representative countries in all cultural clusters, it is somewhat inevitable that the representative cultural objects of some subjects, such as architecture in Nigeria and the DRC and craft in Pakistan and Belgium, are scarce. In follow-up work, we will seek more senior cultural experts from around the world to refine the dataset further.

*Authentic Images Collection.* We identified 752 cultural objects across 30 representative countries in 10 cultural clusters with nine material and nonmaterial cultural subjects. Through manual search, we collected 50 authentic images for each cultural object to build the Unique Cultural Objects from Global Clusters (UCOGC) dataset. To avoid possible erroneous results in manual searches, we checked the characteristics of each cultural object along with literature and books under the guidance of culture professors. This guarantees the correctness of the UCOGC dataset and provides a benchmark for the unique cultural objects around the world.

By comparing the generated images with the corresponding authentic images from the UCOGC, we can evaluate the discrepancies between the generated image dataset and the benchmark dataset, and thus measure the ability of the T2I model to portray the cultural objects accurately. Figure 3 (a) shows an example of data from the UCOGC dataset, and Figure 3 (b) shows images generated from the collected cultural object prompts. Table 1 presents the overall scale of both the UCOGC dataset and the image dataset we generated earlier. Although only one of the four evaluation aspects of cultural representativeness in T2I models: fidelity, is used with UCOGC in this paper, it is obvious that the proposal of UCOGC is of great significance. As the first dataset of cultural objects spanning across cultural clusters around the globe and covering both material and nonmaterial cultural subjects, it provides a benchmark for evaluating the quality of generated culture representativeness in T2I models. Moreover, due to the comprehensiveness and diversity of the UCOGC dataset, it can help other multimodal tasks such as

image search and image captioning, which will effectively improve their performance in cultural representativeness.

## 3.3 Evaluation

We identified and quantified the following four observations: distribution of neutral generation, diversity of expanded generation, the patchwork of elements, and fidelity of cultural objects. For distribution, we use ResNet[21] and ViT[17] transfer learning; for diversity, we introduce the Vendi score[19]; for patchwork, we perform manual annotation; for fidelity, we perform both classical FID[24] and KID[8] scores as well as manual evaluation on the generated dataset of cultural objects.

*3.3.1 ViT and ResNet Transfer Learning.* To analyze the distributional features of images generated with neutral prompts, we can classify these images to identify which regions these images match with. How to build the classifier is the core of the task. Due to the large number of countries and subjects involved in generating images, there is a limit to the number of images generated for each category in the case of a given country. Training the neural network for these limited data tends to have the problems of overfitting and low capacity of generalization. Another approach is to use zero-sample classification such as CLIP[53] and BLIP[35], but the accuracy of these zero-sample classifiers is questionable due to the scarcity of cultural data itself. Therefore, the best way is through transfer learning[50]. The problem of insufficient samples and poor performance of models for specific domains is solved by using pre-trained models to learn the features of the images generated by the model for a given country. There are two dominant architectures in deep learning: convolutional neural network (CNN)[31] and vision transformer (ViT)[17]. We selected two models from each of the two architectures that have outstanding performance: Resnet50V2, a variant of residual neural networks that performs better than ResNet50 and ResNet101 on the ImageNet dataset. ViT-B/16 is the base vision transformer pre-trained on ImageNet-21k[15, 21] (More details of the training process of the models, Appendix C). The images generated for the countries specified under each culture subject were learned separately. The images generated for the neutral prompts without specifying country were then fed into the classifier, and then the distributional preferences of each cultural subject for different regions were obtained. In this course, the 40,500 expanded prompts images were first fed to the pre-trained ViT with ResNet, and the 40,500 neutrally generated images were classified to get the partiality of each model.

*3.3.2 Qualified Human Assessment.* For image generation models, most works use automated tests to validate the fidelity of their models, such as FID, but it is criticized for inadequate assessment of non-ImageNet datasets[10] and being inconsistent with human senses[49]. Human assessment is crucial to validate the performance of text-to-image generation models, yet it suffers from the problems of unreliability and irreducibility[48]. To ensure the quality of the human evaluation, we guided 47 annotators to learn the features of the collected 752 categories of cultural objects (48 categories per person) under training by two professors mentioned before who are experts in the field of world culture studies. After passing a qualification test, using the UCOGC images as a reference, the annotators were asked to start scoring the images generated from the cultural object prompts on a 5-point Likert scale (The interface diagrams for the qualification test and similarity scoring, Figures 6(a) and (b) in Appendix D). After scoring any 10 of the 50 collected images taken from each generated cultural object by three annotators, the mean value is taken as the model's capability to accurately portray the cultural object. Besides, the evaluation process can refer to the standard dataset at any time.

We also found that some cultural subjects simply patchwork the elements of the country. For example, national flag elements were embedded into apparel patterns or performance scenarios. This is due to the model's failure to learn the accurate representation of specific cultural subjects in a given country, resulting from semantic misunderstandings. For the images generated by the expended prompts of the specified country and subject, we counted the number of occurrences of patchwork in each subject by manual recognition. All images generated from expanded prompts were assessed three times, and if two or more people had the sense of occurrence of the patchwork, it was determined that there was a patchwork (Questionnaire as seen by the annotators conducting the patchwork assessment, Figure 7(c) in Appendix D).

*3.3.3 Vendi, FID and KID.*

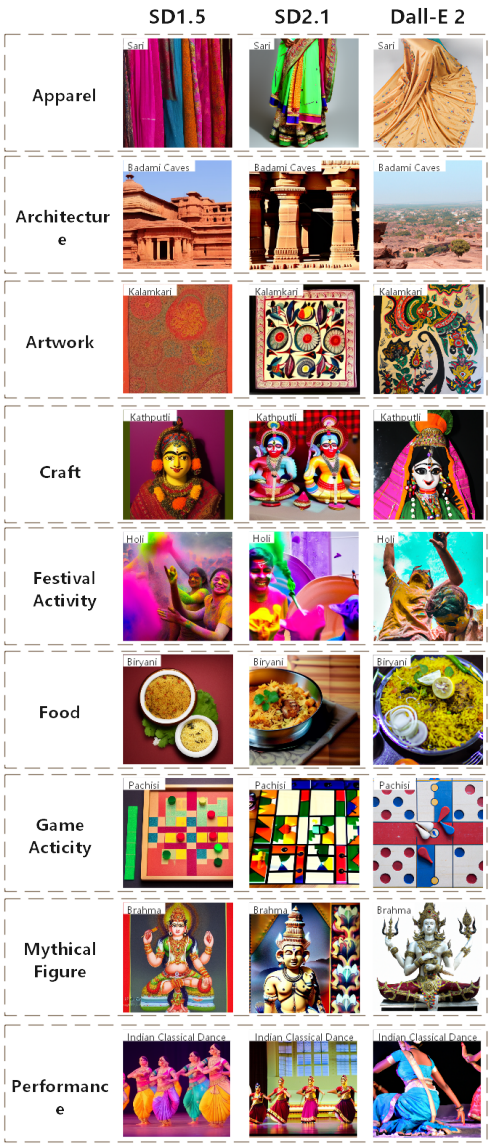
*Vendi score.* In the process of counting the patchwork, we found that subjects in some countries were very stereotyped and lacked diversity compared to other images in several clusters. Some studies have proposed to use Precision/Recall and Density/Coverage[68] to measure the diversity of the generated samples, but this requires comparing with the training samples. As a black-box evaluation, we do not have access to such training data, and the Vendi score[19], a matric based on the exponential of the entropy of the eigenvalues of a similarity matrix, solves this problem. It does not need to refer to the distributions of the sample labels from the original dataset. Under each model, only the generated images are needed, and the Vendi score for them can be computed. With Vendi Score, we implemented a diversity analysis of the images that were generated with an expanded prompt.

*FID and KID Score.* As mentioned, Fréchet inception distance (FID)[24] is one of the most commonly used metrics in the field of image generation, which calculates the Frechette distance between the real and generated sample sets using a pre-trained InceptionV3 network. The disadvantage of FID is that it is biased because its expected value over finite data is not its true value[50]. Kernel Inception Distance (KID)[8], measures the dissimilarity between two probability distributions by employing independently drawn samples from each distribution. KID does not assume a parametric form for the distribution of activation and is unbiased. We selected the two metrics as a comparison to the manual assessment. Divided by subjects, of each model, the FID and KID scores of the generated cultural object dataset compared to the UCOGC dataset are calculated, which can be used to represent the fidelity of the generative content.

# 4 RESULT

After thorough quantitative analysis, we found that the culture of a disadvantaged country is prone to be neglected, some specified


(a) Examples of Cultural Object Images Included in UCOGC Dataset. Illustrated with Apparel and Festival Activity from Five Countries.




(b) Examples of Cultural Object Images Generated by Three T2I Models.

**Figure 3: Comparison of UCOGC dataset and generated dataset.**

subjects often present a stereotype or a simple patchwork of elements, and over half of the cultural objects are mispresented. The following is a detailed results of the assessment and analysis.

