Crossroads of Continents: Automated Artifact Extraction for Cultural Adaptation with Large Multimodal Models

Anjishnu Mukherjee   Ziwei Zhu   Antonios Anastasopoulos
Department of Computer Science, George Mason University
{amukher6,zzhu20,antonis}@gmu.edu

Abstract

In this work, we present a comprehensive three-phase study to examine (1) the effectiveness of large multimodal models (LMMs) in recognizing cultural contexts; (2) the accuracy of their representations of diverse cultures; and (3) their ability to adapt content across cultural boundaries. We first introduce DALLE STREET, a large-scale dataset generated by DALL-E 3 and validated by humans, containing $9,935$ images of 67 countries and 10 concept classes. We reveal disparities in cultural understanding at the sub-region level with both open-weight (LLaVA) and closed-source (GPT-4V) models on DALLE STREET and other existing benchmarks. Next, we assess models' deeper culture understanding by an artifact extraction task and identify over $18,000$ artifacts associated with different countries. Finally, we propose a highly composable pipeline, CULTUREADAPT, to adapt images from culture to culture. Our findings reveal a nuanced picture of the cultural competence of LMMs, highlighting the need to develop culture-aware systems.[1]

1 Introduction

Culture is hard to define and has always been so. Kroeber (1952) explored how the word has evolved to gain different meanings in different contexts. Recent efforts in natural language processing research have seen growing interest in understanding how culture influences language models and human behavior, including language, art, and decision-making (Hershcovich et al., 2022; Adilazuarda et al., 2024; Liu et al., 2024; Ge et al., 2024). As large multimodal models (LMMs; Liu et al., 2023a) intersect more with human life, the need for them to comprehend and respect cultural nuances is crucial. Research in this area explores model alignment with human values, measures of

Figure 1: We introduce a large scale balanced dataset for measuring cultural awareness, a task for extracting implicit cultural artifacts from images and a highly composable pipeline for adapting images to different cultural contexts with fine-grained edits.

cultural awareness, and methods for *cultural adaptation*, where content that may be considered as accurately representing a culture or a country, maybe stereotypically, is modified to represent a different one.

The challenge of assessing the capability of understanding and leveraging cultural knowledge in LMMs, given their multimodal nature – processing both text and image data – is significant. Prior research has primarily investigated LMMs for cultural awareness[2] by examining their performance on culture-related tasks such as region classification from images (Basu et al., 2023; Yin et al., 2023; Pouget et al., 2024), image-caption matching (Liu et al., 2021), and cultural image captioning (Cao et al., 2024). However, these tasks may not fully determine whether LMMs are responding to cultural cues encoded within their training data or merely identifying superficial culture-associated features. Furthermore, there is a lack of in-depth analysis

---

[1]Dataset and code are available: https://github.com/iamshnoo/crossroads

[2]We maintain that LMMs do not inherently possess human values but that their outputs may display cultural knowledge.

of how LMMs utilize cultural knowledge during these tasks and whether they exhibit consistent understanding across various cultural contexts.

To address these gaps, We propose a comprehensive three-phase study to deepen our understanding of cultural awareness in LMMs. First, we develop a large-scale dataset for assessing cultural awareness as measured by the ability of LMMs to recognize and differentiate between cultures, with countries as proxies, in a task setting similar to GeoGuessr (Geoguessr, 2024). Next, we delve deeper into cultural understanding by introducing an artifact extraction task designed to identify implicit cultural artifacts that LMMs use to distinguish between cultures. Finally, based on the insights gained from the artifact extraction task, we propose a cultural adaptation task combining multiple generative models in an end-to-end pipeline to adapt images from one cultural context to another. This pipeline enhances our understanding of LMMs' adaptability and promises an effective data augmentation technique to improve cultural awareness in LMMs. Our main contributions are as follows:

- We introduce, DALLE STREET, a collection of 9, 935 images generated by DALL-E 3, covering 67 countries and 10 cultural concept classes. Compared to datasets like Dollar Street (Rojas et al., 2022), DALLE STREET includes more images from underrepresented geographic regions. We validate the appropriateness of country representations with a human study and will release this dataset publicly.

- We measure how well both open-weight and closed-source LMMs perform on DALLE STREET based on the cultural knowledge encoded in them to understand disparities in performance at the geographic sub-region level for a diverse group of concepts and countries. We include a human study baseline to understand where current models may perform better than humans. We also compare LMM performance on two image datasets collected from different sources.

- We introduce a task for cultural artifact extraction to identify implicit associations. We further develop a method to filter these effectively to discover interesting associations that frequently co-occur for each country.

- We propose a highly composable end-to-end pipeline, CULTUREADAPT, using a chain of LMMs to identify a country for a given image, extract culturally relevant artifacts, ground them

with bounding boxes, and use diffusion-based inpainting to adapt to a target culture. We show examples of this pipeline and include a CLIPScore-based metric to validate its performance.

2 Data

We use three different datasets covering a wide variety of cultural concepts, economic ranges, and different sources: (a) DALLE STREET: synthetic, DALL-E 3 (OpenAI, 2024) generated; (b) Dollar Street: natural, collected photographs; and (c) MaRVL: scraped from the internet under native speaker guidance. We include count statistics for each of the datasets in Table C.2. Overall, our study uses roughly 20k images across 3 datasets, covering 19 geographical sub-regions and 67 countries from the United Nations geoscheme (UN, 2024). Each image is accompanied by a true label for country.

DALLE STREET We consider the same 10 categories and 19 geographical regions as in Dollar Street and include 67 countries.[3] Given these parameters, we generate high resolution images of size 1024x1024 from DALL-E 3 (OpenAI, 2024) for 2 settings - vivid (for generating hyper-real and dramatic images) and natural (for more realistic images). To do so, we follow a templatic prompting approach (prompt shown in Appendix Figure 9) and then sample at least 10 images for each combination of country and concept class. However, due to global content policy filters imposed on the API calls and limited budget, we end up with 9, 935 images (as opposed to the expected 13, 400). We will release this generated dataset under CC BY-SA 4.0.

We perform a qualitative study, where we ask annotators to determine the "appropriateness" of the generated image on a Likert scale from 1 (Strongly Disagree) to 5 (Strongly Agree), where we define appropriateness as the case where the image shown to the annotator is one of the (possibly stereotypical) representations for the given combination of country and category.

Dollar Street (Rojas et al., 2022) This is a dataset of photos of objects and scenes collected by professional and volunteer photographers which specifically focuses on poor and remote environments, containing approximately 200 concept

---

[3]The four extra countries corresponding to new additions to our annotator pool over the timeline of this project.

classes and covering 63 countries. We filter it for images that do not contain multiple labels, classes which do not cover images for all regions, and classes with subjective naming. Then we refer to our group of annotators from diverse backgrounds to choose the top 10 categories by the method of collaborative labeling (Chang et al., 2017), where effectively we simplify our selection of object classes by choosing the ones which all annotators universally agree on as being a relevant dimension for testing cultural awareness. The final data subset that we end up with thus includes 63 countries, 19 geographical regions, $4,137$ images, 10 concept classes (car, cups, mugs and glasses, family snapshots, front door, home, kitchen, plate of food, social drink, wall decoration, and wardrobe). This dataset is available under CC BY-SA 4.0.

MaRVL (Liu et al., 2021) The primary task for this dataset involves validation of statements about image pairs for five languages: Indonesian, Mandarin Chinese, Swahili, Tamil, and Turkish. We use the images and assign country and region labels to create our task setting for cultural awareness.[4] This dataset gives us $4,914$ images curated by native speakers of each of these languages, across 5 distinct geographical regions and it is distributed under CC BY 4.0 license for research purposes.

3 Cultural Awareness (Task 1)

To measure cultural awareness of LMMs, we compare performance on two existing culturally diverse benchmarks and also on DALLE STREET. Later on, we also compare this with the baseline from our human study to understand where LMMs stand at this task in comparison to humans. Overall, we find performance varies across sub-regions but both LLaVA and GPT-4V perform nearly equally well.

3.1 Methods

Given an input image, we prompt two LMMs, LLaVA-NeXT (Liu et al., 2023a), an open-weight model, and GPT-4V vision-preview (OpenAI et al., 2023), a closed-source model, in a zero-shot setting by asking an open-ended question without any possible answer choices: predict the geographical region represented in the image, as per the United Nations geoscheme. We use this geoscheme for three reasons. First, the refusal rate of models is a lot higher when asked about specific country labels. Second, the geoscheme information is

---
[4]MaRVL language to region mapping in Appendix C.1

	GPT-4V	LLaVA
DOLLAR STREET	36.28	**36.83**
DALLE STREET	56.31	**78.05**
MARVL	**41.59**	19.14

Table 1: Overall, LLaVA performs as good as or better than GPT-4V on two out of three datasets. Human accuracy on a subset of DALLE STREET is 42.63.

already present on English Wikipedia, which is usually included in pre-training corpora for most LLMs. And third, sticking to a geoscheme allows us to parse open-ended generations in a structured format. A potential challenge with this approach is that models may not incorporate the most recent changes in the geoscheme, leading to finer grained errors which would be hard to trace (for example, the Wikipedia page for the geoscheme has been updated more than 20 times since the cutoff date for the training data of vision-preview, including some changes to island groups like Channel Islands). We focus on countries which have traditionally been consistently classified without changes to geographic region to avoid these inconsistencies.

