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, Fairfax, VA, USA
{amukher6, zzhu20, antonis}@gmu.edu

Abstract

We present a comprehensive three-phase study to examine (1) the cultural understanding of Large Multimodal Models (LMMs) by introducing 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; (2) the underlying implicit and potentially stereotypical cultural associations with a cultural artifact extraction task; and (3) an approach to adapt cultural representation in an image based on extracted associations using a modular pipeline,* CULTUREADAPT*. We find disparities in cultural understanding at geographic sub-region levels with both open-source (LLaVA) and closed-source (GPT-4V) models on* DALLE STREET *and other existing benchmarks, which we try to understand using over* 18,000 *artifacts that we identify in association to different countries. 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. [17] explored how the word has evolved to gain different meanings in different contexts. Recent efforts in natural language processing have seen growing interest in understanding how culture influences language models and human behavior, including language, art, and decision-making [1, 9, 12, 19]. As large multimodal models (LMMs) intersect more with human life, the need for them to comprehend and respect cultural nuances is crucial. Research in this area focuses on model alignment with human values, assessing *cultural awareness*, and exploring *cultural adaptation*; the goal is to modify content that represents one culture or country, often stereotypically, to reflect a different one, to better suit audiences from different cultural backgrounds.

Figure 1. We introduce a large-scale dataset for measuring cultural awareness, an artifact extraction task for implicit cultural associations, and a modular pipeline for culturally adapting images with fine-grained edits.

The challenge of assessing the capability of understanding and leveraging cultural knowledge in LMMs is significant. Prior research has primarily investigated LMMs for cultural awareness[2] by examining their performance on tasks such as region classification from images [3, 25, 31], image-caption matching [20], and cultural image captioning [4]. However, these tasks do not determine whether LMMs respond to cultural cues encoded within their training data or merely identify superficial cultural associations.

To address these gaps, first, we develop a new large-scale **dataset** to assess 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 [10]. Next, we introduce a **task** designed to identify implicit associations between cultures and artifacts

[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.

[19] that LMMs use to distinguish between cultures. Finally, we propose a **cultural adaptation framework** combining multiple generative models in an end-to-end pipeline to adapt images from one cultural context to another by modifying the underlying implicit associations. Our main contributions are as follows:

- **Dataset:** We introduce DALLE STREET, a collection of 9,935 images generated by DALL-E 3, covering 67 countries and 10 cultural concept classes, with more images from underrepresented geographic regions compared to datasets like DOLLAR STREET [28].
- **Benchmark:** We measure how well humans and multimodal large language models (open- and closed-source) can identify countries for images in DALLE STREET and two other datasets (DOLLAR STREET, MARVL) to study disparities in performance at the geographic subregion level for a diverse group of concepts and countries.
- **Task:** We introduce a task for identifying implicit associations by extracting cultural artifacts from images and filtering them to discover associations that frequently co-occur for each country.
- **Framework:** We propose a modular end-to-end pipeline, CULTUREADAPT (Figure 9), to adapt an image to a target culture by updating identified implicit cultural associations in it using diffusion-based inpainting. We evaluate results by introducing a CLIPScore-based metric.

2. Data

We study around 20k images from three datasets (data statistics in Appendix) covering a wide variety of cultural concepts, economic ranges, and data sources: (a) DALLE STREET: synthetic, DALL-E 3 [23] generated; (b) DOLLAR STREET: natural, collected photographs; and (c) MARVL: web-scraped under native speaker guidance. All the datasets are available under CC BY-SA 4.0 license.

DALLE STREET We use 10 concept classes (car, family snapshots, front door, home, kitchen, plate of food, cups/mugs/glasses, social drink, wall decoration, and wardrobe), 19 geographical regions, and 67 countries (more details in Appendix), similar to DOLLAR STREET. We generate 1024×1024 images with DALL-E 3 [23] in two styles - vivid (hyper-realistic) and natural (realistic), prompting with a template (prompt in Appendix) that specifies the concept class and target country. At least 10 images are sampled per country-concept combination, yielding 9,935 images after filtering out content policy violations from API calls. A qualitative study on a randomly sampled subset of around 300 generated images and 14 participants (Table 1a) shows most annotators agree that the images reflect stereotypical country representations, with less than 1% of images receiving strong disagreement. When unsure, participants tend to neither agree nor disagree about the "appropriateness" of an image. Our annotators also mark *visual cues* in these images, including both explicit (e.g., flags) and implicit (e.g., color schemes) cultural artifacts. This feedback motivates our artifact extraction task and its use in our cultural adaptation framework.

DOLLAR STREET [28] This is a dataset of photos of objects and scenes collected by professional and volunteer photographers. We filter it for images that do not contain multiple labels, classes that do not cover images for all regions, and classes with subjective naming. Then we refer to our group of annotators from diverse backgrounds (details in Appendix) to choose the top 10 categories by the method of collaborative labeling [6], where 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. Our data from this source includes 4,137 images from 63 countries, 19 geographical regions, across 10 concept classes.