## 4.1 Distribution of Neutral Generation

Distributional results can be obtained from classifiers trained by expanding prompts by feeding images generated with neutral prompts. Table 2 illustrates the overall classification results of the ViT and Resnet classifiers for the unspecified generated images. Taking the generated images of expanded prompts for each subject as the training data, one Resnet and one ViT model were trained (totaling 18 classifiers). The images generated with neutral prompts for each subject were then classified with the corresponding transferring models. The distribution probability of the subject in 30 countries can be obtained. There is a certain consistency between the results of the two classifiers (distribution of generated images from neutral prompts after being categorized, Figure 8 in Appendix E). In Stable Diffusion v1.5, for example, Apparel is more likely to favor EST-EUR, Architecture is more likely to favor LAT-EUR, and Mythical figure is more likely to favor ANGLO when no country is

**Table 2: Fine-tuning the pre-trained models based on the images generated from the expanded prompts. The overall classification results were obtained for images generated from neutral prompts. The distribution percentages for each cultural cluster in the table are the means of distribution of the nine subjects across the clusters.**

| | SD1.5 | | SD2.1 | | Dall-E 2 | |
|---|---|---|---|---|---|---|
| | Resnet | ViT | Resnet | ViT | Resnet | ViT |
| CONFUC | 22.90% | 8.50% | 14.30% | 4.50% | 15.20% | 18.40% |
| GERMAN | 8.50% | 13.30% | 2.20% | 5.10% | 6.80% | 3.70% |
| LAT-EUR | 3.00% | 3.90% | 18.90% | 24.40% | 9.80% | 7.00% |
| EST-EUR | 12.60% | 17.40% | 25.40% | 22.00% | 6.30% | 1.60% |
| MID-EAST | 12.50% | 16.60% | 7.90% | 17.60% | 14.40% | 11.90% |
| SE-ASIAN | 8.60% | 6.80% | 6.40% | 3.10% | 3.00% | 0.90% |
| ANGLO | 8.20% | 10.50% | 3.80% | 2.70% | 23.60% | 23.70% |
| LAT-AME | 3.70% | 0.30% | 3.60% | 10.40% | 2.00% | 1.50% |
| NORDIC | 10.20% | 13.70% | 11.80% | 4.80% | 13.60% | 25.30% |
| AFRIC | 9.70% | 9.00% | 5.80% | 5.40% | 5.20% | 6.00% |

specified. However, due to the different features extracted from the pre-training of the two models, there are differences in the results of the classification after training. Here we take the mean value of their classification as the distribution result. To analyze further which countries are being overrepresented and which are being neglected, we compared them against the population[6] and GDP share[7,8] of the corresponding cultural cluster in 2021. As can be visually observed in Figure 4, if a T2I model is fair, the distribution results of images generated with neutral prompts should align with the demographic composition of the world. Clearly, the three models show bias that deviates from population share as they generate cultural content. Observation reveals that the SE-ASIAN and AFRIC clusters are the most heavily overlooked, with the average share distribution of the three T2I models at only 16.5% and 47% of their actual populations (Table 7 in Appendix D). The farther the population share is from the GDP share[33] to the y-axis, which means the lower the GDP per capita is, the less the cultural features of the clusters are present in the generated images. This suggests that the model is more partial to countries with advanced levels of economic development, and such an unfairness has the potential to further exacerbate the underrepresentation of disadvantaged countries.

## 4.2 Diversity of Expanded Generation

We chose the Vendi score, which is highly interpretive and does not require the training dataset to analyze the diversity. Since we discovered that the diversity of the images generated by the expanded prompts is uneven, we quantified it with the Vendi scores. The line graphs of the average Vendi scores of countries and clusters (Figure 9 in Appendix F) show that different countries vary greatly. The higher the Vendi score, the higher the diversity of the dataset. Calculating Kendall's Coefficient of Concordance for these data from

[6]https://www.worldometers.info/
[7]https://data.worldbank.org.cn/
[8]https://unstats.un.org/

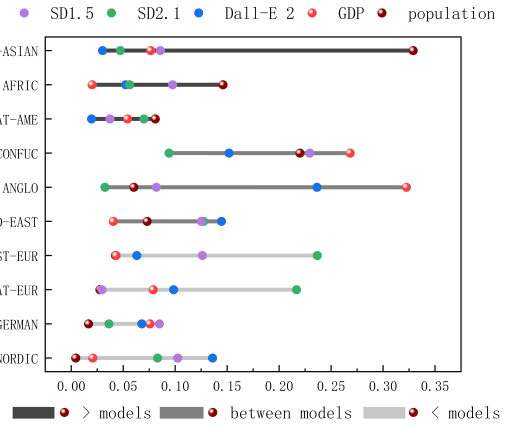
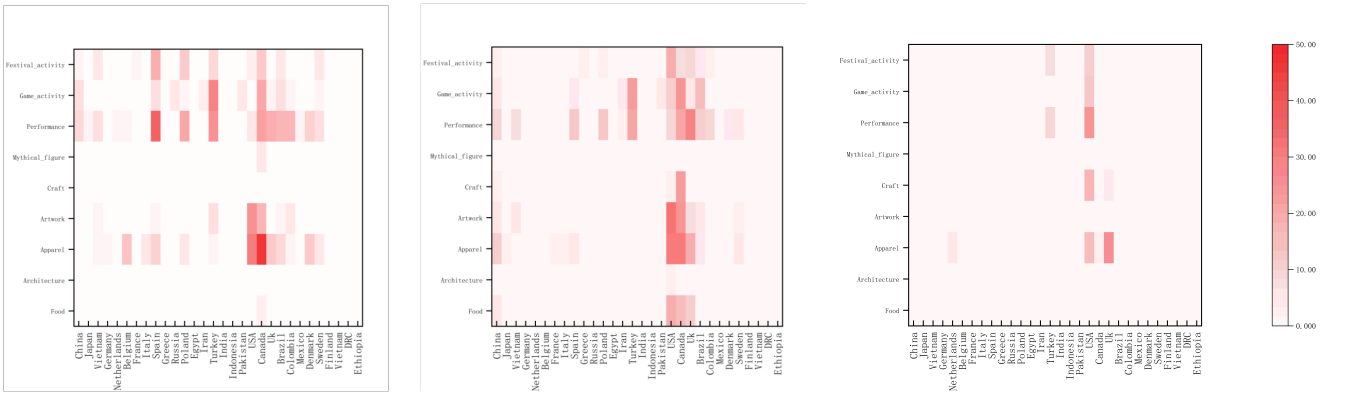



**Figure 4: Categorical share of images generated from neutral prompts for each cluster, and comparison with actual population and GDP share in 2021. Clusters with large populations but lower GDP shares are more likely to be ignored by the model.**

three T2I models, the $p < 0.001$ and Kendall's W is 1[18], which indicates the high agreement of diversity disparity in different countries or cultural clusters. Dall-E v2 has the highest diversity, followed by Stable Diffusion v2.1, and Stable Diffusion v1.5. All three models have the highest diversity in the United States. Table 3 provides the variance of Vendi scores for the 30 countries under each subject. A larger variance indicates greater diversity differences in Vendi scores across different countries within each subject. It can be seen that the Architecture subject exhibited the greatest unevenness in diversity assessments among the countries. To longitudinally compare the diversity disparity of each model across cultural subjects, we visualize the difference between the Vendi scores of each country and its subject's average Vendi score for each model using heat maps (Figure 10 in Appendix F), where the larger the chromatic shift, the greater the fluctuation of the Vendi scores. It can be seen that Dall-E v2 has the largest range of Vendi fluctuation, and Stable Diffusion v2.1 has the smallest, which implies that Dall-E v2 is most influenced by subjects in terms of diversity. Consistent with the previous analysis of variance, some countries' Vendi scores for the same subject vary quite a bit. For example, in three models, Iran's architecture scores low, while Denmark scores high, the USA scores high in performances, and Pakistan scores low (Partial examples of image sets with low Vendi scores, Figure 11 in Appendix F). For mainstream countries, the relevant cultural content is more diverse. Whereas for some less influential countries, the lack of variable data raises the problem of homogeneity, which may further add to stereotypes on them.

## 4.3 Simple Patchwork of Elements

Due to the fact that patchwork is difficult to identify by machines but can be easily recognized by humans, we chose a manual approach to count the number of occurrences of patchwork. Figure 5 shows the number of occurrences of each collocation within



(a) SD1.5 (b) SD2.1 (c) Dall-E 2

**Figure 5: Occurrences of patchwork for the three models. The deeper the color, the more frequently the cultural subject appears in patchwork in this country.**

**Table 3: Variance based on Vendi scores for 30 countries under each subject. All three models have the largest variance over Architecture.**

| | SD1.5 | SD2.1 | Dall-E2 | Mean |
|---|---|---|---|---|
| Apparel | 0.44 | 0.36 | 0.71 | 0.5 |
| Architecture | 1.55 | 0.88 | 1.68 | 1.37 |
| Artwork | 0.34 | 0.58 | 1.18 | 0.7 |
| Handicraft | 0.13 | 0.16 | 1.18 | 0.49 |
| Festival activity | 0.26 | 0.14 | 1.08 | 0.49 |
| Food | 0.15 | 0.19 | 0.66 | 0.34 |
| Game Activity | 0.48 | 0.65 | 1.23 | 0.79 |
| Mythical figure | 0.27 | 0.16 | 0.35 | 0.26 |
| Performance | 0.33 | 0.41 | 0.47 | 0.4 |


**Figure 6: Examples of images that have been determined to appear with patchwork.**

50 images generated by each expanded prompt in the form of a heatmap. Among the three models, apparel and performance have the most serious situation, with the percentage of the occurrences of patchwork being as high as 21.9% and 26.5%, respectively. As can be seen from the color depth of the heatmaps, Canada is the heaviest in Stable Diffusion v1.5 and v2.1, and the US is the heaviest in Dall-E2 v2 for the patchwork situation (e.g., Stable Diffusion v1.5 and Canada, the apparel occurs patchwork 47 times, the game occurs patchwork 20 times, and the performance occurs patchwork 20 times). When the T2I model is unable to understand the subject qualifier together with the region qualifier, it will show a case of patchwork by simply adding some independent elements, such as the national flag, to the subject. Figure 6 illustrates examples of images that are judged to represent a patchwork. The model's lack of learning about certain conjunction features described by more than two words may cause semantic fragmentation, thus generating a patchwork. The patchwork situation reflects the limitations of the current T2I models in terms of the capacity to understand and represent composite concepts.

## 4.4 Fidelity of Cultural Objects

In the previous section, we mentioned the statistically biased problem of the FID score and therefore added the KID score, which performs better in some aspects. However, both FID and KID have the disadvantage of favoring the data trained on ImageNet and are sometimes inconsistent with human perception. Therefore we added qualified human evaluation to solve the deficiency of machine evaluation and chose it as our ultimate criterion for fidelity evaluation.