Metric: Accuracy We process generated text to map it to one of the geographical sub-regions or to a policy violation case, and then compare with true labels by mapping country information to geographical regions, which gives us classification accuracy as a quantitative metric for measuring success. Specifically, we inspect the confusion matrix to better understand on which regions the models make mistakes.

Economic disparities For Dollar Street data, in addition to the ⟨image,country⟩ information, we also have data available for the monthly income of the family corresponding to each image. We use this information to understand differences in performance across economic groups by looking at normalized income quartiles for the 5 broad geographic regions in the data.

Human Study - GeoGuessr We develop an annotation interface similar to Geoguessr (2024) and ask our pool of annotators to label images at either the country or the region level or at continent level. We ask for a maximum of 5 labels and a minimum of 1 per image, to adjust for different familiarity levels with multiple randomly sampled regions/cultures. Appendix C.3 includes detailed descriptions of annotator demographics, interface and instructions provided to annotators.

3

Figure 2: Confusion matrices for GPT-4V on the cultural awareness task for DALLE STREET images. Accurate responses match the true subregion exactly. Special labels include Invalid (no match to region labels or incomplete) and ResponsibleAI (no region label due to policy violation). **Takeaway:** The model shows balanced performance, indicated by a strong leading diagonal. Notably, the model achieves 100% accuracy for Western Asia images, covering Iran, Jordan, Lebanon, Oman, Palestine, and Turkey.

3.2 Results

Contrary to our expectations, LLaVA is not far off from GPT-4V in terms of cultural awareness, as measured in our experimental settings for this task.

Overall comparison We show the aggregated results for our Cultural Awareness Task in Table 1. LLaVA performs as good as GPT-4V on Dollar Street and even outperforms it significantly on the DALLE STREET images. This is a potential indication that LLaVA might be more prone to forming *stereotypical* associations between regions and concepts, because the DALLE STREET images include such associations to distinguish different geographical regions. As we discuss later, it might be possible to trace this down to the training data.

However, we note that on the MaRVL data, LLaVA performs about as good as random guessing, specifically getting incorrect predictions for nearly all of the images corresponding to Swahili and Turkish. Our hypothesis to explain this is based on the fact that the other two datasets have similar distributions about concepts and countries, whereas MaRVL is much more diverse and covers fewer countries but many indigenous concepts, not all of which the model may have seen during training.

Subregion level analysis Figure 2 (and Figure C.4.15) indicates that both GPT-4V and LLaVA
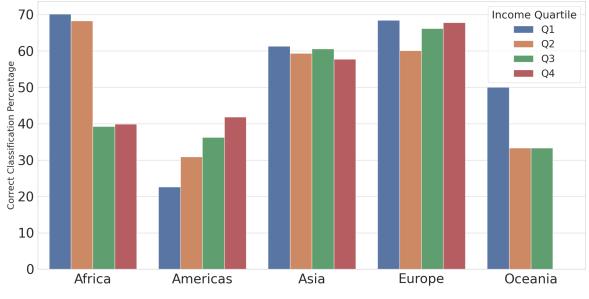
Figure 3: We normalize income data from Dollar Street into region specific quartiles and plot corresponding accuracies for GPT-4V. **Takeaway:** Lower income quartiles (blue, orange bars) are more accurate in Africa and Asia, but higher income quartiles (red, green) are more accurate in Americas; in Europe all quartiles have similar accuracies.

have strong leading diagonals, which is an indicator that they get many predictions correct. GPT-4V has many cases where it does not provide any answer due to violations of the content policy filter. Instead of trying to find workarounds for the filter, we show it as is, because this would represent the experience of any typical user of the system. Very interestingly, both GPT-4V and LLaVA gets all of the answers from Western Asia correct. In general, LLaVA makes a lot of mistakes for different regions, but defaults to South America as the incorrect answer for these cases, whereas for GPT-4V the same translates to the ResponsibleAI policy filter output as the incorrect answer. We include results for the other two datasets in the Appendix (Figure C.4.16 and C.4.17), and find similar trends.

Economic disparity - GPT-4V We utilize the monthly income information available as part of the Dollar Street data, and divide the images into groups based on normalized income quartiles for each broad geographical region - Africa, Asia, Americas, Europe and Oceania (see Figure 3). Lower income quartiles show better performance for Africa and Asia, whereas for America performance improves for higher income groups. This discrepancy reveals the latent biases of the model, which seem to associate Africa with poorer contexts and America with richer ones. For Europe, performance is nearly similar for all quartiles.

4 Automated Extraction of Implicit Cultural Artifacts (Task 2)

To identify the implicit associations that the models may be using for performing Task 1, we propose
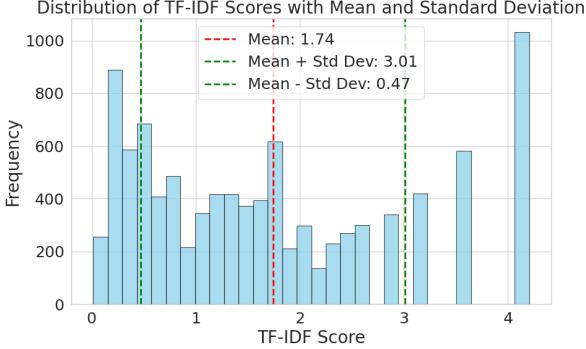
Figure 4: We assign a score for each identified artifact based on how likely it is to co-occur for a given country. Across all such scores, we find the mean and standard deviation to define a range. Scores that lie outside this range would correspond to items that co-occur frequently for a given country, and would most likely represent *stereotypical* implicit associations.

a second task to extract specific cultural artifacts from the images, and discover interesting associations, which are usually stereotypical for the relevant countries.

4.1 Methods

We use the open vocabulary object detection (Zareian et al., 2021) capabilities of GPT-4V `vision-preview`, formulating a detailed prompt (see Appendix C.2) to extract information about the target concept class in a DALLE STREET image, including text descriptions, color,[5] and person count. GPT-4V's strong instruction following capabilities allow the generation of a nearly perfectly validated JSON string, which we lightly post-process, and then provide that as input to GPT-4 `turbo` to get a summary of the objects for any given image. In our initial experiments, GPT-4V outperformed LLaVA significantly in terms of following complex instructions for object detection, so we only report results from GPT-4V.

This process results in many unique cultural artifacts associated with a single country. To narrow this down to the *salient* artifacts that occur relatively more frequently for one country but less for others (i.e. *stereotypical* associations), we follow an approach similar to Jha et al. (2024): we compute the term frequencies of each artifact for each country and also compute document frequency as the number of times an artifact occurs across all countries. The final `tf-idf` score for a given artifact for a particular country is obtained by multiply-
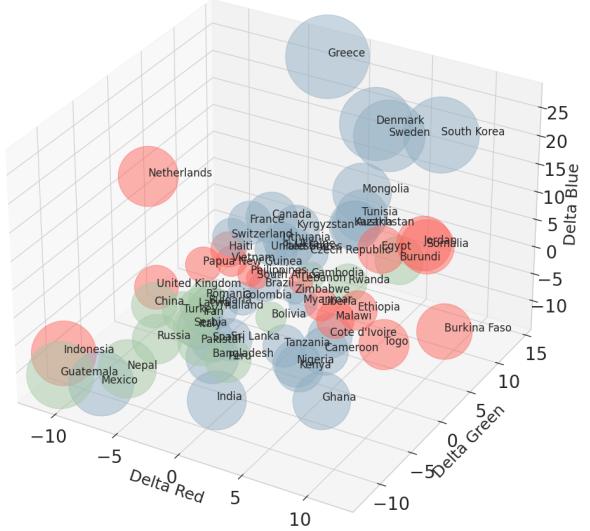

Figure 5: We explore how countries are distributed on a color spectrum by first calculating a global average RGB vector and then defining deltas along each axes aggregated at the country level. **Takeaway**: We find interesting associations - Greece is strongly correlated with blue, Burkina Faso with red.

ing term frequency and the inverse of the document frequency. We then use the distribution of these scores by calculating the mean and standard deviation across all scores, and then perform a qualitative evaluation of both outliers and scores that fit the global distribution.

Color Associations for Countries We calculate the mean RGB vector for each synthetic DALLE STREET image, and then calculate the average to define a global mean vector. We repeat this process at the country level to obtain a mean vector for each country. Then, we find the distance of each country from the global mean and split across the three dimensions of the vector to find colors that are more likely to be associated with some countries. One of the challenges of this approach is that we only consider a limited number of color axes which may not be sufficient to capture finer signals about more complicated color associations, similar to McCarthy et al. (2019). We leave this for future work.