MARVL [20] This is originally a dataset for validation of statements about image pairs curated by native speakers in five languages: Indonesian, Mandarin Chinese, Swahili, Tamil, and Turkish. We assign country and region labels (based on corresponding languages) to get 4,914 images across 5 geographical regions.

3. Human Study

We perform multiple human studies to support different experiments throughout the paper. Our 14 annotators, aged 25–40, are from diverse demographic backgrounds and currently reside in the USA. They have lived in or are native to over 10 countries and all four major regions in our dataset. Approximately 40% identify as female, while the rest identify as male. The group comprises 40% graduate students, with the remainder being working professionals or computer science faculty. Annotators were recruited from computer science labs and diverse social networks, and all provided consent for their data to be used for research. The studies qualify for IRB exemption as no PII is involved. More details for the annotation process are included in the Appendix along with interfaces used for each study.

4. Cultural Awareness (Task 1)

We compare performances of humans, LLaVA and GPT-4V on DALLE STREET, DOLLAR STREET and MARVL in terms of their ability to predict the country given an image. Overall, we find performances vary across sub-regions, but both LLaVA and GPT-4V perform better than humans.

Appropriateness Category	Percentage (%)
Agree	40.79
Neither Agree nor Disagree	34.54
Strongly Agree	19.41
Disagree	4.61
Strongly Disagree	0.66

(a) Results from our human study on appropriateness for generated images show that most participants agree or are neutral, with less than 1% expressing strong disagreement.

Geographical Level	Accuracy (%)
Country Level	22.16
Subregion Level	47.63
Continent Level	77.77
Union Accuracy	78.03
Intersection Accuracy	21.91

(b) Accuracy for the cultural awareness task improves from country to subregion to region level.

Table 1. (a) Perceived appropriateness of generated images by human participants. (b) Cultural awareness accuracy at different geographical levels for a subset of DALLE STREET images.

4.1. Methods

Given an input image, we prompt `LLaVA-NeXT` [21] and GPT-4V `vision-preview` [24] in a zero-shot generative setting by asking an open-ended question without providing answer choices: *Predict the geographical region represented in the image, as per the United Nations geoscheme* [30]. We use this geoscheme for three reasons: (1) models have a higher refusal rate when queried with specific country labels; (2) the geoscheme is included in most LLM pre-training data (English Wikipedia); and (3) it enables structured parsing of open-ended generations. We focus on countries with stable geographic classifications to prevent errors due to geoscheme updates.

Evaluation Metrics We process generated text to map it to one of the geographical sub-regions or a policy violation case and then compare it with true labels by mapping country information to geographical regions, which gives us classification accuracy as a quantitative metric for measuring success. Since this is a typical classification problem, we also inspect the confusion matrix to locate sub-regions with more errors.

Economic disparities For DOLLAR STREET, 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 region-specific normalized income quartiles.

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

Table 2. LLaVA matches or outperforms GPT-4V on two of three datasets. Human accuracy on a DALLE STREET subset is 47.63%.

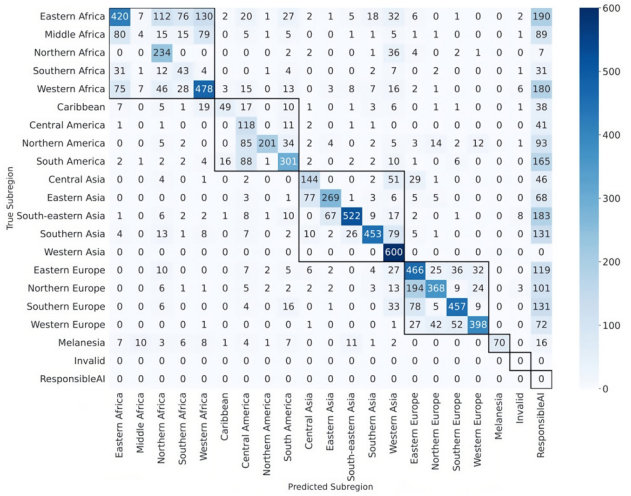

Figure 2. Confusion matrices for GPT-4V on the cultural awareness task for DALLE STREET images. Accurate responses match the true subregion. Special labels include Invalid (no match or incomplete) and ResponsibleAI (policy violation). **Takeaway:** The model performs well, with a strong leading diagonal and 100% accuracy for Western Asia (which covers Iran, Jordan, Lebanon, Oman, Palestine, Turkey).

Human Baseline 14 annotators label images at the country, subregion, or continent level, with 1 to 5 guesses per image, accounting for varying familiarity with different regions. We first evaluate exact match accuracy at the country, then subregion, and finally continent levels. We also consider two cases: union (the correct answer appears at any level) and intersection (the correct answer appears at all levels). Table 1b shows that while country-level accuracy is low, it nearly doubles at each broader geographic level.

4.2. Results

We find similar trends across datasets for both models, with some variations across subregions.