The lower the FID and KID and the higher the human evaluation score, the more similar the generated images are to the authentic images. Table 4 shows the results of FID, KID, and human evaluation scores calculated in the nine cultural subjects, where FID and KID are in general consistent with each other (FID and KID scores line

graphs, Figure 12(a) and (b), Appendix G). Overall, Stable Diffusion v1.5 gives the greatest performance, followed by Dall-E v2, and Stable Diffusion v2.1. In festival activity, the FID of Stable Diffusion v2.1 suddenly declines dramatically and significantly outperforms the other two models. The apparent inconsistency with KID that occurs here may be caused by the failure of the images in this dataset to conform to the normal distribution assumed by FID. Compared to the FID and KID scores of the dataset consisting of generated cultural objects, it could be found that the result differs from those calculated by the models[57, 59] using general data sets such as MS-COCO[36].

For human evaluation, we assessed annotator agreement using the Intraclass Correlation Coefficient[66]. The results for Stable Diffusion v1.5, Stable Diffusion v2.1, and Dall-E v2 were 0.976, 0.985, and 0.989, respectively, indicating high agreement among the raters. The standard deviations for the three models were 0.322, 0.200, 0.228, respectively. Observing that the results of the human evaluation (Figure 12 (c) in Appendix G), there is a fluctuation in the rank order of the mean scores on each subject among different T2I models. It's hard to tell which model is of the best quality in fidelity, with their mean scores all being very close to 3 ('so-so or ok similarity'). Considering all cultural objects across different subjects as a unified collection, varied score distribution for all objects (Table 8 in Appendix G) can be used to reflect the fidelity performance of each model. We can see that the proportion of cultural objects whose scores are lower than 3 in Stable Diffusion v1.5, Stable Diffusion v2.1, and Dall-E v2 are 56.5%,52.7%,43.8% respectively. It means that compared to our benchmark dataset UCOGC, all T2I models present nearly half of the cases where the generated images of the cultural objects do not match with the authentic images. Observing some of the examples (Figure 13 in Appendix G) whose manual evaluation scores below 2 ('fairly dissimilar'), it can be found that these cultural objects are completely mispresented. When the T2I model does not learn enough about the cultural object, it may choose to generate images of other words that are closer to the cultural object in the latent space of embeddings. For example, understanding the Tang suit (a kind of apparel from China) as "suit" and generating images of suit instead of Tang suit, and understanding Drop (a kind of food from the Netherlands) as "drop" associated with water droplets. Although the above words are close in emending distance, they are in fact completely different items, and if the model understands them incorrectly, it will generate fully inconsistent images, which further amplifies the risk of transmitting misleading information about cultural objects.

# 5 DISCUSSION

In the previous sections, by introducing two attributes, culture cluster and subject, and collecting generated dataset and UCOGC dataset, we investigate the cultural representativeness of T2I models in four dimension: distribution, diversity, patchwork, and fidelity. In this section, we will discuss the following three aspects: causes of the underrepresentation and impacts, dataset release, limitations and future work. These discussions are intended to provide insights and comments to stimulate improvements of T2I models on the performance of cultural representativeness.

## 5.1 Causes of the Underrepresentation and Impacts

Our research reveals four manifestations of cultural underrepresentation in T2I models, and it is worthy to discuss the causes and potential impacts behind them.

The first issue revolves around the overrepresentation and neglect of specific countries and regions, which may be due to an imbalance in the composition of the training data. This bias seems to lead to models favoring economically developed countries, which in turn may further marginalize disadvantaged regions and reduce their presence in the global culture. Models may overemphasize the cultural characteristics of powerful cultural clusters and reduce the presentation of other areas. This is not in line with the principles of diversity and inclusion.

The second issue relates to differences in diversity between T2I models across different cultural clusters and subjects. The reason why such differences occur seems due to the poor and the homogenization of some data especially to disadvantaged cultures. Uneven diversity may lead to disregarding the variety of some cultural subjects in a country, ultimately limiting the richness of cultural representation. The less diversity of some cultural clusters may reinforce stereotypes about themself.

The third issue relates to the problem of patchwork in the generated images, which suggests that the T2I model may encounter challenges when trying to understand and accurately represent specific cultural objects. These limitations seem to stem from inherent difficulties of models in capturing nuanced cultural features precisely, highlighting the current boundaries of their understanding and expressive capabilities.

The last issue centers on the fact that different models have problems generating images that match actual cultural objects. This phenomenon may originate from the fact that T2I models have flaws in the semantic understanding of certain cultural objects. The models may not have enough learning samples of cultural objects or tend to associate certain cultural objects with words that are close but are varied in latent space, resulting in the generation of erroneous images. Erroneous images generated by the model may convey incorrect cultural information, jeopardizing the accurate expression of culture and increasing the potential risk of cultural misconception.

These issues carry the potential for far-reaching consequences, encompassing cultural misunderstandings, unequal cultural representation, the possible reinforcement of stereotypes, and the dissemination of potentially misleading cultural information. In light of these challenges, inspiring improvements in T2I models' cultural representation becomes all the more imperative.

## 5.2 Datasets Release

In our research, we have developed two crucial datasets to evaluate and improve the cultural representativeness of T2I models. The generated dataset is instrumental in assessing the performance of T2I models in cultural responsiveness across cultural clusters and subjects. It allows for an extensive evaluation of model capabilities and provides insights into areas of improvement. As a benchmark dataset, the UCOGC dataset can be used to evaluate and enhance the cultural representativeness of any T2I model, as well as aid

**Table 4: The results of FID, KID, and human evaluation scores calculated in the nine cultural subjects. The lower the FID and KID and the higher the human evaluation score, the more similar the generated images are to the authentic images.**

| | SD1.5 | | | SD2.1 | | | DallE-2 | | |
|---|---|---|---|---|---|---|---|---|---|
| | Human | FID | KID | Human | FID | KID | Human | FID | KID |
| Food | 3.279 | 69.551 | 0.044 | 3.546 | 99.128 | 0.077 | 3.488 | 65.913 | 0.043 |
| Apparel | 2.562 | 68.230 | 0.033 | 2.755 | 80.266 | 0.048 | 2.978 | 74.106 | 0.037 |
| Architecture | 3.063 | 59.842 | 0.034 | 3.189 | 77.455 | 0.050 | 3.510 | 72.404 | 0.046 |
| Craft | 2.951 | 50.933 | 0.011 | 3.631 | 82.148 | 0.040 | 3.488 | 54.487 | 0.017 |
| Artwork | 3.263 | 52.108 | 0.014 | 3.280 | 87.552 | 0.048 | 3.076 | 53.541 | 0.021 |
| Mythical figure | 2.546 | 67.221 | 0.029 | 2.238 | 83.619 | 0.048 | 2.556 | 75.457 | 0.040 |
| Performance | 3.319 | 77.658 | 0.056 | 3.482 | 86.485 | 0.065 | 3.589 | 99.879 | 0.074 |
| Game activity | 2.230 | 63.150 | 0.017 | 2.530 | 82.890 | 0.041 | 2.575 | 78.214 | 0.027 |
| Festival activity | 3.099 | 86.467 | 0.053 | 2.722 | 0.002 | 0.049 | 2.245 | 85.723 | 0.046 |
| Mean | 2.924 | 66.129 | 0.032 | 3.041 | 75.505 | 0.052 | 3.056 | 73.303 | 0.039 |

other multimodal tasks. Access datasets as well as other associated data at https://github.com/Hi-2048/chi.git.

## 5.3 Limitations and Future Work

While we have made significant efforts to compile a diverse and comprehensive dataset for cultural objects, it's important to acknowledge that some subjects in specific regions remain underrepresented due to the inherent diversity of cultures worldwide. For example, for the food subject in India, in addition to choosing Naan, Tandoor Chicken and Biryani, we can find more foods unique to the region. However, more often than not, we cannot even find three representative cultural objects in many countries. For instance, only one or two types of representative architecture of Nigeria can be found. To address this limitation, we will involve collaborating with senior cultural experts from various regions to refine and expand the dataset, ensuring a more balanced representation of cultural objects.

In our pursuit of maintaining evaluation quality, we opted for offline volunteer recruitment facilitated with expert guidance rather than crowdsourcing online. While this approach has its merits, we acknowledge potential concerns regarding representativeness. Therefore, in future research endeavors, we plan to explore the use of internet crowdsourcing to ensure a broader and more representative pool of evaluators while maintaining the involvement of domain experts. This will allow us to strike a balance between expertise and diversity in our human assessments.

KID and FID are common parameters for evaluating T2I models, but their inconsistency with human perception is a common problem because image resizing and compression hardly lower the perceived quality but lead to larger changes in FID scores. When the automatic evaluation is not reliable, human evaluation is our ultimate criterion for fidelity evaluation but it hinders reusability. For future work, developing hybrid approaches that combine both automatic and manual evaluation could potentially be more robust and efficient.

Although our study has primarily examined established models such as DALL-E v2 and Stable Diffusion as exemplars of text-to-image generation, it is important to acknowledge the emergence of new models like Imagen, which hold promise in this field but these models are still not accessible to the public. Consequently, we defer the investigation of these alternative models to future research endeavors. Exploring these newer models will contribute to a more thorough comprehension of the text-to-image generation landscape and its potential for enhancing cultural representation.

# 6 CONCLUSION

Our research delves into the intriguing realm of Text-to-Image generative models and their profound implications for cultural representativeness. While these models have garnered considerable attention for their creative potential, they also inherit biases from their training data, perpetuating cultural misrepresentations. We highlight the significance of addressing cultural biases within T2I models, which have largely been overlooked in previous research. Our investigation unveils three critical findings: first, there exists an evident bias in generated images when a specific country is unspecified, with Latin-American and South-East Asian cultures being particularly neglected; second, certain cultural subjects, such as performance and architecture, tend to exhibit a simplistic patchwork of elements or stereotypical representations; third, a notable portion of featured cultural objects is inaccurately portrayed in Stable Diffusion v1.5 and v2.1, as well as DALL-E v2. As we move forward, it is imperative to ensure that these models offer a more inclusive and nuanced portrayal of cultures worldwide, fostering cross-cultural understanding and appreciation.