Counting the Number of People In our initial observations, we noticed that DALL-E 3 tends to generate different population densities in images from different countries for otherwise identical prompts. To explore this, our object detection prompt also reports the count of people for each image, split in two buckets (between 1 and 10 people, or more than 10 people). We further divide the first bucket into two smaller buckets based on
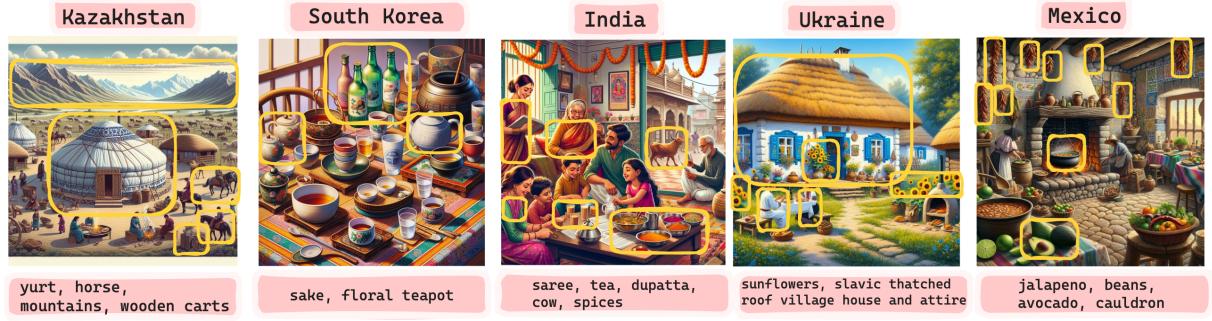
Figure 6: We identify more than 18,000 unique cultural artifacts across all countries as part of our second task, and then filter them to find salient ones. This figure shows strongest correlated artifacts for 5 randomly picked countries. More examples in Appendix C.4.13.

the exact count (less than 5 people, or between 5 and 10 people), to get 3 buckets overall. We process this output to consider all possible variations of these terms (for example, people, person, man, woman, etc) and aggregate statistics at the country level to find the distribution of population densities in generated images.

4.2   Results

We analyse all the DALL-E 3 generated images using GPT-4V to discover hidden associations that these LMMs makes between countries and cultural artifacts.

Selected Examples of Associations within Standard Deviation   Figure 6 (and Figure C.4.13) illustrates some of the randomly sampled interesting cultural artifacts we discover, and Figure 4 shows the distribution of the tf-idf scores we calculate for each country-artifact pair. Artifacts that fall outside the range as denoted by the standard deviation lines from the mean, would be very strongly associated with a particular country and are thus very important to more closely look at, to consider if they are stereotypical (especially negatively) or not. We include statistics for the number of artifacts we identify along with more examples of interesting associations in Appendix C.4 Table 6.

Color associations   We find that models tend to not only associate particular cultural artifacts with countries, but also colors. Figure 5 (and Figure C.4.18) shows a subset of all the countries in the DALLE STREET images versus the delta values in the RGB vector components from the global average vector. Once again, for some countries, these delta values fall well outside the range denoted by the standard deviation lines from the mean and these represent strong presence of that color component in images for that country. For example,

Figure 7: We explore people count associations being made by DALL-E 3 and GPT-4V, and show some selected countries where generated images in DALLE STREET have more than 10 detected people. **Takeaway:** African countries typically fall into high-person-count buckets in our experiments.

Greece is very strongly linked to the colors blue and green, whereas Indonesia and Netherlands are associated with red. An explanation for why this may be happening may be inferred from inspecting the generated images. Images of Greece typically feature a lot of sea, blue decorations, and blue patterns (usually *stereotypical*) whereas Netherlands is famous for its red tulip fields, which is also an artifact association that we are able to obtain from our previous analysis.

People-Count Associations   We parse the information obtained about population density per DALL-E 3 image from our object detection prompt for GPT-4V, to get 3 buckets overall (between 1 and 5 people, between 5 and 10 people or more than 10 people). Figure 7 (and C.4.19) shows a sample of countries for each bucket versus the number of images for that country which fall into that bucket.

A general trend is that images usually fall into the two extreme buckets, and we have very few occurrences of the middle bucket. Cameroon gets

Figure 8: Our CULTUREADAPT pipeline identifies the fine-grained elements of an image that should be modified for cultural adaptation. Note that any in-painting technique can be used. Compare CULTUREADAPT with DALL-E 3, which generates a completely different image. More examples of edits in Appendix C.4.14.

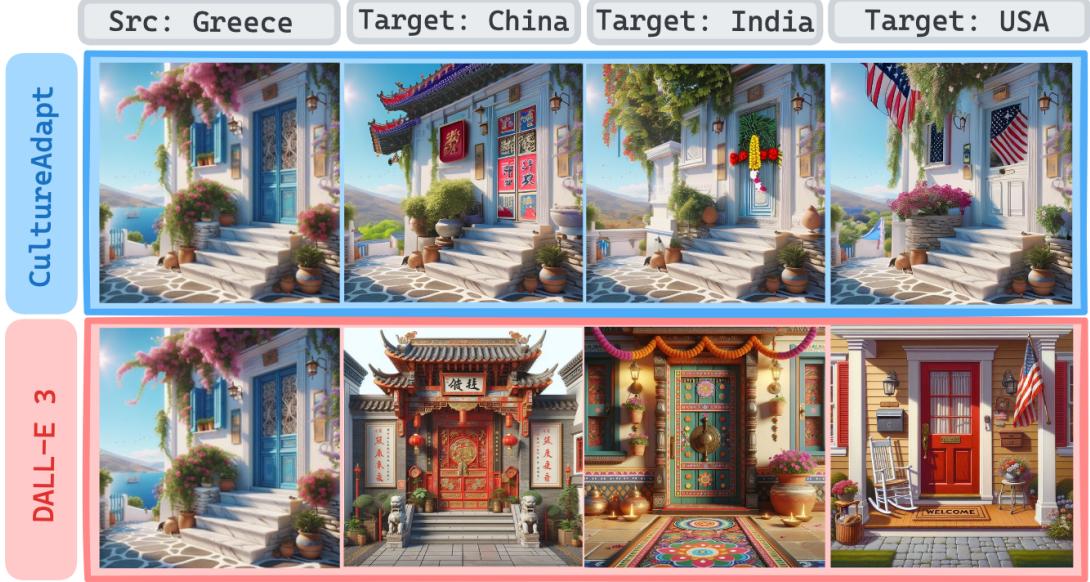the highest number of occurrences in the bucket of more than 10 people counts, despite other countries like China and India having much higher population densities in the real world. In general, African countries fall into the high-person-count buckets, while European ones fall largely in the few-people one. This might reflect the models' perception of societies as largely individualistic or collectivist. That said, as there might be errors by GPT-4V in the person-counting process in the images –which we have not validated except from looking at a handful of examples– we recommend taking these results with a grain of salt, and leave this further exploration for future work.

5 Cultural Adaptation (Task 3)

We propose a highly composable method, CUL-TUREADAPT, for fine-grained editing that aims to use the discovered cultural artifacts from Task 2, and adapts a given source image to a target culture. This has many potential downstream applications. For example, in generative AI applications, if the generated content is detected to not follow the specified country label, then our pipeline can be used to update it in near real-time. Another useful application of our pipeline would be as a source of data augmentation for pre-training corpora. We can use the pipeline to create culturally augmented data and also include translations to create a parallel corpus across languages and cultures which can then be used for continued pre-training to mitigate identified issues.

5.1 Methods: CULTUREADAPT

Recent work on cultural translation (Khanuja et al., 2024; Li and Zhang, 2023; Fung et al., 2024) defines different approaches for adapting images or text from one culture to another. We propose a simple automated and highly composable pipeline that can extract culture specific entities from an image, find the most salient ones, ground them with bounding boxes, and then modify only these objects without affecting the rest of the image. After this cultural adaptation process, we use CLIPScore (Hessel et al., 2021) as a reference-free metric to measure the image-country compatibility using CLIP (Radford et al., 2021) (treating the name of the country as the caption).

Our pipeline is built with fully customizable components. For example, we currently use GPT-4V vision-preview for open-vocab object detection, Grounding DINO (Liu et al., 2023b) for grounding objects in bounding boxes and Stable Diffusion 2 inpainting (Rombach et al., 2021) for updating these grounded object masks. But, we could also use something like Tag2Text (Huang et al., 2023) or RAM (Zhang et al., 2023) to generate captions for an image, and then extract tags from these captions corresponding to objects, which can then be passed into Grounded SAM (Ren et al., 2024) to obtain bounding boxes and segmentation masks, which will finally be used by any

inpainting model like Stable Diffusion 3 (AI, 2024) or MimicBrush (Chen et al., 2024). We demonstrate the feasibility of our approach.

Evaluation Let image I_1 originally correspond to country C_1. We use CULTUREADAPT to translate it to I_2 for country C_2. The CLIPScore of an image-country pair is denoted as $S(I, C)$. We define two delta values from this as follows:

$$\Delta_1 = S(I_2, C_1) - S(I_1, C_1)$$
$$\Delta_2 = S(I_2, C_2) - S(I_1, C_2)$$

If $\Delta_1 < 0$ and $\Delta_2 > 0$, then it would imply that after cultural adaptation, the image is closer to the target country C_2 than the source country C_1. This is exactly the success criterion for our task.