Overall comparison LLaVA performs as good as GPT-4V on DOLLAR STREET and outperforms it significantly on the DALLE STREET images (Table 2). This indicates that LLaVA *may* have implicitly learned *stereotypical* associations between regions and concepts because the DALLE STREET images include such associations (Section 5). However, on the MARVL data, LLaVA performs

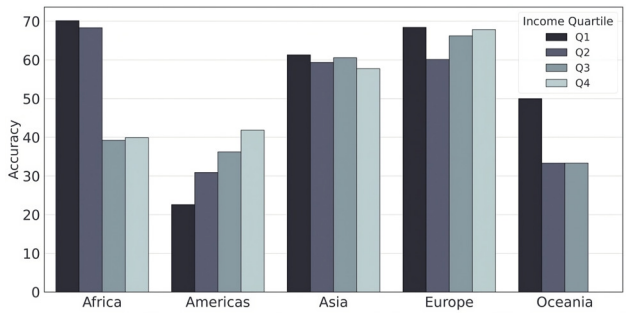

Figure 3. We normalize DOLLAR STREET income data into region-specific quartiles and plot accuracies for GPT-4V. **Takeaway:** Lower income quartiles (Q1, Q2) show higher accuracy in Africa and Asia, while higher quartiles (Q3, Q4) perform better in the Americas. In Europe, accuracy is similar across all quartiles.

about as good as random guessing. This *may* be because MARVL covers specific indigenous concepts that the model may not have seen before.

Subregion Level Analysis GPT-4V performs well on DALLE STREET, with a strong leading diagonal indicating many correct predictions (Figure 2). However, it often provides no answer due to content policy violations. Notably, both models accurately predict all Western Asian images (figures in Appendix). LLaVA tends to default to South America for incorrect answers, while GPT-4V defaults to policy violations. Similar trends are observed in other datasets (figures in Appendix).

Economic Disparity Using income data from DOLLAR STREET, we group results by normalized income quartiles across Africa, Asia, Americas, Europe, and Oceania (GPT-4V - Figure 3, LLaVA results in Appendix). Performance is better for lower-income quartiles in Africa and Asia, while it improves with higher-income groups for the Americas. This *might* indicate that the model defaults to associating Africa with poorer contexts and America with wealthier ones. For Europe, performance remains consistent across all quartiles.

5. Extracting Implicit Associations of Cultures and Artifacts (Task 2)

We propose to extract cultural artifacts (material items) from the generated images to identify the implicit associations the models may use for Task 1. We find associations that are usually stereotypical (and not truly representative) for the relevant countries. This provides a better understanding of the models, enabling us to develop our approach for Task 3 for cultural adaptation of images.

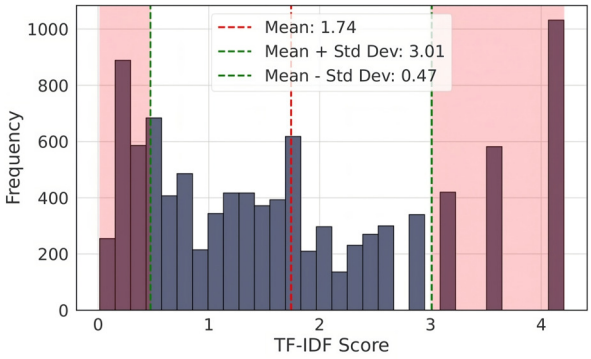

Figure 4. We score each artifact based on its likelihood of co-occurrence for a country. Scores outside the mean and standard deviation range (red) indicate frequent co-occurrences, representing implicit (potentially *stereotypical*) associations.

5.1. Methods

We use GPT-4V `vision-preview` for open vocabulary object detection [32], using a detailed prompt (prompt in Appendix) to extract information about concept classes in DALLE STREET images, including descriptions, color,[3] and person count. GPT-4V's strong instruction-following capabilities result in nearly perfect JSON outputs, which we lightly post-process and summarize using GPT-4 `turbo` (prompt in Appendix). This process yields many unique associations for each country. In our initial experiments, GPT-4V significantly outperformed LLaVA, hence we report GPT-4V results.

Salient associations To identify *salient* artifacts that appear more frequently in one country than others (potentially *stereotypical* associations), we follow an approach similar to [15]: 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, to calculate a `tf-idf` score by multiplying term frequency and the inverse of the document frequency. We then perform a qualitative evaluation of outliers from the distribution of these scores.

Evaluations Extracting cultural artifacts is a novel task with no prior work or established metrics for quantitative evaluation. We explored several approaches, but each had limitations. As discussed before, our DALLE STREET validation includes visual cues marked by annotators, consisting of names and bounding boxes. A simple metric could compare these names with objects extracted by GPT-4V, but annotators often provided descriptive labels (e.g., "*Mongol-looking structure*") instead of specific terms (e.g., "*yurt*"), making semantic matching challenging. Bounding boxes

[3]This refers to object and person appearance, not race.