# ACKNOWLEDGMENTS

We thank Gang Liu and Yongqin Jiao for their cultural insights to help us select cultural objects and build a standard dataset, crowdsourcing workers for manual evaluations, and the anonymous reviewers for their valuable feedback on the draft. This work was supported in part by the National Natural Science Foundation of China under Grant 62162021 and 62362026; in part by the specific research fund of The Innovation Platform for Academicians of Hainan Province under Grant YSPTZX202314, in part by the Key Project of Hainan Province under Grant ZDYF2023GXJS158.

# REFERENCES


[1] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Large language models associate Muslims with violence. *Nat. Mach. Intell.* 3, 6 (2021), 461–463. https://doi.org/10.1038/s42256-021-00359-2

[2] Jaimeen Ahn and Alice Oh. 2021. Mitigating Language-Dependent Ethnic Bias in BERT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 533–549. https://doi.org/10.18653/v1/2021.emnlp-main.42

[3] Björn Barz and Joachim Denzler. 2021. WikiChurches: A Fine-Grained Dataset of Architectural Styles with Real-World Challenges. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, Joaquin Vanschoren and Sai-Kit Yeung (Eds.). https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/eccbc87e4b5ce2fe28308fd9f2a7baf3-Abstract-round2.html

[4] Miguel Basáñez, Ronald Inglehart, and Alejandro Moreno. 1998. Human values and beliefs: A cross-cultural sourcebook. *Ann Arbor: University of Michigan Press* (1998).

[5] Abhipsa Basu, R. Venkatesh Babu, and Danish Pruthi. 2023. Inspecting the Geographical Representativeness of Images from Text-to-Image Models. *CoRR* abs/2305.11080 (2023). https://doi.org/10.48550/arXiv.2305.11080 arXiv:2305.11080

[6] Shruti Bhargava and David A. Forsyth. 2019. Exposing and Correcting the Gender Bias in Image Captioning Datasets and Models. *CoRR* abs/1912.00578 (2019). arXiv:1912.00578 http://arxiv.org/abs/1912.00578

[7] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2023. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023*. ACM, 1493–1504. https://doi.org/10.1145/3593013.3594095

[8] Mikolaj Binkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. 2018. Demystifying MMD GANs. *CoRR* abs/1801.01401 (2018). arXiv:1801.01401 http://arxiv.org/abs/1801.01401

[9] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (Eds.). 4349–4357. https://proceedings.neurips.cc/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html

[10] Ali Borji. 2022. Pros and cons of GAN evaluation measures: New developments. *Comput. Vis. Image Underst.* 215 (2022), 103329. https://doi.org/10.1016/j.cviu.2021.103329

[11] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

[12] Yang Trista Cao and Hal Daumé III. 2020. Toward Gender-Inclusive Coreference Resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 4568–4595. https://doi.org/10.18653/v1/2020.acl-main.418

[13] Jaemin Cho, Abhay Zala, and Mohit Bansal. 2022. DALL-EVAL: Probing the Reasoning Skills and Social Biases of Text-to-Image Generative Models. *arXiv preprint arXiv:2202.04053* (2022).

[14] Jean-Rémy Conti, Nathan Noiry, Stéphan Clémençon, Vincent Despiegel, and Stéphane Gentric. 2022. Mitigating Gender Bias in Face Recognition using the von Mises-Fisher Mixture Model. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (Eds.). PMLR, 4344–4369. https://proceedings.mlr.press/v162/conti22a.html

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[16] Prafulla Dhariwal and Alexander Quinn Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 8780–8794. https://proceedings.neurips.cc/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *CoRR* abs/2010.11929 (2020). arXiv:2010.11929 https://arxiv.org/abs/2010.11929

[18] Andy P Field. 2005. Kendall's coefficient of concordance. *Encyclopedia of Statistics in Behavioral Science* 2 (2005), 1010–11.

[19] Dan Friedman and Adji Bousso Dieng. 2022. The Vendi Score: A Diversity Evaluation Metric for Machine Learning. *CoRR* abs/2210.02410 (2022). https://doi.org/10.48550/arXiv.2210.02410 arXiv:2210.02410

[20] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. 2022. Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XV (Lecture Notes in Computer Science, Vol. 13675)*, Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer, 89–106. https://doi.org/10.1007/978-3-031-19784-0_6

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity Mappings in Deep Residual Networks. *CoRR* abs/1603.05027 (2016). arXiv:1603.05027 http://arxiv.org/abs/1603.05027

[22] Melissa Heikkilä. 2022. The viral AI avatar app Lensa undressed me—without my consent. *Technology Review. https://www. technologyreview. com/2022/12/12/1064751/the-viral-ai-avatar-app-lensa-undressed-me-without-my-consent* (2022).

[23] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women Also Snowboard: Overcoming Bias in Captioning Models. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part III (Lecture Notes in Computer Science, Vol. 11207)*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer, 793–811. https://doi.org/10.1007/978-3-030-01219-9_47

[24] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 6626–6637. https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html

[26] Geert Hofstede. 1980. Culture's Consequences. Beverly Hills, Calif.

[27] Geert Hofstede. 1984. *Culture's consequences: International differences in work-related values*. Vol. 5. sage.

[28] Robert J House, Paul J Hanges, Mansour Javidan, Peter W Dorfman, and Vipin Gupta. 2004. *Culture, leadership, and organizations: The GLOBE study of 62 societies*. Sage publications.

[29] Wei-Lin Hsiao and Kristen Grauman. 2021. From Culture to Clothing: Discovering the World Events Behind A Century of Fashion Images. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 1046–1055. https://doi.org/10.1109/ICCV48922.2021.00110

[30] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, Seoul, Republic of Korea, April 18-23, 2015*, Bo Begole, Jinwoo Kim, Kori Inkpen, and Woontack Woo (Eds.). ACM, 3819–3828. https://doi.org/10.1145/2702123.2702520

[31] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012).

[32] Alfred Louis Kroeber. 1925. *Handbook of the Indians of California*. Vol. 78. US Government Printing Office.

[33] Simon Kuznets. 1946. *National income*. National Bureau of Economic Research New York.

[34] Shahar Levy, Koren Lazar, and Gabriel Stanovsky. 2021. Collecting a Large-Scale Gender Bias Dataset for Coreference Resolution and Machine Translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 2470–2480. https://doi.org/10.18653/v1/2021.findings-emnlp.211

[35] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. *CoRR* abs/2201.12086 (2022). arXiv:2201.12086

https://arxiv.org/abs/2201.12086

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 740–755.

[37] Zhixuan Liu, Youeun Shin, Beverley-Claire Okogwu, Youngsik Yun, Lia Coleman, Peter Schaldenbrand, Jihie Kim, and Jean Oh. 2023. Towards Equitable Representation in Text-to-Image Synthesis Models with the Cross-Cultural Understanding Benchmark (CCUB) Dataset. *CoRR* abs/2301.12073 (2023). https://doi.org/10.48550/arXiv.2301.12073 arXiv:2301.12073

[38] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2023. Stable Bias: Analyzing Societal Representations in Diffusion Models. *CoRR* abs/2303.11408 (2023). https://doi.org/10.48550/arXiv.2303.11408 arXiv:2303.11408

[39] Thomas Manzini, Yao Chong Lim, Alan W. Black, and Yulia Tsvetkov. 2019. Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 615–621. https://doi.org/10.18653/v1/n19-1062

[40] Nicole Meister, Dora Zhao, Angelina Wang, Vikram V Ramaswamy, Ruth Fong, and Olga Russakovsky. 2022. Gender artifacts in visual datasets. *arXiv preprint arXiv:2206.09191* (2022).

[41] Danaë Metaxa, Michelle A. Gan, Su Goh, Jeff T. Hancock, and James A. Landay. 2021. An Image of Society: Gender and Racial Representation and Impact in Image Search Results for Occupations. *Proc. ACM Hum. Comput. Interact.* 5, CSCW1 (2021), 26:1–26:23. https://doi.org/10.1145/3449100

[42] Margaret Mitchell, Dylan K. Baker, Nyalleng Moorosi, Emily Denton, Ben Hutchinson, Alex Hanna, Timnit Gebru, and Jamie Morgenstern. 2020. Diversity and Inclusion Metrics in Subset Selection. In *AIES '20: AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, February 7-8, 2020*, Annette N. Markham, Julia Powles, Toby Walsh, and Anne L. Washington (Eds.). ACM, 117–123. https://doi.org/10.1145/3375627.3375832

[43] Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 5356–5371. https://doi.org/10.18653/v1/2021.acl-long.416

[44] Ranjita Naik and Besmira Nushi. 2023. Social Biases through the Text-to-Image Generation Lens. *CoRR* abs/2304.06034 (2023). https://doi.org/10.48550/arXiv.2304.06034 arXiv:2304.06034

[45] Tarek Naous, Michael J. Ryan, and Wei Xu. 2023. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models. *CoRR* abs/2305.14456 (2023). https://doi.org/10.48550/arXiv.2305.14456 arXiv:2305.14456

[46] David Amat Olóndriz, Ponç Palau Puigdevall, and Adrià Salvador Palau. 2021. FooDI-ML: a large multi-language dataset of food, drinks and groceries images and descriptions. *CoRR* abs/2110.02035 (2021). arXiv:2110.02035 https://arxiv.org/abs/2110.02035

[47] Edward Ongweso. 2019. Racial Bias in AI Isn't Getting Better and Neither Are Researchers Excuses.