5.2 Results

We show examples of using CULTUREADAPT in Figure 8, where we start with a DALLE STREET image for the concept of front door in the country Greece. Then we culturally translate these to 3 other countries, China, India and USA. For all of these images shown, $\Delta_1 < 0$ and $\Delta_2 > 0$ (values in Appendix C.4.3), implying that quantitatively we have measured a culture change for all 3 cases. We include more examples of CULTUREADAPT in Appendix C.4.14.

We further contrast these with images generated for the same concept, country pairs from DALL-E 3 to highlight differences in the two approaches. Images generated from DALL-E 3 may be more representative of the target country as per CLIP-Score but they are stylistically very different from the original source image (i.e lack object consistency before and after the cultural translation process), thus making them less usable for different applications where this property is necessary.

6 Human Studies

We choose a small subset of our data, about 300 images randomly sampled, and perform a couple of human studies with 14 people. Here we present highlight findings due to space, with more details and annotator demographics in Appendix C.3.

Country Representation in DALLE STREET We examine whether annotators consider specific images to be stereotypical representations of a country. Figure C.4.20a shows that most participants agree or strongly agree, with only 2 out of 304 images receiving disagreement. When unsure,

participants tend to neither agree nor disagree about the appropriateness of an image, . This study indicates that our generated images effectively represent country-level associations. Our participants also *mark* this information with visual cues (Table C.7), from explicit (e.g. flags) to implicit (e.g. color choice, nuanced cultural artifacts). Based on this analysis, we develop our framework for cultural artifact extraction and cultural adaptation using DALLE STREET images.

Human baseline for Cultural Awareness We look at accuracy metrics from 5 different perspectives. For our study, we allowed users to select a minimum of 1 label and a maximum of 5 per image. These labels can be at the country level, or at the continent level or subcontinent level. So, we can evaluate exact match accuracy first at the country level, then at the subregion level and finally at the continent level. In addition, we also consider the case of union (correct answer at one of the three levels is present in the human provided labels), and the case of intersection (correct answer at all three levels are present in the human provided labels). Figure C.4.20b shows that on average participants have a low accuracy at the country level, but nearly double performance every time we go up a geographical level, which is along the lines of what we expect.

7 Conclusion

This study addresses the critical need for cultural awareness in Large Multimodal Models (LMMs) by introducing a comprehensive framework to evaluate and enhance their cultural competence. We create a large-scale, culturally diverse dataset of $9,935$ images across 67 countries and 10 concept classes, facilitating robust evaluation of LMMs on cultural awareness tasks. Further, we introduce a cultural artifact extraction task to identify over $18,000$ cultural artifacts that co-occur frequently with these countries, revealing significant insights into the implicit cultural associations encoded in these models. We also propose a cultural adaptation task and a highly composable pipeline CULTUREADAPT to adapt images across cultural contexts with fine-grained edits. Overall, this work emphasizes the importance of developing culturally sensitive AI systems and provides a foundational benchmark for future research towards improvement in cultural representation.

Limitations

Coverage   Our work focuses on a limited number of concept classes across a small portion of all countries in the world. While we are able to cover most geographic regions and sub-regions with this data, it is important to ensure wider coverage over time to ensure that no culture is left behind.

Open Vocabulary Object Detector   Our pipeline for automated cultural adaptation is only as good as its components. The first and most crucial component of meaningful object tag extraction is approached using GPT-4V in our work, but open-source alternatives like RAM and Tag2Text are slowly catching up to cover more wider variety of classes.

Stable Diffusion Biases   While we are using Stable Diffusion for inpainting masked objects only, our approach does not guarantee that the object generated from the diffusion process will not reflect any harmful *stereotypes*. Our CLIPScore based metric for measuring country deltas is also preliminary and limited by the strength of the underlying CLIP model embedddings and its inherent biases as well.

Mitigation Strategies

An important aspect of studying differences in culture awareness of LMMs is understanding how to develop approaches to mitigate them. This may involve looking more closely at pre-training corpora and developing efficient techniques for data augmentation for the same, so that secondary approaches like prompting in local languages can extract more aligned responses. While exploring all of these is beyond the current scope of our paper, we present a survey of the most relevant recent work that attempts to do so.

Attributing model behavior to specific training examples   Methods like Park et al. (2023) aim to better understand issues with training data by assigning scores to data points which cause the model to behave in a certain way. This is useful as a first step to trace training data that reflects potentially harmful *stereotypes* from the list of automatically identified cultural artifacts that we identify. Recent work (Hall et al., 2023) has looked at the LAION corpus (Schuhmann et al., 2022) to determine that geographical regions may only co-occur with more stereotypical image captions. For example, a Greek person having dinner in Greece is less likely to caption a photo of that scene as "having food in Greece" but a tourist is more likely to mention something like "when in Greece, you must moussaka" in a caption for the same image, leading to perpetuation of stereotypes when these image-caption pairs are used for training models. Similarly, color associations that we find in this paper may simply be because images were scraped from travel websites which usually portray a more vibrant view of a country than the native viewpoint.

Data Augmentation   There are many interesting approaches to generate higher quality data to mitigate cultural differences. Dollar Street for example uses photographers as a source of collecting this data manually, whereas MaRVL uses crowdsourcing. Other automated approaches include knowledge acquisition by scraping from internet sources (Fung et al., 2024), simulating conversations between LLMs (Li et al., 2024b), semantic data augmentation (Li et al., 2024a) and value preference alignment by collecting preferences during conversations with LLMs (Kirk et al., 2024). For multimodal data augmentation, methods include approaches based on finding synonyms from semantic graphs of image captions (Li and Zhang, 2023) and end-to-end pipelines for image trans-creation (Khanuja et al., 2024). Our approach CULTUREADAPT aims to contribute to this area of research by automatically identifying cultural artifacts and providing a composable pipeline for finer-grained edits to these artifacts.

Prompting in local languages   Hall et al. (2023) find that images generated with non-English languages tend to struggle with prompt consistency and continue to show stereotyped representations, whereas Pouget et al. (2024) find that pre-training with global, unfiltered data before fine-tuning on English content can improve cultural understanding without sacrificing performance on popular benchmarks. These contrasting findings imply the need to dedicate more efforts to find the impact of pre-training on data that has been augmented for visual modality using pipelines such as ours along with translated image captions for a multimodal parallel corpora (Etxaniz et al., 2024) to eliminate geographical stereotypes. Improving the quality of the text encoders used in LMMs is another important angle to look at in this light.

Acknowledgements

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Ashutosh Dwivedi, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in llms: A survey.

Stability AI. 2024. Stable diffusion 3 released. https://stability.ai/news/stable-diffusion-3.

Abhipsa Basu, R. Venkatesh Babu, and Danish Pruthi. 2023. Inspecting the geographical representativeness of images from text-to-image models.

Yong Cao, Wenyan Li, Jiaang Li, Yifei Yuan, Antonia Karamolegkou, and Daniel Hershcovich. 2024. Exploring visual culture awareness in gpt-4v: A comprehensive probing.

Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. CHI '17, page 2334–2346, New York, NY, USA. Association for Computing Machinery.

Xi Chen, Yutong Feng, Mengting Chen, Yiyang Wang, Shilong Zhang, Yu Liu, Yujun Shen, and Hengshuang Zhao. 2024. Zero-shot image editing with reference imitation.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2024. Bertaqa: How much do language models know about local culture? Preprint, arXiv:2406.07302.

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively multi-cultural knowledge acquisition & lm benchmarking.

Xiao Ge, Chunchen Xu, Daigo Misaki, Hazel Rose Markus, and Jeanne L Tsai. 2024. How culture shapes what people want from ai. ArXiv preprint, abs/2403.05104.

Geoguessr. 2024. Geoguessr: A geography game. https://www.geoguessr.com.

Melissa Hall, Candace Ross, Adina Williams, Nicolas Carion, Michal Drozdzal, and Adriana Romero Soriano. 2023. Dig in: Evaluating disparities in image generations with indicators for geographic diversity.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza

Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xinyu Huang, Youcai Zhang, Jinyu Ma, Weiwei Tian, Rui Feng, Yuejie Zhang, Yaqian Li, Yandong Guo, and Lei Zhang. 2023. Tag2text: Guiding vision-language model via image tagging.

Oana Ignat, Gayathri Ganesh Lakshmy, and Rada Mihalcea. 2024. Cross-cultural inspiration detection and analysis in real and llm-generated social media data.

Akshita Jha, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Rida Qadri, Chandan K. Reddy, and Sunipa Dev. 2024. Visage: A global-scale analysis of visual stereotypes in text-to-image generation.

Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. 2024. An image speaks a thousand words, but can everyone listen? on image transcreation for cultural relevance.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The prism alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models.

A. L. Kroeber. 1952. Culture: A Critical Review of Concepts and Definitions. The Museum, Cambridge, Mass. Retrieved from https://nrs.lib.harvard.edu/urn-3:fhcl:30362985. Accessed 11 June 2024.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. Culturellm: Incorporating cultural differences into large language models.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024b. Culturepark: Boosting cross-cultural understanding in large language models.

Zhi Li and Yin Zhang. 2023. Cultural concept adaptation on multimodal reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 262–276, Singapore. Association for Computational Linguistics.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2024. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2023b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection.