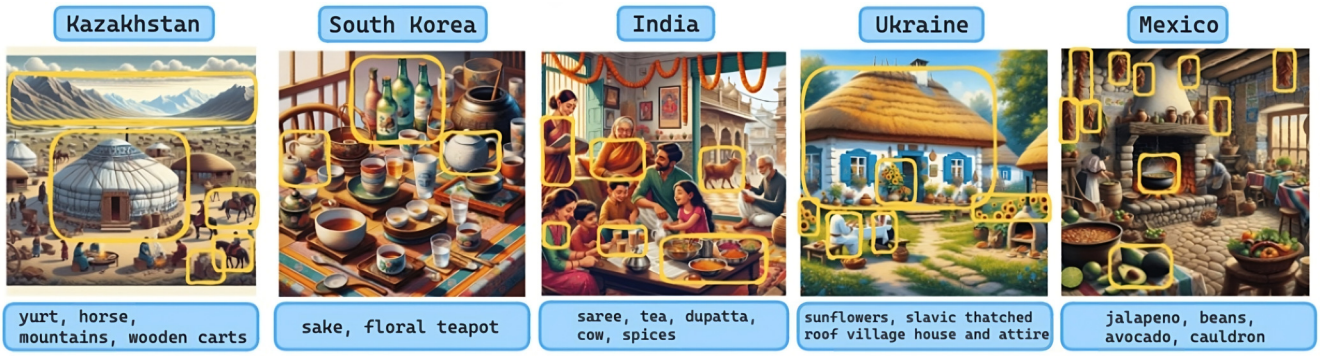

Figure 5. We identify more than 18,000 unique cultural artifacts across all countries as part of our second task and then filter them to find salient ones. This figure shows the strongest correlated artifacts for 5 randomly picked countries. We include more such examples of country-object associations in the Appendix.

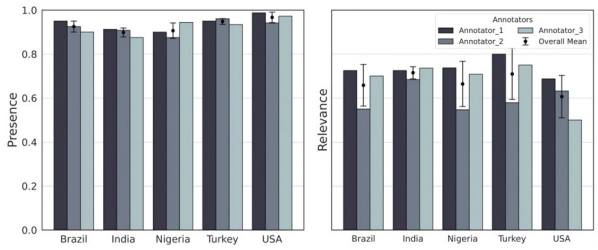

Figure 6. Human validation of a random sample of generated artifacts shows (1) low hallucination (high presence scores) and (2) more than a random fraction of artifacts usually display cultural relevance.

were also imprecise, often covering multiple cues, limiting their usefulness. Given these challenges, we focus on salient association identification and qualitative evaluations. We sampled 100 images and asked annotators to verify if (1) a cultural artifact is present and (2) if it is indeed culturally relevant. Over 90% of the artifacts are deemed to be present, but only 60–70% are marked relevant, indicating that not all salient associations are necessarily stereotypical (Figure 6). We also compared human-labeled visual cues with model outputs, finding many similarities (examples in Appendix). Future work could develop large-scale annotations for reference-based quantitative metrics.

Color Associations for Countries We calculate the mean RGB vector for each DALLE STREET image, then average them to get a global mean vector. We repeat this process at the country level and measure each country's distance from the global mean, identifying colors more strongly associated with specific countries across three (RGB) dimensions.

Counting the Number of People We observed that DALL-E 3 generates varying population densities across countries for identical prompts. To explore this, we use an object detection prompt (included in Appendix) to count

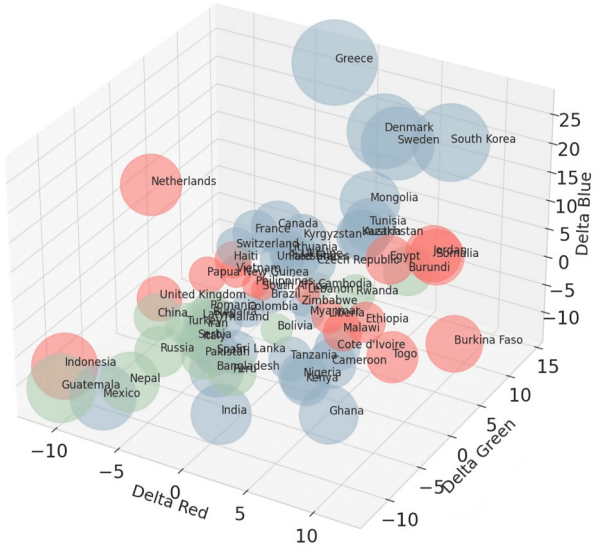

Figure 7. We explore how countries are distributed on a color spectrum by first calculating a global average RGB vector and then defining deltas along each axes aggregated at the country level. **Takeaway**: We find interesting associations - Greece is strongly correlated with blue, and the Netherlands with red/orange.

people in each image, split into three buckets: less than 5, 5 to 10, and more than 10 people. We process related terms (e.g., people, person, man, woman) and aggregate country-level statistics to analyze population density distributions in the generated images. Annotators validate a random sample of these counts, finding general consistency, though they often do not reflect real-world population densities accurately.