[48] Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. 2023. Toward Verifiable and Reproducible Human Evaluation for Text-to-Image Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 14277–14286. https://doi.org/10.1109/CVPR52729.2023.01372

[49] Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. 2023. Toward Verifiable and Reproducible Human Evaluation for Text-to-Image Generation. *CoRR* abs/2304.01816 (2023). https://doi.org/10.48550/arXiv.2304.01816 arXiv:2304.01816

[50] Sinno Jialin Pan and Qiang Yang. 2010. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* 22, 10 (2010), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

[51] Edoardo Maria Ponti, Goran Glavas, Olga Majewska, Qianchu Liu, Ivan Vulic, and Anna Korhonen. 2020. XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 2362–2376. https://doi.org/10.18653/v1/2020.emnlp-main.185

[52] Benjamin Grant Purzycki, Coren Apicella, Quentin D Atkinson, Emma Cohen, Rita Anne McNamara, Aiyana K Willard, Dimitris Xygalatas, Ara Norenzayan, and Joseph Henrich. 2016. Cross-cultural dataset for the evolution of religion and morality project. *Scientific Data* 3, 1 (2016), 1–12.

[53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. http://proceedings.mlr.press/v139/radford21a.html

[54] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020. Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing. In *AIES '20: AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, February 7-8, 2020*, Annette N. Markham, Julia Powles, Toby Walsh, and Anne L. Washington (Eds.). ACM, 145–151. https://doi.org/10.1145/3375627.3375820

[55] Pranjal Singh Rajput and Shivangi Aneja. 2021. IndoFashion: Apparel Classification for Indian Ethnic Clothes. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 3935–3939. https://doi.org/10.1109/CVPRW53098.2021.00440

[56] Vikram V. Ramaswamy, Sunnie S. Y. Kim, and Olga Russakovsky. 2021. Fair Attribute Classification Through Latent Space De-Biasing. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 9301–9310. https://doi.org/10.1109/CVPR46437.2021.00918

[57] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *CoRR* abs/2204.06125 (2022). https://doi.org/10.48550/arXiv.2204.06125 arXiv:2204.06125

[58] Peter J Reynolds. 1979. Iron-age farm: the Butser experiment. *(No Title)* (1979).

[59] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 10674–10685. https://doi.org/10.1109/CVPR52688.2022.01042

[60] Candace Ross, Boris Katz, and Andrei Barbu. 2021. Measuring Social Biases in Grounded Vision and Language Embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, 998–1008. https://doi.org/10.18653/v1/2021.naacl-main.78

[61] Michael Roy. 2017. Cathy O'Neil. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York: Crown Publishers, 2016. 272p. Hardcover, $26 (ISBN 978-0553418811). *College & Research Libraries* 78, 3 (2017), 403. https://doi.org/10.5860/crl.78.3.403

[62] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *NeurIPS*. http://papers.nips.cc/paper_files/paper/2022/hash/ec795aeadae0b7d230fa35cbaf04c041-Abstract-Conference.html

[63] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS*. http://papers.nips.cc/paper_files/paper/2022/hash/a1859debfb3b59d094f3504d5ebb6c25-Abstract-Datasets_and_Benchmarks.html

[64] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. *CoRR* abs/2111.02114 (2021). arXiv:2111.02114 https://arxiv.org/abs/2111.02114

[65] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. 2017. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536* (2017).

[66] Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin* 86, 2 (1979), 420.

[67] Thomas Smits and Melvin Wevers. 2023. A multimodal turn in Digital Humanities. Using contrastive machine learning models to explore, enrich, and analyze digital visual historical collections. *Digital Scholarship in the Humanities* 38, 3 (03 2023), 1267–1280. https://doi.org/10.1093/llc/fqad008 arXiv:https://academic.oup.com/dsh/article-pdf/38/3/1267/51309490/fqad008.pdf

[68] George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Leigh Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, J. Eric T. Taylor, and Gabriel Loaiza-Ganem. 2023. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *CoRR* abs/2306.04675

(2023). https://doi.org/10.48550/arXiv.2306.04675 arXiv:2306.04675

[69] Yolande A. A. Strengers, Lizhen Qu, Qiongkai Xu, and Jarrod Knibbe. 2020. Adhering, Steering, and Queering: Treatment of Gender in Natural Language Generation. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, Regina Bernhaupt, Florian 'Floyd' Mueller, David Verweij, Josh Andres, Joanna McGrenere, Andy Cockburn, Ignacio Avellino, Alix Goguey, Pernille Bjøn, Shengdong Zhao, Briane Paul Samson, and Rafal Kocielnik (Eds.). ACM, 1–14. https://doi.org/10.1145/3313831.3376315

[70] Lukas Struppek, Dominik Hintersdorf, Felix Friedrich, Manuel Brack, Patrick Schramowski, and Kristian Kersting. 2023. Exploiting Cultural Biases via Homoglyphs in Text-to-Image Synthesis. arXiv:2209.08891 [cs.CV]

[71] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth M. Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 1630–1640. https://doi.org/10.18653/v1/p19-1159

[72] Alexander Toet, Daisuke Kaneko, Inge De Kruijf, Shota Ushiama, Martin G Van Schaik, Anne-Marie Brouwer, Victor Kallen, and Jan BF Van Erp. 2019. CROCUFID: A cross-cultural food image database for research on food elicited affective responses. *Frontiers in psychology* 10 (2019), 58.

[73] Schrasing Tong and Lalana Kagal. 2020. Investigating Bias in Image Classification using Model Explanations. *CoRR* abs/2012.05463 (2020). arXiv:2012.05463 https://arxiv.org/abs/2012.05463

[74] Edward Burnett Tylor. 1871. *Primitive culture: Researches into the development of mythology, philosophy, religion, art and custom*. Vol. 2. J. Murray.

[75] Eddie L. Ungless, Björn Ross, and Anne Lauscher. 2023. Stereotypes and Smut: The (Mis)representation of Non-cisgender Identities by Text-to-Image Models. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, 7919–7942. https://doi.org/10.18653/v1/2023.findings-acl.502

[76] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural Discrete Representation Learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 6306–6315. https://proceedings.neurips.cc/paper/2017/hash/7a98af17e63a0ac09ce2e96d03992fbc-Abstract.html

[77] Jialu Wang, Yang Liu, and Xin Eric Wang. 2021. Are Gender-Neutral Queries Really Gender-Neutral? Mitigating Gender Bias in Image Search. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 1995–2008. https://doi.org/10.18653/v1/2021.emnlp-main.151

[78] Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns. *Trans. Assoc. Comput. Linguistics* 6 (2018), 605–617. https://doi.org/10.1162/tacl_a_00240

[79] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. 2022. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. *Trans. Mach. Learn. Res.* 2022 (2022). https://openreview.net/forum?id=AFDcYJKhND

[80] Miao Zhang and Rumi Chunara. 2022. Fair contrastive pre-training for geographic images. *CoRR* abs/2211.08672 (2022). https://doi.org/10.48550/arXiv.2211.08672 arXiv:2211.08672

[81] Yanzhe Zhang, Lu Jiang, Greg Turk, and Diyi Yang. 2023. Auditing Gender Presentation Differences in Text-to-Image Models. *CoRR* abs/2302.03675 (2023). https://doi.org/10.48550/arXiv.2302.03675 arXiv:2302.03675

[82] Yi Zhang and Jitao Sang. 2020. Towards Accuracy-Fairness Paradox: Adversarial Example-based Data Augmentation for Visual Debiasing. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann (Eds.). ACM, 4346–4354. https://doi.org/10.1145/3394171.3413772

[83] Dora Zhao, Angelina Wang, and Olga Russakovsky. 2021. Understanding and Evaluating Racial Biases in Image Captioning. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 14810–14820. https://doi.org/10.1109/ICCV48922.2021.01456

[84] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, 2979–2989. https://doi.org/10.18653/v1/d17-1323

# A SELECTED REPRESENTATIVE COUNTRIES IN EACH CULTURAL CLUSTER

**Table 5: List of the countries we chose to represent their cultural clusters. Three countries with the largest populations are primarily used to represent their cultural clusters, with the only exceptions being Latin America and Eastern Europe.**

| Cultural Cluster | Country |
|---|---|
| CONFUCIAN | China |
| | Japan |
| | Vietnam |
| GERMAN | Germany |
| | Netherlands |
| | Belgium |
| LATIN-EUROPEAN | France |
| | Italy |
| | Spain |
| EASTERN EUROPEAN | Greece |
| | Russia |
| | Poland |
| MIDDLE EASTERN | Egypt |
| | Iran |
| | Turkey |
| SOUTH-EAST ASIAN | India |
| | Indonesia |
| | Pakistan |
| ANGLO-SAXON | the United States of America |
| | Canada |
| | the United Kingdom |
| LATIN-AMERICAN | Brazil |
| | Columbia |
| | Mexico |
| NORDIC | Denmark |
| | Switzerland |
| | Finland |
| AFRICAN | Nigeria |
| | the Democratic Republic of the Congo |
| | Ethiopia |