Arya D. McCarthy, Winston Wu, Aaron Mueller, William Watson, and David Yarowsky. 2019. Modeling color terminology across thousands of languages. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2241–2250, Hong Kong, China. Association for Computational Linguistics.

OpenAI. 2024. Dall·e 3 technical report. https://cdn.openai.com/papers/dall-e-3.pdf. [Accessed: June 9, 2024].

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, and et al. 2023. Gpt-4 technical report.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. 2023. Trak: Attributing model behavior at scale.

Angéline Pouget, Lucas Beyer, Emanuele Bugliarello, Xiao Wang, Andreas Peter Steiner, Xiaohua Zhai, and Ibrahim Alabdulmohsin. 2024. No filter: Cultural and socioeconomic diversity in contrastive vision-language models.

Haoyi Qiu, Wenbo Hu, Zi-Yi Dou, and Nanyun Peng. 2024. Valor-eval: Holistic coverage and faithfulness evaluation of large vision-language models.

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*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. 2024. Grounded sam: Assembling open-world models for diverse visual tasks.

William Gaviria Rojas, Sudnya Diamos, Keertan Ranjan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. 2022. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. In *Neural Information Processing Systems*.

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

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.

Katie Seaborn, Yuto Sawa, and Mizuki Watanabe. 2024. Coimagining the future of voice assistants with cultural sensitivity. *ArXiv preprint*, abs/2403.17599.

Hansa Srinivasan, Candice Schumann, Aradhana Sinha, David Madras, Gbolahan Oluwafemi Olanubi, Alex Beutel, Susanna Ricco, and Jilin Chen. 2024. Generalized people diversity: Learning a human perception-aligned diversity representation for people images.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020–2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

UN. 2024. Methodology: Standard country or area codes for statistical use (m49).

Da Yin, Feng Gao, Govind Thattai, Michael Johnston, and Kai-Wei Chang. 2023. Givl: Improving geographical inclusivity of vision-language models with pre-training methods.

Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. 2021. Open-vocabulary object detection using captions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 14393–14402. Computer Vision Foundation / IEEE.

Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, Yandong Guo, and Lei Zhang. 2023. Recognize anything: A strong image tagging model.

A Related work

Recent spike in interest for culturally aware NLP has produced many relevant works, which has inspired different aspects of our research.

Li and Zhang (2023) uses semantic graphs to develop an annotation free data augmentation approach for augmenting cultural components of captions. The main challenge of this approach is that the cultural artifacts identified for replacement are copy-pasted into the target image, which leads to inconsistencies at the object boundaries. Similarly, Khanuja et al. (2024) define the image transcreation task formally in the context of culture, and develop pipelines for cultural translation. The edited images from these pipelines tend to vary in structure, and may sometimes look very different from the source image. Our proposed CULTUREADAPT pipeline overcomes both of these challenges, by using automatically generated object labels grounded with bounding boxes to create precise semantic masks which are then filled in with a diffusion based model in an inpainting setting.

Qiu et al. (2024) calculates co-occurence statistiscs for features extracted from images and Jha et al. (2024) assigns importance scores to attributes to determine those which co-occur more frequently for an identity group over others. We explore both of these ideas in our work to understand which cultural artifacts are more likely to co-occur for a given country. Pouget et al. (2024) and Hall et al. (2023) both explore the two real-world datasets, MaRVL and Dollar Street, specifically developing reliable metrics to measure geo-localization and object consistency across regions. Liu et al. (2024) defines a taxonomy for culturally aware and adapted NLP which we borrow terminology from, whereas Adilazuarda et al. (2024) provides a survey of the current state of research in this sub-field and also what is yet to be done.

Other works also explore downstream tasks, for example, Ignat et al. (2024) looks at cross-cultural inspiration detection in social media data, Seaborn et al. (2024) develops voice assistants with cultural sensitivity and Srinivasan et al. (2024) improves cultural diversity of search results.

B DALLE STREET generation prompt

> **Dalle Street Generation Prompt**
>
> A typical scene of **{category}** in **{country}**, culturally accurate and detailed.

Figure 9: We use a simple prompt that includes information about the concept class and the target country using a template, to generate our large scale dataset of DALL-E 3 images.

C Appendix

C.1 Dataset details

Dollar Street concept classes (10) car, cups, mugs and glasses, family snapshots, front door, home, kitchen, plate of food, social drink, wall decoration and wardrobe.

Dollar Street countries (63) South Africa, Serbia, Indonesia, Brazil, Kenya, India, Nigeria, France, Kazakhstan, United States, Philippines, Mexico, Sri Lanka, Netherlands, Thailand, Colombia, Pakistan, China, Russia, Egypt, Iran, United Kingdom, Romania, Spain, Turkey, Ukraine, Italy, Czech Republic, Denmark, Ethiopia, Jordan, Burundi, Burkina Faso, Malawi, Somalia, Zimbabwe, Haiti, Cote d'Ivoire, Myanmar, Papua New Guinea, Liberia, Cambodia, Bangladesh, Rwanda, Nepal, Palestine, Tunisia, Cameroon, Bolivia, Ghana, Vietnam, Guatemala, Mongolia, South Korea, Kyrgyzstan, Lebanon, Tanzania, Switzerland, Sweden, Canada, Peru, Austria and Togo.

Concept classes (10) for DALLE STREET car, cups, mugs and glasses, family snapshots, front door, home, kitchen, plate of food, social drink, wall decoration and wardrobe.

Countries in our DALLE STREET (67) Austria, Bangladesh, Bolivia, Brazil, Bulgaria, Burkina Faso, Burundi, Cambodia, Cameroon, Canada, China, Colombia, Cote d'Ivoire, Czech Republic, Denmark, Egypt, Ethiopia, France, Ghana, Greece, Guatemala, Haiti, India, Indonesia, Iran, Italy, Jordan, Kazakhstan, Kenya, Kyrgyzstan, Latvia, Lebanon, Liberia, Lithuania, Malawi, Mexico, Mongolia, Myanmar, Nepal, Netherlands, Nigeria, Pakistan, Palestine, Papua New Guinea, Peru, Philippines, Romania, Russia, Rwanda, Serbia, Somalia, South Africa, South Korea, Spain, Sri Lanka, Sweden, Switzerland, Tanzania, Thailand, Togo, Tunisia, Turkey, Ukraine, United Kingdom, United States, Vietnam, Zimbabwe

MaRVL Country to Language Mappings In the context of the MaRVL dataset, various languages are mapped to specific sub-regions based on the countries where these languages are predominantly spoken. The mapping is as follows:

- `"id"`: The language code for Indonesian, which is primarily spoken in **Indonesia**, corresponds to the **South-eastern Asia** sub-region.

- `"sw"`: The language code for Swahili, used in countries such as **Tanzania**, **Kenya**, and **Rwanda**, is mapped to the **Eastern Africa** sub-region.

- `"ta"`: The language code for Tamil, spoken in **India** and **Sri Lanka**, is associated with the **Southern Asia** sub-region.

- `"tr"`: The language code for Turkish, which is the official language of **Turkey**, falls under the **Western Asia** sub-region.

- `"zh"`: The language code for Chinese, predominantly spoken in **China**, is linked to the **Eastern Asia** sub-region.

Table 2: Dataset Statistics

Sub-region	Eastern Africa	Eastern Asia	South-eastern Asia	Southern Asia	Western Asia	Caribbean	Central America	Central Asia	Eastern Europe	Melanesia
MaRVL	875	1107	1091	924	917	-	-	-	-	-
Dollar Street	310	313	578	839	128	56	12	20	136	14
DALL-E 3 Images	1052	438	840	742	600	160	176	280	741	147
Total	2237	1858	2509	2505	1645	216	188	300	877	161

Sub-region	Middle Africa	Northern Africa	Northern America	Northern Europe	South America	Southern Africa	Southern Europe	Western Africa	Western Europe	Total
MaRVL	-	-	-	-	-	-	-	-	-	4914
Dollar Street	107	81	317	51	447	60	223	262	183	4137
DALL-E 3 Images	303	289	465	736	605	139	740	888	594	9935
Total	410	370	782	787	1052	199	963	1150	777	18986

C.2 Prompt details

Prompt used for Object Detection with GPT-4V We use a detailed prompt for GPT-4V to extract objects, colors and counts from images generated with DALL-E 3.