5.2. Results

We analyze all the DALL-E 3 generated images to discover implicit associations between countries and cultural artifacts. Our method effectively surfaces implicit associations but also highlights challenges with stereotype reinforcement and demographic inaccuracies.

Artifact Associations Our analysis shows that some cultural artifacts are strongly tied to specific countries, with certain artifacts exceeding one standard deviation from the mean `tf-idf` score (Figure 4). While these associations can offer cultural insights, they often reflect negative stereotypes, such as the over-representation of palm trees for tropical regions, which overlooks broader diversity. This suggests that while our method captures implicit associations, it also underscores the need to refine models to avoid reinforcing such stereotypes.

Color Associations Models not only associate cultural artifacts with countries but also colors. In Figure 7, the `RGB` delta values for several countries in DALLE STREET fall outside the standard deviation. For instance, Greece is strongly associated with blue and the Netherlands with red, *likely* due to recurring elements like blue seas in Greek images and red tulip fields in Dutch ones.

People-Count Associations Most images fall into the extreme buckets (less than 5 or more than 10 people), with few in the middle (Figure 8, more counts in Appendix), often misrepresenting actual population densities. In general, African countries tend to fall into high person-count buckets, while European ones are on the low person-count end, *possibly* reflecting the model's perception of collectivist versus individualistic societies.

6. Cultural Adaptation (Task 3)

We propose a method to edit a given image for a target culture by modifying the detected salient implicit associations between countries and artifacts.

6.1. Methods

Recent works on cultural translation [8, 16, 18] define different approaches for adapting images or text from one culture to another.

CULTUREADAPT Our pipeline (Figure 9) uses GPT-4V for open-vocab object detection to extract implicit cultural associations from the source image. Next, we use Grounding DINO [22] to ground these objects with bounding boxes, which we convert into masks. We then create an inpainting prompt by adding the target country to the list of detected objects and use Stable Diffusion 2 `inpainting` [29] to edit[4] the image by filling in the masked pixels. We evaluate our method using CLIPScore [13] to measure image-country similarity (treating the name of the country as the caption) and cosine similarity of DINO-ViT [5] embeddings for measuring structural preservation. Our choice of metrics is inspired from [16].

[4]We acknowledge that the resulting image may represent common stereotypes because the underlying artifacts may be implicitly associated with a stereotypical view of the country.

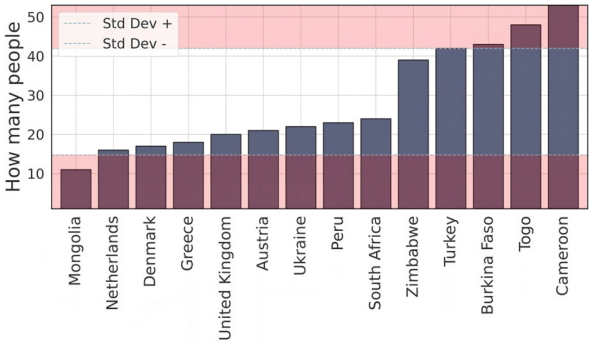

Figure 8. We explore people count associations being made by DALL-E 3 and GPT-4V, and show some selected countries where generated images in DALLE STREET have more than 10 detected people. **Takeaway:** African countries typically fall into high-person-count buckets in our experiments.

Modularity Our pipeline is modular, allowing components to be easily swapped to improve performance over time. For example, we could use Tag2Text [14] or RAM [33] for image captioning, extract object tags, and then use Grounded SAM [27] for bounding boxes and segmentation masks. These can then be passed to inpainting models like Stable Diffusion 3 [2] or MimicBrush [7].

Baseline The closest related work is by [16], which proposes three methods for image editing. We compare our approach with their two most relevant methods, providing qualitative examples (Figure 9) and quantitative evaluations. Additionally, we conduct a qualitative study assessing human preferences for layout preservation and cultural relevance changes (Figure 11) for country pairs common to both studies. We also perform extensive statistical testing to compare structural similarity and editing success, using CLIPScore-based metrics, for both approaches.

Evaluation Let image I_1 correspond to country C_1 and I_2 be its adaptation for country C_2. The CLIPScore for an image-country pair is denoted as $S(I, C)$. We define two deltas:

$$\Delta_1 = S(I_2, C_1) - S(I_1, C_1) \quad (1)$$

$$\Delta_2 = S(I_2, C_2) - S(I_1, C_2) \quad (2)$$

If $\Delta_1 < 0$ and $\Delta_2 > 0$, it indicates successful adaptation, where I_2 is closer to C_2 than C_1. Our primary metric, M_1, tracks how often this condition is met. A secondary metric, M_2, compares $\Delta_2 - \Delta_1$ to evaluate success, though it's less ideal for cases where Δ_1 is positive. We also use