## B SELECTED CULTURAL OBJECTS AND THE MISPRESENTED ONES

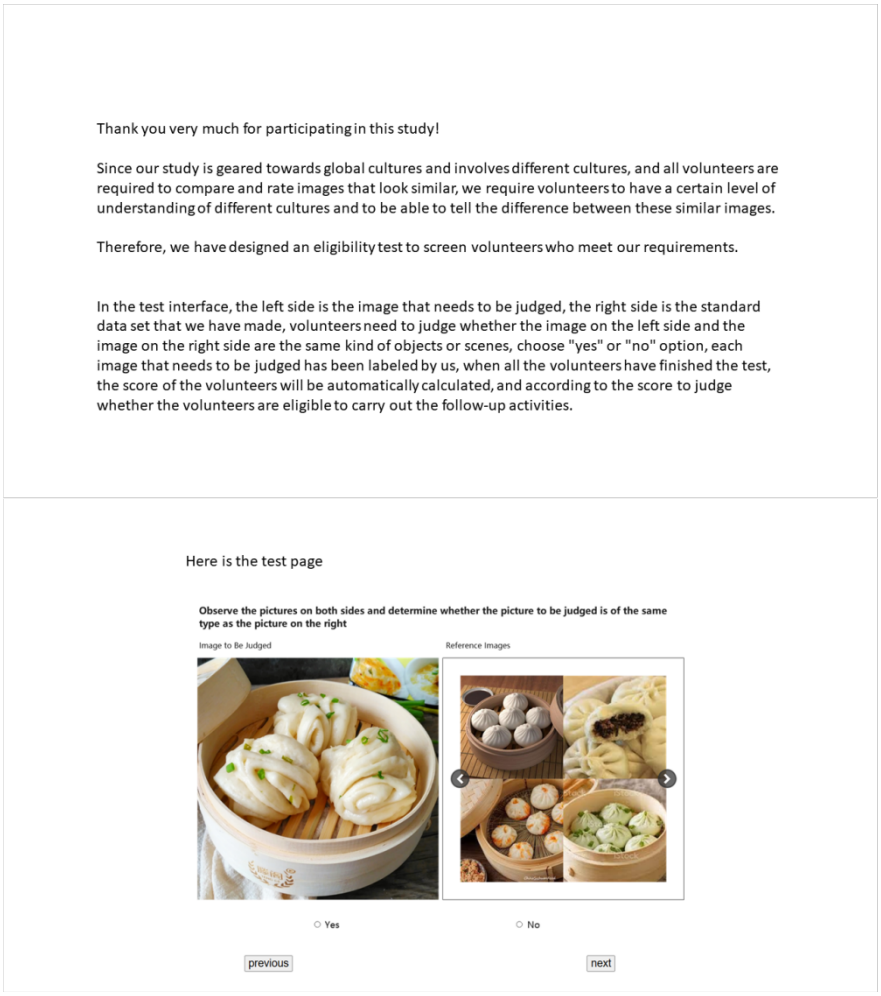
**Table 6: The 752 cultural objects across 30 representative countries in ten cultural clusters with nine material and nonmaterial cultural subjects were identified and verified by experts. We labeled the low-quality cultural objects that rated less than 3 ('so-so or ok similarity') by the qualified human evaluation in at least two of the three T2I models with '*'. For complete cultural objects see our supplementary materials or data release website.**

| Country | Food | Architecture | Apparel | |
|---|---|---|---|---|
| China | Mantou; Baozi; Shao Bing* | Siheyuan*; The Great Wall; Tulou | Hanfu; Tang Suit*; Cheongsam | |
| Japan | Sushi; Onigiri (Rice Ball); Tempura | Torii; Tenshu*; Five-storied Pagoda | Kimono; Yukata; Hakama; | |
| Vietnam | Bánh mì; pho bò; Pork and shrimp rolls-Goi cuon | Tháp Chăm*; Đình*; Long Son* | Áo Dài; Áo tu thân*; Nón lá* | |
| Germany | German sausage; Black Forest cake; Sauerkraut mit Schweinshaxe | Black Forest House*; Speicherstadt; Fachwerkhaus* | Dirndl; Lederhosen; Trachten | |
| Netherlands | Haring'Hollandse Nieuwe*; Pannenkoeken; Drop* | windmill; Hofjes*; Dutch Brick Architecture | Kraplap*; Klomp*; Poffer* | |
| Belgium | Stoemp*; Waffles; Couques de Dinant* | Belfry of Bruges; Cathedral of Our Lady, Antwerp; Grand Place/Grote Markt* | | |
| France | Baguette; Foie Gras; Macaron | Norman Timber Framing*; Alsatian Timber Houses; Parisian High-Ceiling Apartments | Breton costume*; Basque Costume*; Can-can dress | |
| Italy | Pizza; Spaghetti; Lasagna | Roman Colosseum; Cinque Terre; Venetian Bell Towers | Calabrian Costume*; Costume Veneziano; Sardinian Costume | |
| Spain | Jamón ibérico; Gazpacho; Empanada | Arabesque Arch; Spanish Plaza; Andalusian Patio | Traje de flamenca*; Manila shawl; Traje de luces* | |
| Greece | Soutzoukakia*; Tomatokeftedes; Bougatsa | Greek Temples; Olympia Archaeological Site; Whitewashed Houses* | Thracian clothing*; Amalía dress*; Evzonas uniform* | ..... |
| Russia | хлеб*; Lobster Bisque; Tula Gingerbread | Russian Orthodox Churches; Khrushchyovka*; Russian Wooden Houses | Kosovorotka*; Sarafan*; Kokoshnik | |
| Poland | Kielbasa; Golabki*; Zupa czosnkowa* | Zakopane Style Architecture | Kontusz*; Folk costumes of Podhale | |

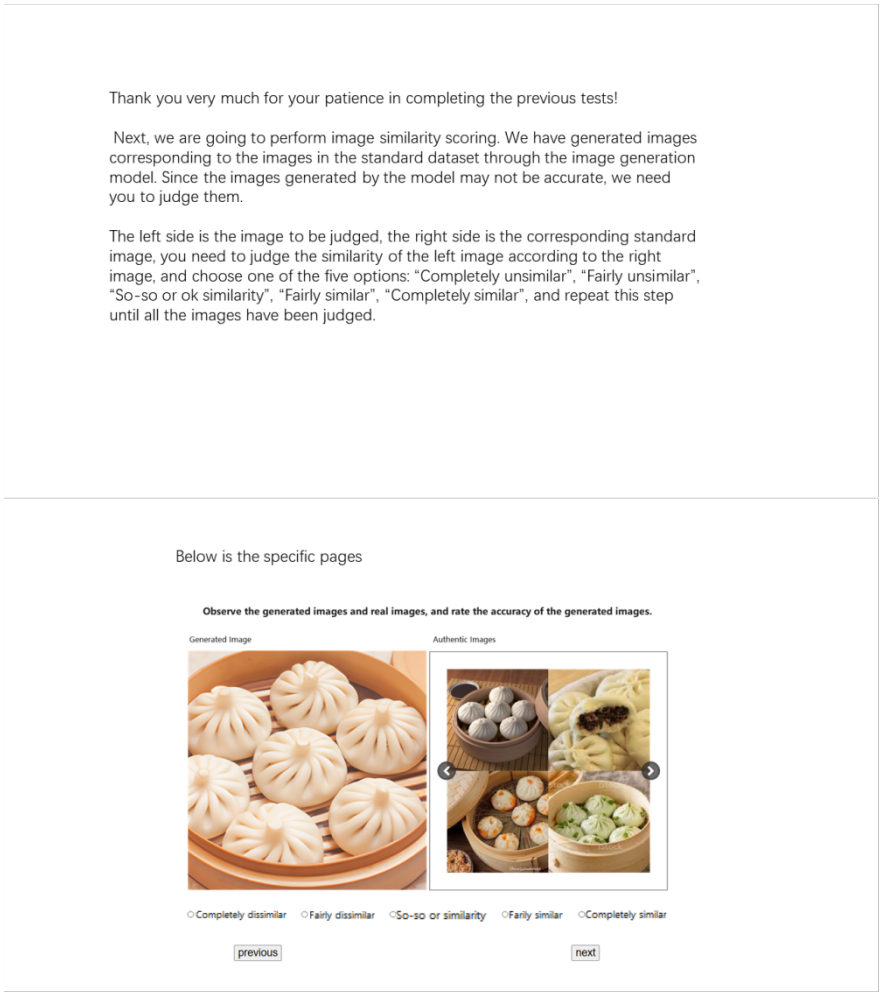
## C CONFIGURATION OF TRANSFER LEARNING

Both ResNet50V2 and ViT-B/16 transferring architecture were extended by appending layers, including flattening, batch normalization, and multiple densely connected layers with GELU activation functions. The model was finalized with a softmax activation layer comprising 'n_classes' output units. The learning rate was set to 1e-4, and optimization was carried out using the Rectified Adam optimizer from TensorFlow Addons. The training was conducted with batch sizes of 32 samples each. The model underwent a maximum of 100 training epochs. The complete code is available on our Supplementary Materials or Data Release website.