GPT-4V Object Detection Prompt

```
Give me a json output of the items you see in this image in both the foreground and background.
Output the objects as a JSON with two fields: 'relevant_objects' for objects pertinent to the
image category concept and 'other_objects' for all additional detected objects. Be as specific
as possible. Within each field, for each detected object, include sub-fields describing object
attributes like color, count, and anything else that is appropriate. For example, for buildings
describe the architectural style in a sentence, for people describe clothing and headgear (if
multiple colors and headgears are present, include the top three), for food items describe the
exact type of food and include a brief recipe description, for pictures of rooms include objects
in the background like mountains outside a window or paintings on the wall portraying something
specific like a landmark or a particular type of scenery. For the counts of items, if the number
of items is less than 10, give me exact numbers otherwise say more than 10.
```
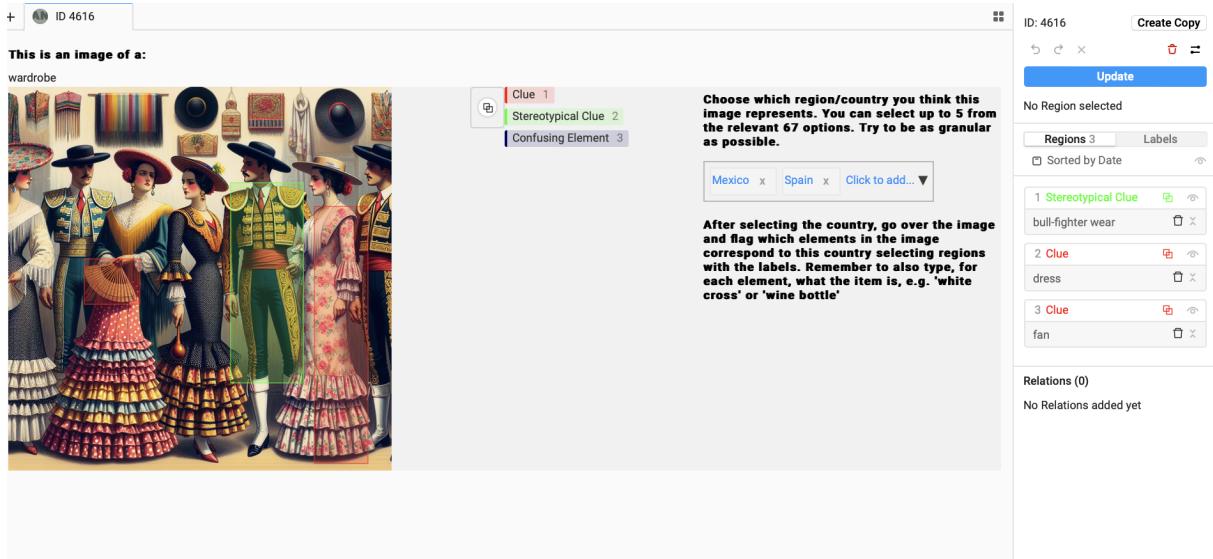

Figure 10: Annotation Interface for Study 1

Prompt used for processing generated objects   We use a detailed prompt for GPT-4 to process the dictionaries generated with the previous prompt into a simplified list along with some parsing rules to ensure correctness of the data structure.

```
GPT-4 Processing Object Detection Outputs

You will be provided a dictionary of items for the country {country} and concept {concept}. Summarize
the dictionary into a list of comma-separated list of items with their respective colors. For
example, [red apple, blue car, green tree, house with a red roof and tinted glasses]. Strictly
follow the output format requested. The dictionary is as follows:
```

C.3   Human Study - Interface and Annotators

Annotator Demographics   All annotators are demographically located in the USA and are between 25-40 years old. Roughly 40% identify as female, and the rest as male. In terms of an education background, 40% of the annotators are graduate students, whereas the rest includes working professionals from different backgrounds and also computer science faculty. In total, we have 14 annotators, recruited from different computer science labs at an university and also from a diverse set of social connections for this study. All the annotators have agreed to consent for using this data for research purposes. Our study qualifies for exemption from IRB as no PII is involved.

General Instructions for Human Study   We use https://labelstud.io (Tkachenko et al., 2020–2022) to perform our human study. For each annotator, we create 2 tabs corresponding to the 2 studies, and ask them to solve them in numerical order to avoid potential biases from seeing true labels first from the second study. Time taken to complete the first study is usually 2 hours, and the time taken to complete the second study is typically 30 minutes. All annotators will be compensated for their time with a $20 gift card upon completion of the task.

Task 1 Instructions   For every image, you have to make atleast 1 guess for the geographical region label, along with atleast 1 corresponding clue. If you are not sure what the clue is, add a question mark symbol at the end of it - example, headgear? bread?). Do not reverse image search or look anything up. Answer only using your knowledge or instinct. You can try guessing sub-region/region if you are not sure about country. You can use the knowledge of the fact that the image is generated by an AI conditioned on the provided prompt above the image. Note that you do not have to be correct!

Once the label is done, you need to add at least one bounding box somewhere in the image (it can be very specific and small or very broad or even the entire image) and then label that bounding box as either a

14

	Δ_1	Δ_2
Greece → China	-8.48	5.07
Greece → India	-3.47	0.11
Greece → USA	-6.13	5.81

Table 3: Delta values for cultural adaptation from source country to target country. A successful cultural adaptation is indicated by $\Delta_1 < 0$ and $\Delta_2 > 0$.

clue or a stereotypical clue or a confusing element and then add a text description for the bounding box from the interface on the right (for example, a bounding box for a basket of baguettes can be a clue for France and the text description may be either something specific like "baguette" or something generic like "bread?"). The difference between stereotypical clue and regular clue is that stereotypical would be something like baguettes for France or "naan" for India or specific clothing styles for some country whereas a regular clue is something that you are using to make your guess but you don't know enough about your guess to know what stereotypical clues might be associated with it, for example, sand for island countries.

Figure 11: Annotation Interface for Study 2

Task 2 Instructions   For every image, select a single rating for "appropriateness" = "this image is one of the possible (stereo)typical representation of the mentioned category for the mentioned country". Then select at least one clue corresponding to your rating, similar to the first study.

C.4   Additional Results

Performance of Human Study participants on the Cultural Awareness Task   We include anonymized performance statistics of each of our annotators to show differences in performance at the individual data point level for countries, subregions and continents.

How many artifacts did we identify using GPT-4V   Table 5 shows the counts across each of the 67 countries for the number of artifacts identified. `adj` implies unique artifacts are a combination of

User	Country Level	Subregion Level	Continent Level
User 1	46.53	70.14	90.97
User 2	16.67	31.94	78.47
User 3	11.19	31.47	70.63
User 4	32.81	50.78	72.66
User 5	21.13	51.41	75.35
User 6	10.87	32.61	71.01
User 7	28.67	64.34	84.62
User 8	22.14	43.57	69.29
User 9	7.09	34.04	70.92
User 10	26.43	62.86	83.57
User 11	20.71	50.71	87.14
User 12	4.86	18.05	75.69
User 13	3.57	17.85	72.14
User 14	6.47	37.41	75.53

Table 4: User accuracies across country, subregion and continent levels, rounded to two decimal places. At the country level, accuracy varies from 46% to about 4%, so the subregion level accuracies are a more reliable indicator of performance even for humans. Continent is the most generic label, and has very high accuracies from all participants.

words and adjectives that appear before it to quantify count or color. `no_adj` implies only the raw words identified. Figure 4 includes a distribution of the TD-IDF scores for these (country, artifact) pairs and high scores that lie outside the range as given by the mean and the standard deviation of the distribution (i.e larger than 3.01 or smaller than 0.47) would imply strongly correlated artifacts for a given country, and there exists 4019 such items from this data. Further human filtering can be done to remove common words like table or mailbox or dresses to find unique and interesting associations like pretzels in Austria, zinnias in Bolivia and many more, some of which we report in Table 6.

Table 5: Artifact Statistics

	Austria	Bangladesh	Bolivia	Brazil	Bulgaria	Burkina Faso	Burundi	Cambodia	Cameroon	Canada
adj	248	283	279	264	264	288	292	276	251	292
no_adj	154	148	141	157	138	153	141	161	133	170

	China	Colombia	Cote d'Ivoire	Czech Republic	Denmark	Egypt	Ethiopia	France	Ghana	Greece
adj	275	268	270	276	278	276	252	278	265	270
no_adj	124	134	146	160	168	152	139	157	137	151

	Guatemala	Haiti	India	Indonesia	Iran	Italy	Jordan	Kazakhstan	Kenya	Kyrgyzstan
adj	272	301	264	312	246	262	280	257	270	269
no_adj	163	160	147	172	123	135	158	139	149	147

	Latvia	Lebanon	Liberia	Lithuania	Malawi	Mexico	Mongolia	Myanmar	Nepal	Netherlands
adj	269	244	276	267	281	266	282	292	290	260
no_adj	156	133	163	159	152	133	156	155	156	146

	Nigeria	Pakistan	Palestine	Papua New Guinea	Peru	Philippines	Romania	Russia	Rwanda	Serbia
adj	279	254	270	279	265	276	272	244	288	250
no_adj	152	144	137	141	137	152	149	144	161	143

	Somalia	South Africa	South Korea	Spain	Sri Lanka	Sweden	Switzerland	Tanzania	Thailand	Togo
adj	291	272	279	263	250	260	265	272	273	279
no_adj	165	165	144	149	150	157	155	145	154	140

	Tunisia	Turkey	Ukraine	United Kingdom	United States	Vietnam	Zimbabwe	Total		
adj	249	254	267	281	273	291	311	18212		
no_adj	133	132	148	181	162	163	166	10035		

More examples of edits using CULTUREADAPT We include more examples of edits across different concept classes and source-target pairs using our CULTUREADAPT pipeline in Figure 14. As can be seen, the pipeline is only constrained by the two bottlenecks of object detection and diffusion based inpainting, which sometimes may detect objects incorrectly or not generate consistent images of human faces for example.
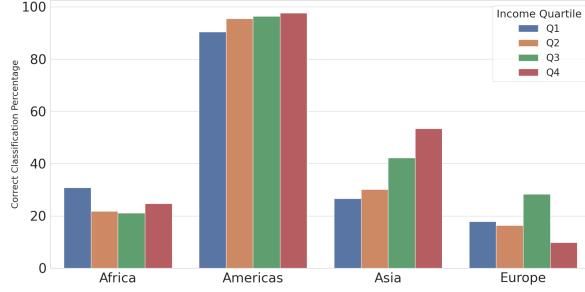
Figure 12: We normalize income data from Dollar Street into region specific quartiles and plot corresponding accuracies for LLaVA.