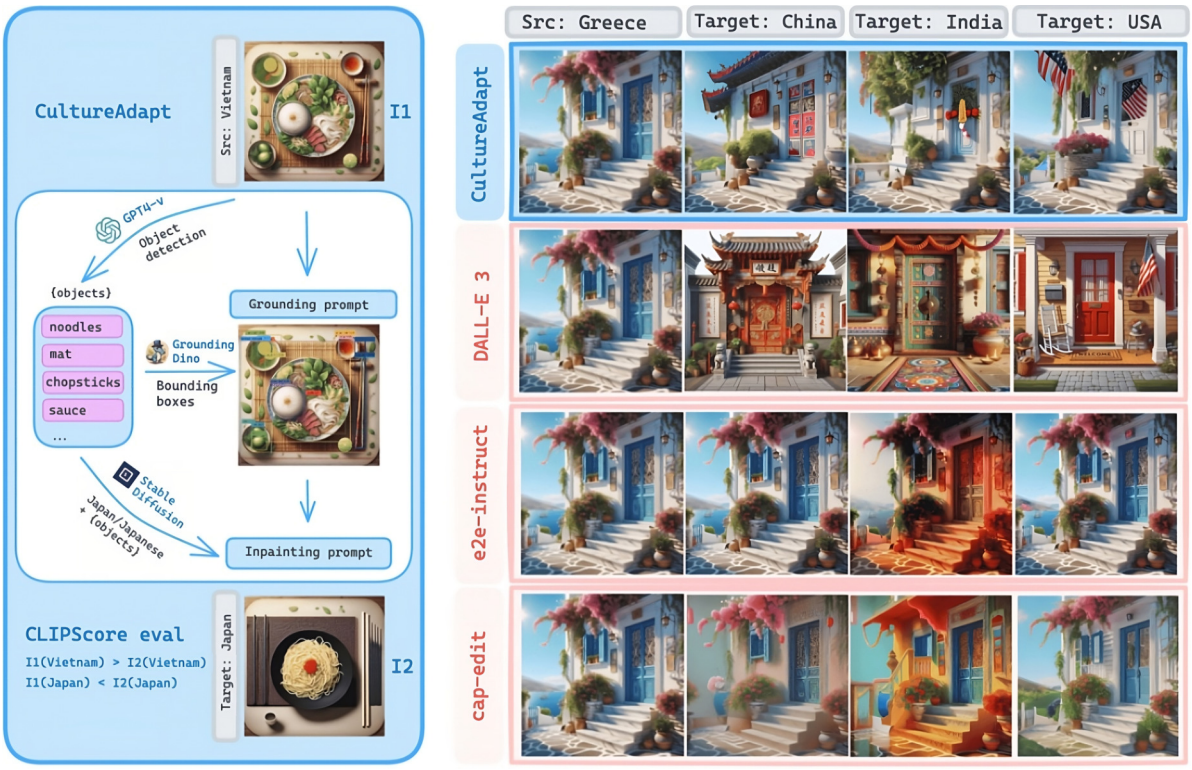

Figure 9. The CULTUREADAPT pipeline first identifies objects in an image using GPT-4V and grounds them with bounding boxes from Grounding DINO. These masks, along with a prompt containing object names and the target country, are used by Stable Diffusion for inpainting and cultural adaptation. **Takeaway:** Unlike DALL-E 3 (which generates a completely new image) or other editing methods [16] that struggle with cultural adaptation, CULTUREADAPT makes precise, meaningful edits. (Additional examples in the Appendix.)

another metric, `SSIM`, to measure structural similarity between I_1 and I_2 via DINO-ViT embeddings. A good edit performs well across editing and similarity metrics.

6.2. Results

Empirical findings indicate that our method works well both qualitatively and quantitatively.

Qualitative comparisons In Figure 9, we show CULTUREADAPT applied to a randomly selected image, for adapting from Greece to China, India, and the USA. By visually contrasting with results from DALL-E 3, we see that our approach preserves structural similarity better, and by comparing with the baselines `e2e-instruct` and `cap-edit`, we demonstrate our method's ability to make meaningful edits. Our Appendix includes more examples.

Comparison with baseline We compare CULTUREADAPT with `cap-edit` (Figure 10) for 20 country pairs (complete results in Appendix). Both methods produce images similar to the source, but CULTUREADAPT edits are overall more culturally relevant.

To test this empirically, we compare similarity scores using the Wilcoxon signed-rank test (Shapiro-Wilk: $p = 5.99 \times 10^{-39}$, indicating non-normality). `cap-edit` has a slightly higher mean similarity 0.97 than CULTUREADAPT 0.94, with a statistically significant difference ($p = 6.02 \times 10^{-215}$) for ($\alpha = 0.05$). The bootstrapped 95% confidence interval for the mean difference is [0.0298, 0.0333], which does not include 0, supporting this result.