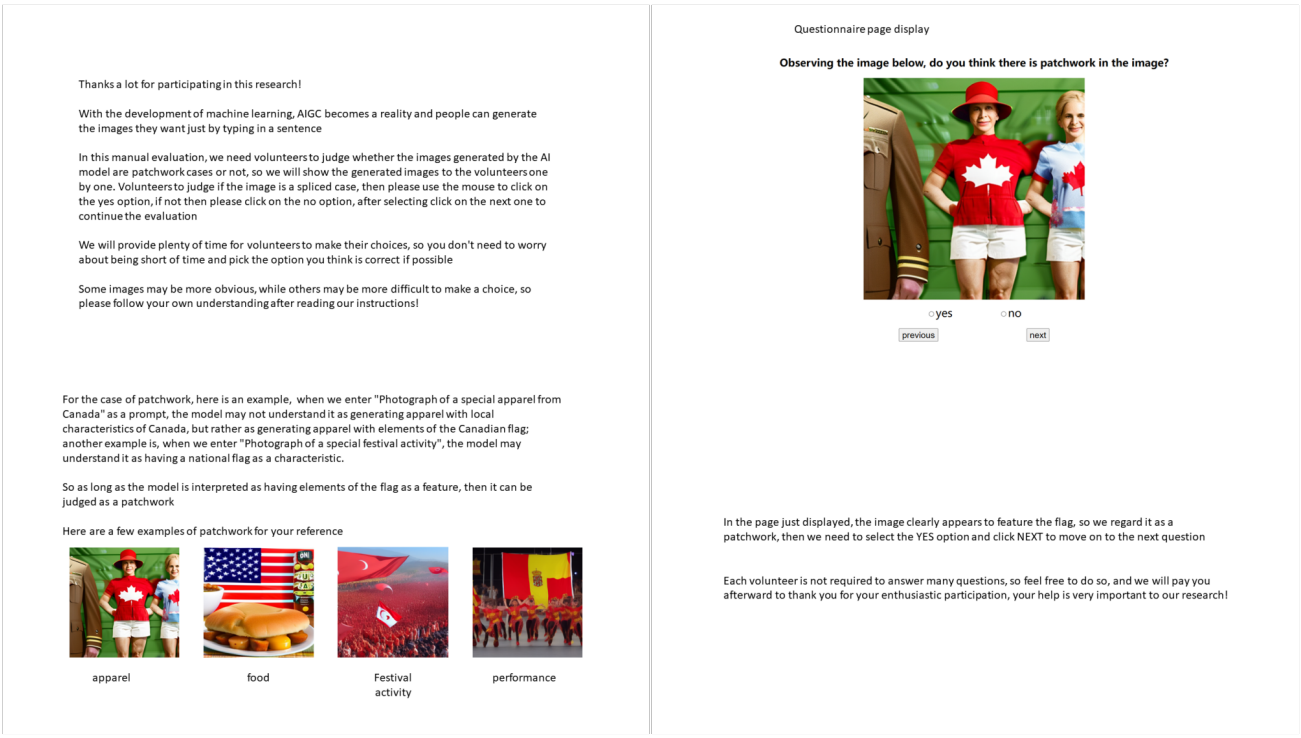
## D QUESTIONNAIRES




**(a) Example of qualification tests page. For each cultural object, ten images were randomly displayed and rated. Each cultural object was evaluated by three individuals. The evaluation process can refer to the standard dataset at any time.**



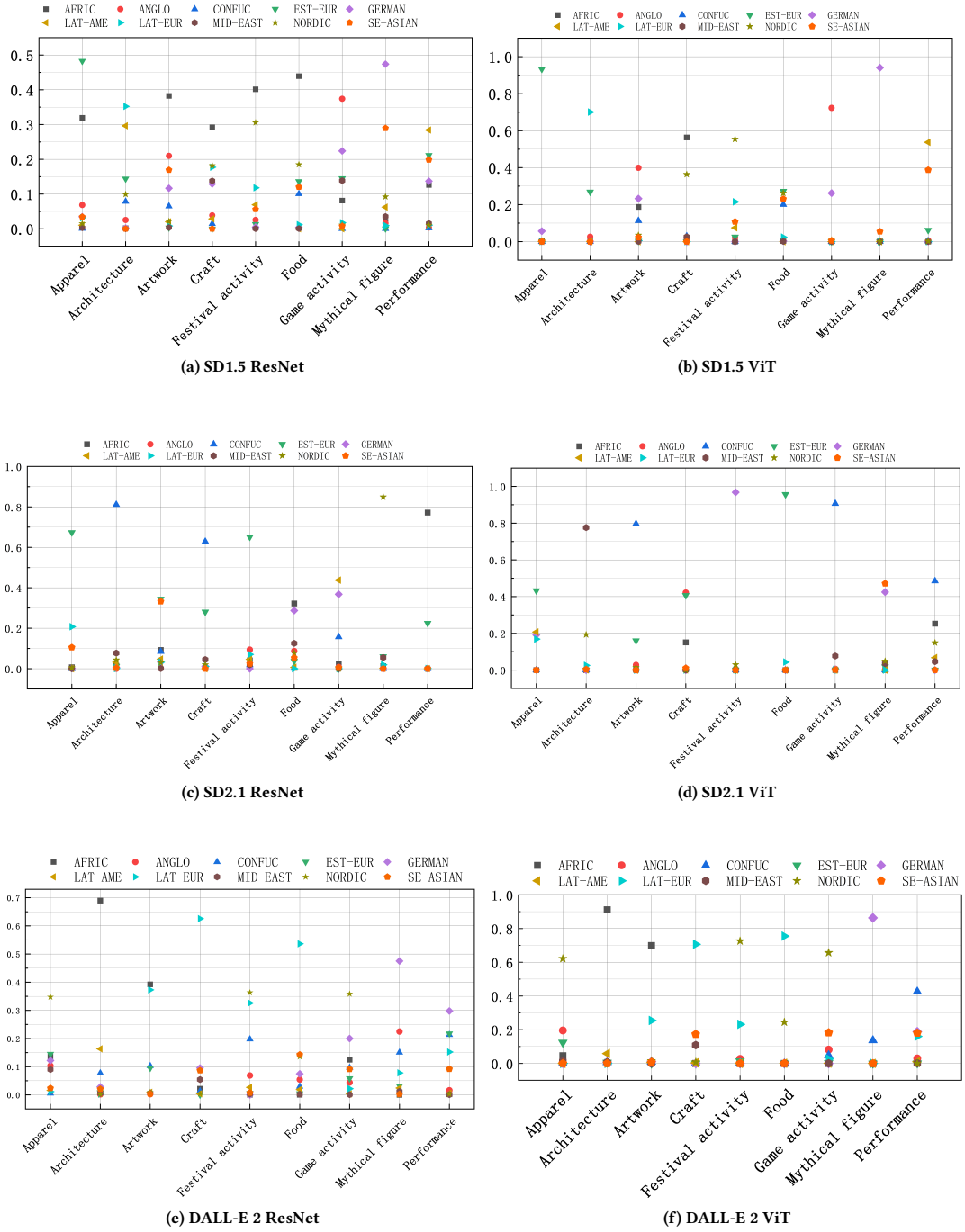

**(b) Example of similarity scoring page. All images generated from expanded prompts are assessed three times, and if two or more people had the sense of occurrence of the patchwork, it is determined that there is a patchwork.**




**(c) Example of patchwork evaluation page. Each image was evaluated by three people.**

**Figure 7: The human assessment page designed for this study.**

# E DISTRIBUTION OF NEUTRAL GENERATION




(a) SD1.5 ResNet

(b) SD1.5 ViT

(c) SD2.1 ResNet

(d) SD2.1 ViT

(e) DALL-E 2 ResNet

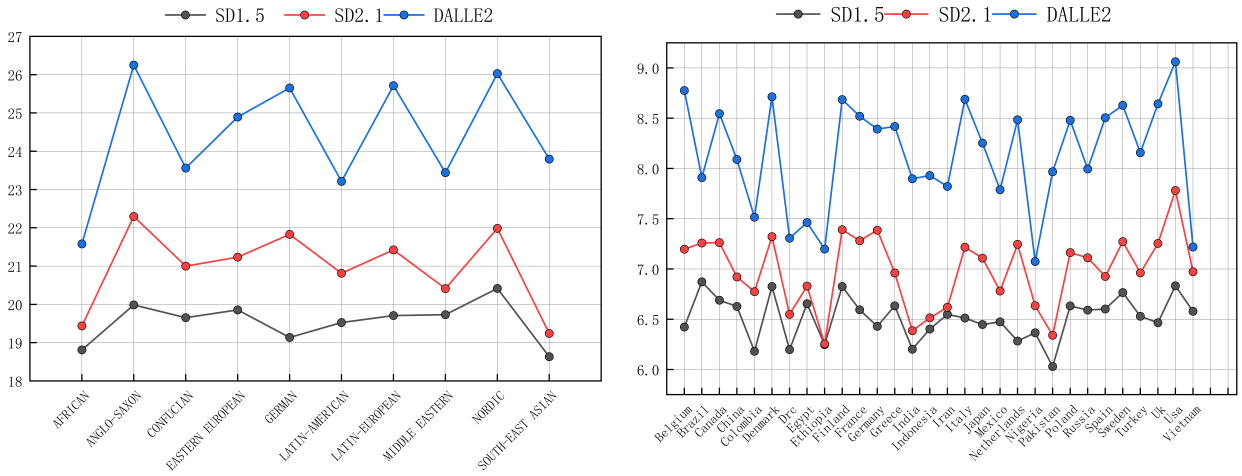
(f) DALL-E 2 ViT

**Figure 8: The distribution probability of the cultural subject in 30 countries. Taking the generated images of expanded prompts for each subject as the training data, the images generated with neutral prompts for each subject are then classified with the corresponding Resnet or ViT transferring models.**

**Table 7: The distribution of the population share, GDP share, and generated distribution on three T2I models in the global cultural clusters. The total share of images generated by three T2I models across cultural subjects shows a neglect of some cultural clusters with high population shares but relatively low GDP shares.**

| | Population | GDP | Dall-E 2 | SD2.1 | SD1.5 |
|---|---|---|---|---|---|
| CONFUC | 22.00% | 26.83% | 15.18% | 9.41% | 22.94% |
| GERMAN | 1.66% | 7.56% | 6.79% | 3.63% | 8.49% |
| LAT-EUR | 2.75% | 7.88% | 9.85% | 21.67% | 2.96% |
| EST-EUR | 4.25% | 4.31% | 6.31% | 23.67% | 12.62% |
| MID-EAST | 7.30% | 4.04% | 14.45% | 12.75% | 12.52% |
| SE-ASIAN | 32.90% | 7.65% | 3.01% | 4.72% | 8.57% |
| ANGLO | 6.02% | 32.24% | 23.64% | 3.24% | 8.19% |
| LAT-AME | 8.08% | 5.41% | 1.96% | 7.00% | 3.72% |
| NORDIC | 0.43% | 2.07% | 13.59% | 8.30% | 10.25% |
| AFRIC | 14.60% | 2.01% | 5.23% | 5.62% | 9.74% |