Table 6: Interesting associations and their explanations for various countries.
Note that these associations are extracted from LMM generations and may not always be accurate.

Country	Interesting Associations and Explanations
Austria	**Dirndl**: A traditional dress worn in Austria and parts of Germany. **Pretzel**: A type of baked bread product, often associated with German-speaking countries. **Lederhosen**: Traditional leather shorts worn by men in the Alpine regions.
Bangladesh	**Lungi**: A traditional garment worn by men, usually a wraparound skirt. **Kurti**: A traditional garment worn by women, often paired with leggings or a skirt. **Harmonium**: A musical instrument commonly used in South Asian music.
Bolivia	**Zinnias**: A type of flower native to the region, known for its bright colors and significance in local celebrations. **Llama**: A domesticated South American camelid, significant in Bolivian culture. **Chullos**: Knitted hats, typically with ear flaps, that are traditional to the Andes.
Brazil	**Bikini**: Associated with the famous beaches of Brazil. **Lychee**: A tropical fruit found in Brazil. **Samba**: A Brazilian music genre and dance style.
Bulgaria	**Spanakopita**: A savory pastry filled with spinach and feta cheese. **Moussaka**: A layered dish with eggplant, potatoes, and minced meat. **Terracotta**: Refers to clay-based unglazed or glazed ceramic.
Cameroon	**Kaftans**: A type of long robe worn in many African countries. **Fufu**: A dough-like food made from cassava or yams. **Savanna**: A type of ecosystem common in Cameroon, characterized by grassland with scattered trees.
Canada	**Poutine**: A dish consisting of fries topped with cheese curds and gravy. **Moose**: A large mammal found in Canada. **Snowmobile**: A vehicle designed for travel on snow, common in Canadian winters.
China	**Changshan**: A traditional Chinese garment for men. **Baozi**: A type of Chinese steamed bun with fillings. **Lion Dance**: A traditional dance in Chinese culture performed during the Lunar New Year and other cultural events.

Continued on next page

17

Country	Interesting Associations and Explanations
Ethiopia	**Injera**: A sourdough flatbread and a staple food in Ethiopia. **Wat**: A traditional Ethiopian stew. **Shawl**: Often worn by Ethiopian women as part of traditional attire.
France	**Camembert**: A famous French cheese. **Baguette**: A long, thin loaf of French bread. **Beret**: A soft, round, flat-crowned hat associated with French culture.
Germany	**Oktoberfest**: An annual beer festival and cultural event in Munich. **Bratwurst**: A type of German sausage. **Dirndl**: Traditional dress worn by women during Oktoberfest and other occasions.
Greece	**Toga**: A garment worn in ancient Greece. **Dolma**: A dish made of grape leaves stuffed with rice or meat. **Moussaka**: A layered dish with eggplant, meat, and béchamel sauce.
India	**Sari**: A traditional garment worn by women. **Lassi**: A yogurt-based drink. **Rangoli**: A form of art created on the floor using colored rice, sand, or flower petals.
Japan	**Kimono**: A traditional Japanese garment. **Sushi**: A popular Japanese dish. **Tatami**: A type of mat used as a flooring material in traditional Japanese rooms.
Mexico	**Sombrero**: A wide-brimmed hat traditionally worn in Mexico. **Tacos**: A traditional Mexican dish. **Guacamole**: A Mexican avocado-based dip or spread.
Morocco	**Tagine**: A North African dish named after the earthenware pot in which it is cooked. **Kaftan**: A long robe worn in Morocco. **Mint Tea**: A popular beverage in Morocco, often served as a welcoming gesture.
Nepal	**Topi**: A traditional hat worn in Nepal. **Himalayas**: The mountain range running across Nepal. **Dal Bhat**: A traditional Nepalese dish consisting of lentils and rice.
Peru	**Chullo**: A traditional hat with earflaps. **Llama**: A significant animal in Peruvian culture. **Ponchos**: Traditional clothing made from wool.
Thailand	**Tuk-tuk**: A common form of transportation in Thailand. **Pad Thai**: A popular Thai noodle dish. **Elephant**: An animal deeply ingrained in Thai culture and symbolism.
Togo	**Kente Cloth**: A traditional fabric made of silk and cotton, known for its vibrant colors and patterns. **Yam Festival**: A major cultural festival celebrating the harvest of yams. **Agbadza Dance**: A traditional dance performed during festivals and ceremonies.
Tunisia	**Shisha**: A popular water pipe used for smoking flavored tobacco.

Country	Interesting Associations and Explanations
	Harissa: A spicy chili paste that is a staple in Tunisian cuisine. **Mosaic Art**: Intricate and colorful tile art that is significant in Tunisian culture.
Turkey	**Evil Eye**: A common talisman believed to protect against negative energy. **Baklava**: A sweet pastry made of layers of filo filled with nuts and honey. **Whirling Dervishes**: A religious dance performed by Sufi practitioners.
Ukraine	**Pysanky**: Traditional Ukrainian Easter eggs decorated with intricate designs. **Borscht**: A beet soup that is a key part of Ukrainian cuisine. **Vyshyvanka**: Traditional Ukrainian embroidered shirts.
United Kingdom	**Afternoon Tea**: A British tradition involving tea and a variety of snacks. **Red Telephone Box**: Iconic public telephone booths found throughout the UK. **Fish and Chips**: A classic British dish of battered fish and fried potatoes.
United States	**Route 66**: A historic highway symbolizing the American road trip. **Thanksgiving**: A national holiday celebrating the harvest and other blessings. **Statue of Liberty**: A symbol of freedom and democracy in the US.
Vietnam	**Ao Dai**: A traditional Vietnamese dress for women. **Pho**: A Vietnamese noodle soup that is a staple dish. **Conical Hat (Non La)**: A traditional hat made of bamboo and palm leaves.
Zimbabwe	**Mbira**: A traditional musical instrument also known as the thumb piano. **Great Zimbabwe**: The ruins of an ancient city, significant in Zimbabwean history. **Victoria Falls**: One of the largest and most famous waterfalls in the world, located on the border between Zimbabwe and Zambia.

Table 7: Cultural artifacts for various countries based on human annotations.
Note that these artifacts are based on subjective perceptions of our human annotators and may not be completely accurate always.

Country	Cultural Artifacts
Austria	beer, sausage, dirndl
Bangladesh	rice, saree, fish
Bolivia	colorful clothes, poncho, hats
Brazil	brazilian flag, tropical fruit, colorful pottery
Bulgaria	clothing, rugs, door
Burkina Faso	dry, black people, straw basket
Burundi	rice, beans, bananas
Cambodia	buddhism, buddhist art, clothing
Cameroon	african people, bananas, beans
Canada	maple leaf, canadian flag, poutine
China	characters, chinese food, lanterns
Colombia	coffee, rice, avocado
Cote d'Ivoire	black people, dry, african outfit
Czech Republic	beer, dress, czech
Denmark	danish flag, beer, windmill
Egypt	hieroglyphs, egyptian art, islamic clothing
Ethiopia	coffee, colors, clay pots
France	baguette, cheese, wine
	Continued on next page

Country	Cultural Artifacts
Ghana	black people, african necklaces, clothing
Greece	blue and white, sea, olives
Guatemala	mayan art, tortilla, beans
Haiti	black people, rice, beans
India	naan, curry, sari
Indonesia	buddhism, rice, clothing
Iran	islamic art, kebab, persian rug
Italy	pizza, pasta, wine
Jordan	clothing, arabic, islamic art
Kazakhstan	clothing, houses, islamic art
Kenya	african people, african art, corn
Kyrgyzstan	clothing, islamic art, rugs
Latvia	clothing, beer, bread
Lebanon	arabic clothing, hummus, bread
Liberia	rice, black people, palm trees
Lithuania	clothing, food, beer
Malawi	hut, corn, black people
Mexico	sombrero, tequila, tortilla
Mongolia	yurt, dumplings, clothing
Myanmar	buddhist art, rice, pagoda
Nepal	buddhist elements, hindu elements, rice
Netherlands	windmill, cheese, dutch clothing
Nigeria	rice, yams, african clothing
Pakistan	clothing, curry, sombrero
Palestine	arabic art, hummus, bread
Papua New Guinea	black people, tropical fruit, coconut
Peru	inca clothing, machu picchu, andes mountains
Philippines	rice, tropical vegetation, cooking
Romania	clothing, sheep, ceramic pots
Russia	fur hat, warm clothes, vodka
Rwanda	african art, beans, dark-skinned people
Serbia	clothing, beer, sausages
Somalia	islamic art, banana, rice
South Africa	african art, corn, hat
South Korea	korean characters, kimchi, korean dress
Spain	flamenco, paella, bull fighting
Sri Lanka	buddhist art, buddhist symbols, spicy food
Sweden	northern european clothing, fish, snowy landscape
Switzerland	alps, swiss cheese, chocolate
Tanzania	african art, rice, meat
Thailand	buddhist art, thai food, clothing
Togo	african clothing, cloth patterns, wood carvings
Tunisia	arabic art, couscous, arched doorways
Turkey	turkish coffee, rugs, kebabs
Ukraine	clothing patterns, flower designs, ukrainian food
United Kingdom	pubs, fish and chips, tea
United States	american flag, burgers, jeans
Vietnam	conical hats, pho, pagodas
Zimbabwe	thatched huts, african clothing, animal carvings

C.5 LLM hyperparameters

We discuss the generation settings we used for our experiments, and also the associated costs and hardware.