For editing metrics, CULTUREADAPT outperformed `cap-edit` in a statistically significant way. For M_1, 54% of samples met the condition versus 50% for `cap-edit` ($p = 7.43 \times 10^{-5}$, McNemar's test). For M_2, CULTUREADAPT had a higher mean score (3.11 vs. 2.68), with the difference significant per the Wilcoxon test ($p = 4.39 \times 10^{-11}$), non-normality confirmed by the Shapiro-Wilk test ($p = 1.23 \times 10^{-16}$). The bootstrapped 95% confidence interval for M_2 differences supports our conclusion ($[-0.560, -0.307]$). In a human study of 100 images, 3 participants rated both methods equally preferable on structure preservation and cultural relevance (Figure 11).

Error Analysis We identify two common error modes (Figure 12): (1) multiple similar objects to be edited in the source image lead to masks covering a significant portion of the image, often in an overlapping manner, resulting in major changes as the diffusion process needs to in-

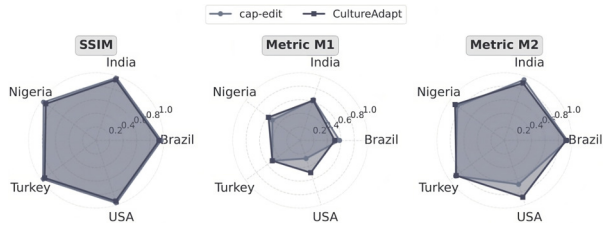

Figure 10. Our method performs better than other approaches in terms of editing metrics while still maintaining comparable structural similarities.

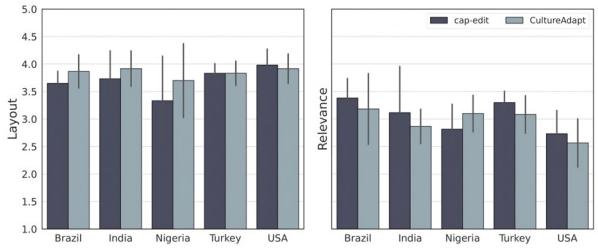

Figure 11. When asked to compare editing outputs from CULTUREADAPT to the existing baseline `cap-edit`, our annotators have nearly similar preferences for both methods in terms of (1) the ability of the methods to maintain layout similarity and (2) cultural relevance of the edited image to the target country.

paint from scratch, and (2) when editing realistic DALL-E 3 images, the edited objects sometimes lose realism. Replacing Grounding DINO with segmentation models (e.g., Grounded SAM) helps mitigate the first issue, but the second remains an open problem, as noted by others [11]. Other less obvious issues include not maintaining the correct count or orientation of objects and not generating human faces correctly, as the underlying diffusion model is trained with a privacy filter.

7. Related Work

The growing interest in culturally aware NLP has inspired various aspects of our research.

[18] propose a data augmentation approach using semantic graphs to enhance cultural components in captions. However, their method often results in inconsistencies at object boundaries when cultural artifacts are copied and pasted into images. Similarly, [16] formalize the task of image transcreation, but their pipelines can produce images that differ significantly from the source or not at all. Our CULTUREADAPT pipeline maintains image coherence by generating semantic masks with bounding boxes and using diffusion-based inpainting.

[26] calculate co-occurrence statistics from image features, and [15] assign importance scores to attributes that frequently co-occur for identity groups. We build on these ideas to identify salient cultural artifacts likely to co-occur

Figure 12. Common error cases with our method occur when there are multiple objects in the image or when the image is generated in a realistic style with DALL-E 3.

for a given country. [19] provide a taxonomy for culturally aware NLP, from which we adopt terminology (cultural *artifacts* and *adaptation*). [25] and [11] both explore the two real-world datasets, MARVL and DOLLAR STREET, specifically developing reliable metrics to measure geo-localization and object consistency across regions.

8. Conclusion

This study addresses the critical need for cultural awareness in Large Multimodal Models 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 benchmarking of LMMs on cultural awareness tasks. Further, we introduce an artifact extraction task to identify over 18,000 artifacts that co-occur frequently with these countries, revealing significant insights into the implicit cultural associations encoded in these models. We also propose CULTUREADAPT, a pipeline 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 toward improvement in cultural representation.

Acknowledgements

We are grateful to the anonymous reviewers for their constructive feedback. This project was generously supported by the National Science Foundation under grant IIS-2327143 and by the Microsoft Accelerate Foundation Models Research (AFMR) grant program. Computational resources for experiments were provided by the Office of Research Computing at George Mason University (URL: https://orc.gmu.edu) and funded in part by grants from the National Science Foundation (Awards Number 1625039 and 2018631).