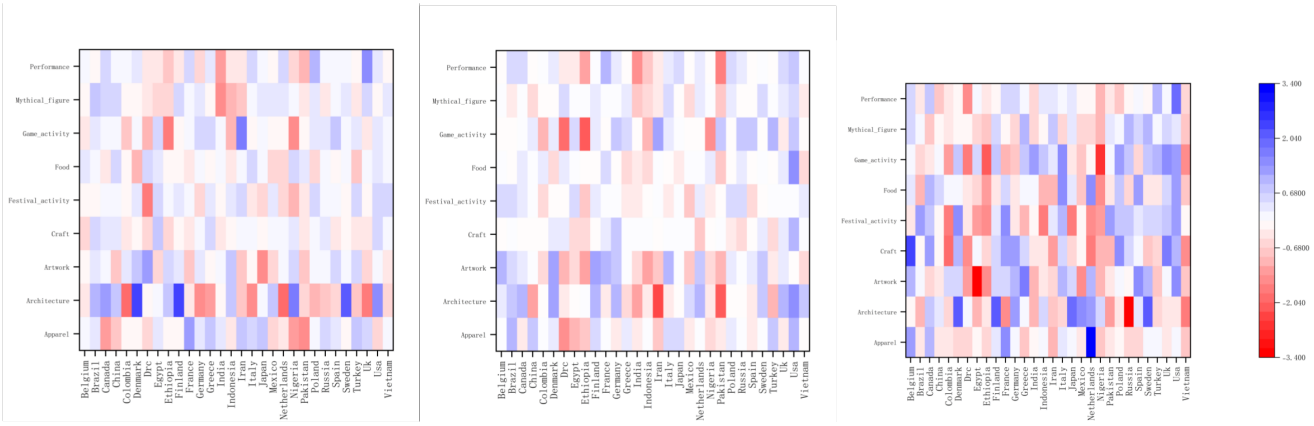
# F DIVERSITY OF EXPANDED GENERATION




(a) Vendi scores for each cultural cluster

(b) Vendi scores for each country

**Figure 9: The line graphs of the average Vendi scores of countries and clusters in each cultural subject.**

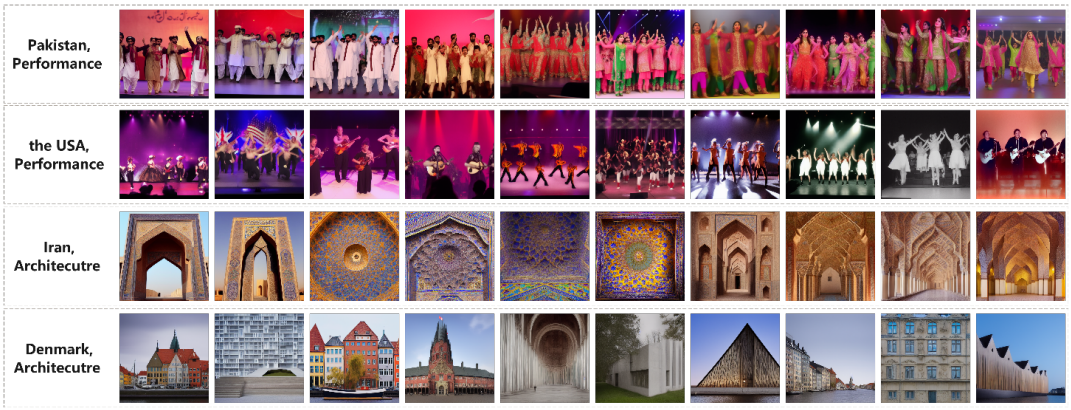



(a) SD1.5

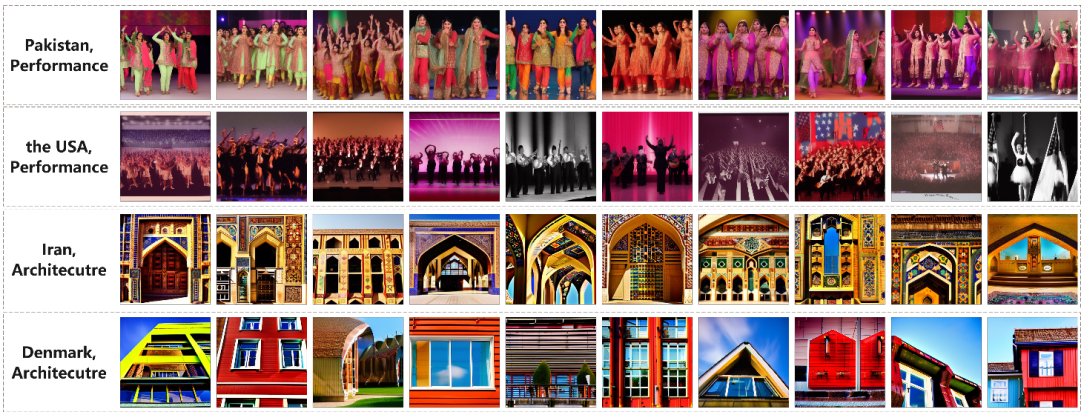
(b) SD2.1

(c) DALL-E 2

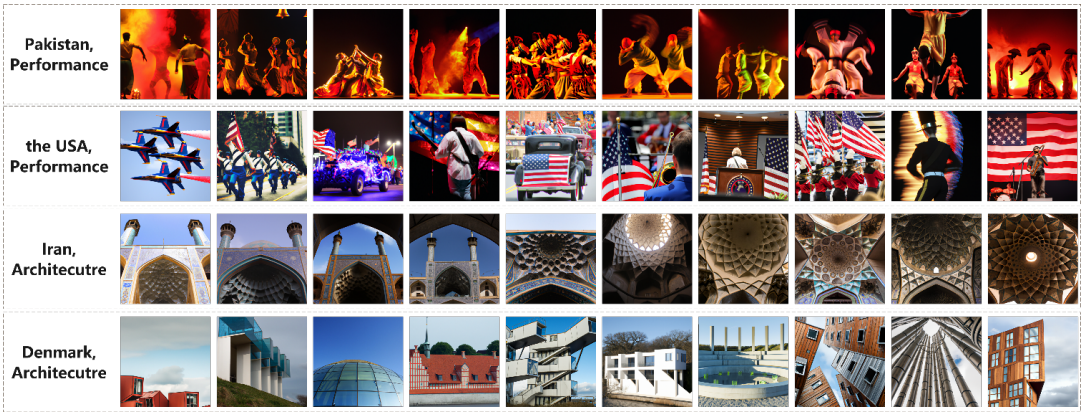
**Figure 10: Heat maps that visualize the difference between the Vendi scores of each country and its subject's average Vendi score for each model. The larger the chromatic shift, the greater the fluctuation of the Vendi scores.**




(a) SD1.5

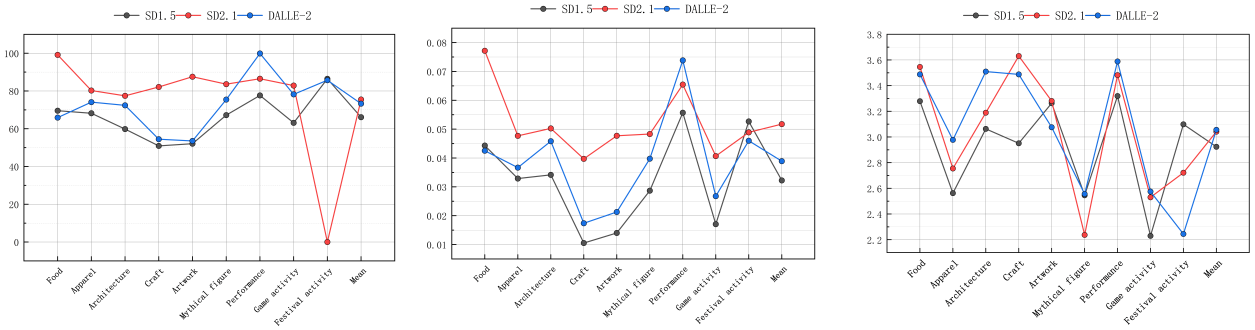



(b) SD2.1




(c) DALL-E 2

# G FIDELITY OF CULTURAL OBJECTS

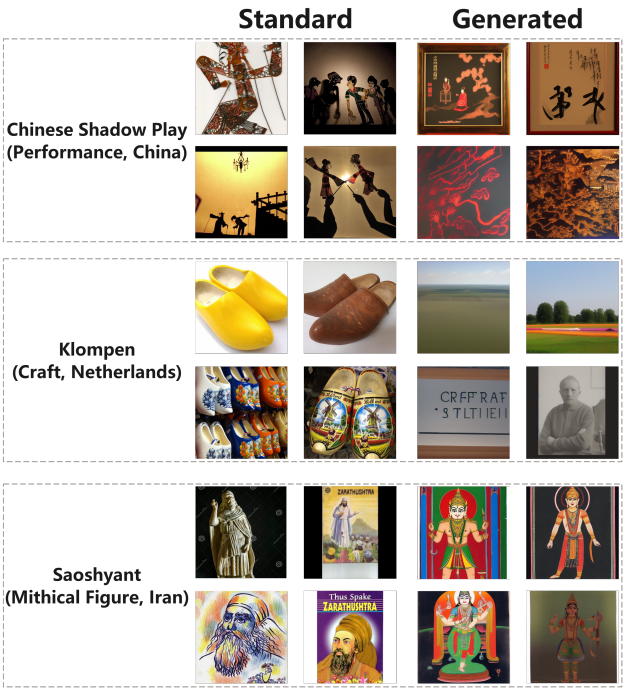



(a) FID (b) KID (c) Human

**Figure 11: FID, KID, manual scores for the three models SD1.5, SD2.1, and Dall-E 2**

**Table 8: The percentage of manual score**

| | 1-2 | 2-3 | 3-4 | 4-5 |
|---|---|---|---|---|
| SD1.5 | 30.05% | 19.41% | 18.62% | 31.91% |
| SD2.1 | 20.48% | 22.87% | 19.15% | 37.50% |
| Dall-E 2 | 28.13% | 15.60% | 16.67% | 39.60% |




(a) SD1.5


(b) SD2.1


(c) DALL-E 2

**Figure 12: Examples of low quality images**