Generation settings
- DALL-E 3 images were generated for `vivid` and `natural` settings for `standard` quality and size 1024 x 1024
- GPT-4 and GPT-4V generations were obtained for temperature = 0.7, top_p = 0.95, no frequency or presence penalty, no stopping condition other than the maximum number of tokens to generate, max_tokens = 300.
- LLaVA generations were obtained for temperature = 1.0 and top_p = 1.0, no penalties, and max_tokens = 128. The reason for using a slightly higher temperature and top_p is to have more consistent outputs. In our initial experiments, LLaVA did not perform as well in terms of following instructions at the same temperatue setting as GPT-4V.
- For Grounding DINO, we use `ShilongLiu/GroundingDINO` from Hugging Face and set both box and text thresholds to 0.25 for the grounded box generations.
- For Stable Diffusion, we use `stabilityai/stable-diffusion-2-inpainting` from Hugging Face, and replace the autoencoder with `stabilityai/sd-vae-ft-mse`. We also use a `DPMSolverMultistepScheduler` for speeding up the generation process. We add `"intricate details. 4k. high resolution. high quality."` to the end of our prompt to get high quality images.

Computation budget
- We spent about $800 in total for DALL-E 3 generations. This was funded by a grant from Microsoft Azure.
- We spent about $700 in total for GPT-4V `vision-preview` and GPT-4 `turbo` inference and across all experiments.
- For experiments with LLaVA, Stable Diffusion and Grounding DINO, we used a single instance of a Multi-Instance A100 GPU with 40GB of GPU memory, 3/7 fraction of Streaming Multiprocessors, 2 NVIDIA Decoder hardware units, 4/8 L2 cache size, and 1 node.
- Total emissions for API based models are estimated to be 25 kgCO$_2$ eq, of which 100 percents were directly offset by the cloud provider. Total emissions for our on-premise GPU usage is estimated to be less than 5 kgCO$_2$ eq. Estimations were conducted using the MachineLearning Impact calculator (Lacoste et al., 2019).

Figure 13: We identify more than 18,000 unique cultural artifacts across all countries as part of our second task, and then filter them to find salient ones. This figure shows strongest correlated artifacts for 20 randomly picked countries.
Note that these associations are extracted from LMM generations and may not always be accurate.

Figure 14: We show examples of edits made using our CULTUREADAPT pipeline across 4 different concept classes and 12 pairs of unique source, target combinations to illustrate both cases where our pipeline excels and also where it is limited by the parts it is composed of. For all of these edits, our metric success criteria of $\Delta_1 < 0$ and $\Delta_2 > 0$ is satisfied.

Figure 15: Confusion matrices for GPT-4V and LLaVA on the cultural awareness task for DALLE STREET images.

Figure 16: Confusion matrices for GPT-4V and LLaVA on the cultural awareness task for Dollar Street images.

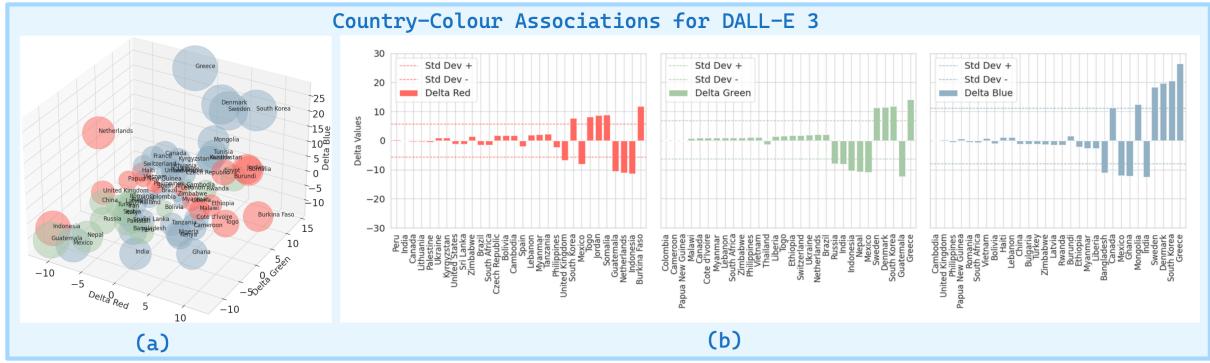Figure 17: Confusion matrices for GPT-4V and LLaVA on the cultural awareness task for MaRVL images.

Figure 18: We explore how countries are distributed on a color spectrum by first calculating a global average RGB vector for DALLE STREET images and then defining deltas along each axes aggregated at the country level. **Takeaway**: We find interesting associations - Greece is strongly correlated with blue, Burkina Faso with red.
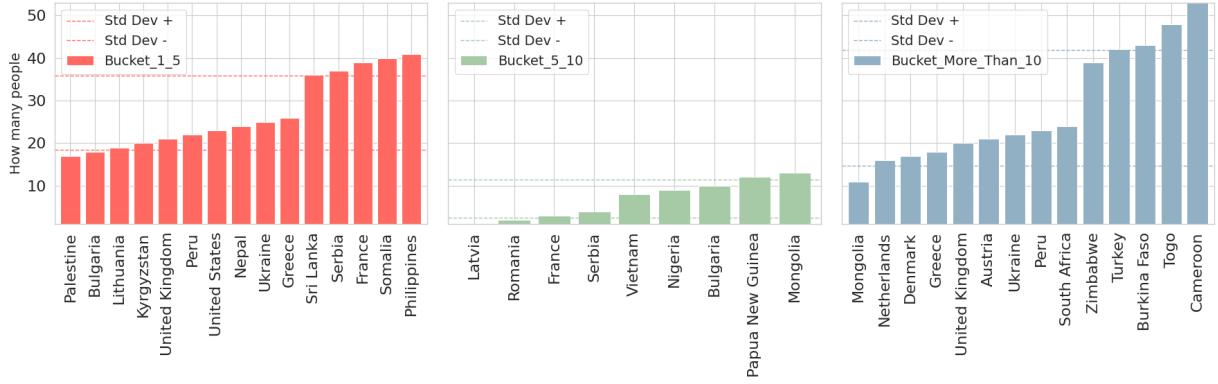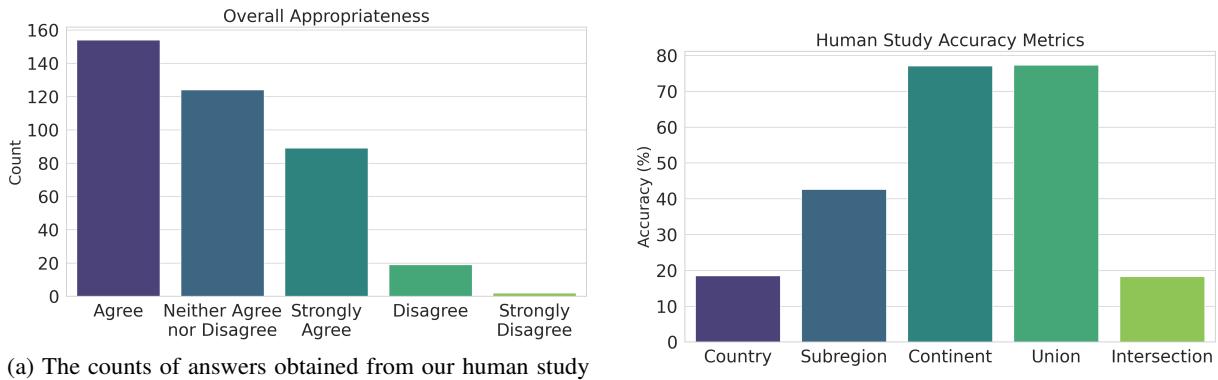

Figure 19: Here, we look at buckets of people counts in DALLE STREET images aggregated at the country level, each of the subplots representing one bucket. **Takeaway**: Counts of people in images may not always accurately reflect population densities of the corresponding countries to scale.

(a) The counts of answers obtained from our human study about the appropriateness of generated images shows that mostly people agree or are neutral, with only 2 disagreements overall.

(b) We measure accuracy for the cultural awareness task at the country, subregion and region level and find increasing performance in this order across our annotator pool.

Figure 20: (a) Appropriateness of generated images as perceived by human study participants. (b) Accuracy of cultural awareness at various geographical levels across our annotator pool for a subset of randomly sampled DALLE STREET images.