References

[1] Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Ashutosh Dwivedi, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. Towards measuring and modeling "culture" in llms: A survey. *ArXiv preprint*, abs/2403.15412, 2024. 1

[2] Stability AI. Stable diffusion 3 released. https://stability.ai/news/stable-diffusion-3, 2024. Accessed: September 5, 2024. 6

[3] Abhipsa Basu, R. Venkatesh Babu, and Danish Pruthi. Inspecting the geographical representativeness of images from text-to-image models. *ArXiv preprint*, abs/2305.11080, 2023. 1

[4] Yong Cao, Wenyan Li, Jiaang Li, Yifei Yuan, Antonia Karamolegkou, and Daniel Hershcovich. Exploring visual culture awareness in gpt-4v: A comprehensive probing. *ArXiv preprint*, abs/2402.06015, 2024. 1

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9630–9640. IEEE, 2021. 6

[6] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 2334–2346, New York, NY, USA, 2017. Association for Computing Machinery. 2

[7] Xi Chen, Yutong Feng, Mengting Chen, Yiyang Wang, Shilong Zhang, Yu Liu, Yujun Shen, and Hengshuang Zhao. Zero-shot image editing with reference imitation. *ArXiv preprint*, abs/2406.07547, 2024. 6

[8] Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. Massively multi-cultural knowledge acquisition & lm benchmarking. *ArXiv preprint*, abs/2402.09369, 2024. 6

[9] Xiao Ge, Chunchen Xu, Daigo Misaki, Hazel Rose Markus, and Jeanne L Tsai. How culture shapes what people want from ai. *ArXiv preprint*, abs/2403.05104, 2024. 1

[10] Geoguessr. Geoguessr: A geography game. https://www.geoguessr.com, 2024. Accessed: September 5, 2024. 1

[11] Melissa Hall, Candace Ross, Adina Williams, Nicolas Carion, Michal Drozdzal, and Adriana Romero Soriano. Dig in: Evaluating disparities in image generations with indicators for geographic diversity, 2023. 8

[12] 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. Challenges and strategies in cross-cultural NLP. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland, May 2022. Association for Computational Linguistics. 1

[13] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 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, 2021. Association for Computational Linguistics. 6

[14] Xinyu Huang, Youcai Zhang, Jinyu Ma, Weiwei Tian, Rui Feng, Yuejie Zhang, Yaqian Li, Yandong Guo, and Lei Zhang. Tag2text: Guiding vision-language model via image tagging. *ArXiv preprint*, abs/2303.05657, 2023. 6

[15] Akshita Jha, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Rida Qadri, Chandan K. Reddy, and Sunipa Dev. Visage: A global-scale analysis of visual stereotypes in text-to-image generation. *ArXiv preprint*, abs/2401.06310, 2024. 4, 8

[16] Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. An image speaks a thousand words, but can everyone listen? on image transcreation for cultural relevance. *ArXiv preprint*, abs/2404.01247, 2024. 6, 7, 8

[17] A. L. Kroeber. *Culture: A Critical Review of Concepts and Definitions*. The Museum, Cambridge, MA, 1952. Retrieved from https://nrs.lib.harvard.edu/urn-3:fhcl:30362985. Accessed: June 11, 2024. 1

[18] Zhi Li and Yin Zhang. Cultural concept adaptation on multimodal reasoning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 262–276, Singapore, 2023. Association for Computational Linguistics. 6, 8

[19] Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art. *ArXiv preprint*, abs/2406.03930, 2024. 1, 2, 8

[20] Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 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, 2021. Association for Computational Linguistics. 1, 2

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *ArXiv preprint*, abs/2304.08485, 2023. 3

[22] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *ArXiv preprint*, abs/2303.05499, 2023. 6

[23] OpenAI. Dall·e 3 technical report. https://cdn.openai.com/papers/dall-e-3.pdf, 2024. Accessed: June 9, 2024. 2

[24] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, and et al. Gpt-4 technical report. *ArXiv preprint*, abs/2303.08774, 2023. 3

[25] Angéline Pouget, Lucas Beyer, Emanuele Bugliarello, Xiao Wang, Andreas Peter Steiner, Xiaohua Zhai, and Ibrahim Al-

abdulmohsin. No filter: Cultural and socioeconomic diversity in contrastive vision-language models. *ArXiv preprint*, abs/2405.13777, 2024. 1, 8

[26] Haoyi Qiu, Wenbo Hu, Zi-Yi Dou, and Nanyun Peng. Valor-eval: Holistic coverage and faithfulness evaluation of large vision-language models, 2024. 8

[27] 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. Grounded sam: Assembling open-world models for diverse visual tasks. *ArXiv preprint*, abs/2401.14159, 2024. 6

[28] William Gaviria Rojas, Sudnya Diamos, Keertan Ranjan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. In *Neural Information Processing Systems (NeurIPS 2022)*, 2022. 2

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *ArXiv preprint*, abs/2112.10752, 2021. 6

[30] United Nations. Methodology: Standard country or area codes for statistical use (m49), 2024. Accessed: September 5, 2024. 3

[31] Da Yin, Feng Gao, Govind Thattai, Michael Johnston, and Kai-Wei Chang. Givl: Improving geographical inclusivity of vision-language models with pre-training methods. *ArXiv preprint*, abs/2301.01893, 2023. 1

[32] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. 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, 2021. 4

[33] 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. Recognize anything: A strong image tagging model. *ArXiv preprint*, abs/2306.03514, 2023. 6