Beyond Aesthetics: Cultural Competence in Text-to-Image Models

Nithish Kannen[♠], **Arif Ahmad**[‡,*], **Marco Andreetto**[♠], **Vinodkumar Prabhakaran**[♠], **Utsav Prabhu**[♠], **Adji Bousso Dieng**[¶,§], **Pushpak Bhattacharyya**[‡], **Shachi Dave**[♠]

[♠]Google Research, [§]Google DeepMind, [‡]IIT Bombay, [¶]Princeton
Correspondence: {nitkan, shachi}@google.com

Figure 1: Images from a SOTA T2I model demonstrating its lack of cultural diversity: (a) and (b) and cultural awareness: (c) and (d). (a) Images for "*High definition photo of a monument*" lack architectural and global diversity. (b) Images for "*Image of Nigerian dish*" lack the rich diversity in Nigerian cuisine. (c) "*Image of Jagannath Temple from India*" produces an incorrect depiction of the temple. (d) "*Image of Japanese dish Kabayaki*" produces an incorrect and cartoonized photo.

Abstract

Text-to-Image (T2I) models are being increasingly adopted in diverse global communities where they create visual representations of their unique cultures. Current T2I benchmarks primarily focus on faithfulness, aesthetics, and realism of generated images, overlooking the critical dimension of cultural competence. In this work, we introduce a framework to evaluate cultural competence of T2I models along two crucial dimensions: *cultural awareness* and *cultural diversity*, and present a scalable approach using a combination of structured knowledge bases and large language models to build a large dataset of cultural artifacts to enable this evaluation. In particular, we apply this approach to build CUBE (CUltural BEnchmark for Text-to-Image models), a first-of-its-kind benchmark to evaluate cultural competence of T2I models.[2] CUBE covers cultural artifacts associated with 8 countries across different geo-cultural regions and along 3 concepts: cuisine, landmarks, and art. CUBE consists of 1) CUBE-1K, a set of high-quality prompts that enable the evaluation of cultural awareness, and 2) CUBE-CSpace, a larger dataset of cultural artifacts that serves as grounding to evaluate cultural diversity. We also introduce cultural diversity as a novel T2I evaluation component, leveraging quality-weighted Vendi score. Our evaluations reveal significant gaps in the cultural awareness of existing models across countries and provide valuable insights into the cultural diversity of T2I outputs for under-specified prompts. Our methodology is extendable to other cultural regions and concepts, and can facilitate the development of T2I models that better cater to the global population.

[*]Work done while Arif Ahmad was a student researcher at Google Research.
[2] https://github.com/google-research-datasets/cube

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

https://doi.org/10.52202/079017-0439

1 Introduction

Text-to-image (T2I) generative capabilities have advanced rapidly in recent years, exemplified by models such as Imagen 2 (Saharia et al., 2022), and DALLE-3 (Betker et al., 2023). As powerful tools for creative expression and communication, they have the potential to revolutionize numerous industries such as digital arts, advertising, and education. However, their widespread adoption across the globe raises important ethical and social considerations (Bird et al., 2023; Weidinger et al., 2023), in particular, in ensuring that these models work well for all people across the world (Qadri et al., 2023; Mim et al., 2024). While early T2I model evaluations focused on photo-realism (Saharia et al., 2022) and faithfulness(Hu et al., 2023; Cho et al., 2024; Huang et al., 2023), recent work has demonstrated various societal biases that they reflect (Cho et al., 2023a; Bianchi et al., 2023; Luccioni et al., 2024). However, the predominantly mono-cultural development ecosystems of these models risks unequal representation of cultural awareness in them, potentially exacerbating existing technological inequalities (Prabhakaran et al., 2022). While the term "culture" has a myriad definitions across disciplines (Rapport & Overing, 2002), in this paper we focus on cultures formed within societies demarcated geographically through national boundaries (similar to Li et al. (2024c)), rather than cultures defined through organizational or other socio-demographic categories. This focus stems from our aim to assess global disparities in the capabilities of T2I models. Such disparities are shown to perpetuate harmful stereotypes about cultures (Jha et al., 2024; Basu et al., 2023), as well as cause the erasure and suppression of sub- and co-cultures (Qadri et al., 2023), and limit their utility across geo-cultural contexts (Mim et al., 2024). While recent work has focussed on biases and stereotypes these models propagate (Jha et al., 2024; Basu et al., 2023), not much work has looked into how competent these models are in capturing the richness and diversity of various cultures.

Gaps in cultural competence may manifest primarily along two aspects of model generations: (i) *cultural awareness*: failure to recognize or generate the breadth of concepts/artifacts associated with a culture (Figure 1(c) and 1(d)), and (ii) *cultural diversity*: the tendency to adopt an oversimplified and homogenized view of a culture that associates (and generates) a narrow set of concepts/artifacts within that culture (Figure 1(b)) or across global cultures (Figure 1(a)). While the lack of cultural awareness in text to image models has been documented before (Hutchinson et al., 2022; Ventura et al., 2023), a major challenge in effectively assessing it at scale is the lack of resources that have a broad representation of cultural artifacts. Similarly, while dataset diversity has also been identified as an important part of the data-centric AI agenda (Oala et al., 2023) and has been investigated for text (Chung et al., 2023) and image modalities (Srinivasan et al., 2024; Dunlap et al., 2023), there has been limited focus on diversity of model generations (Lahoti et al., 2023), especially for T2I models. While works studying diversity of image generations focus on visual similarity (Hall et al., 2024; Zameshina et al., 2023), we study the diversity of generated cultural artifacts (aka *cultural diversity*).

In this paper, we present CUBE: **CU**ltural **BE**nchmark, a first-of-its-kind benchmark designed to facilitate the evaluation of cultural competence of T2I models along two axes: cultural awareness and cultural diversity. We build this benchmark at the country level (in line with recent works (Jha et al., 2024; Li et al., 2024c)), encompassing eight countries and representing three different concepts of cultural artifacts chosen as concepts of clear visual elements, and hence of importance to T2I models. We employed a large-scale extraction strategy that leverages a Knowledge Graph (KG) augmented with a Large Language Model (LLM) to build a broad-coverage compilation of country-specific artifacts to ground our evaluation. CUBE consists of a) **CUBE-1K** - a carefully curated subset of 1000 artifacts, made into prompts that enable evaluation of cultural awareness (Nguyen et al., 2023), and b) **CUBE-CSpace** - a collection of $\sim$300K cultural artifacts spanning the 8 countries and 3 concepts we consider, with a potential to be scaled to other concepts and countries. Furthermore, we introduce cultural diversity (CD) as a new evaluation component for T2I models, adapting the quality-weighted Vendi score (Nguyen & Dieng, 2024).We detail the CUBE curation process in Section 3, cultural awareness evaluation in Section 4 and cultural diversity evaluation in Section 5. To summarize, our main contributions are:

- A new T2I CUltural BEnchmark (**CUBE**), that assesses the cultural competence of T2I models along two key dimensions: (1) Cultural Awareness and (2) Cultural Diversity. We curate a dataset of 300K cultural artifacts spanning three concepts with a potential to be scaled to other concepts.
- An extensive human evaluation measuring the faithfulness and realism of T2I-generated cultural artifacts across eight countries and three concepts, revealing substantial gaps in cultural awareness.
- A novel T2I evaluation component leveraging the quality-weighted Vendi score that satisfies the desirable properties to assess cultural diversity in T2I models.

https://doi.org/10.52202/079017-0439

Table 1: **Overview of text-to-image benchmarks**. Existing benchmarks focus only on faithfulness and realism as evaluation aspects and overlook the cultural skill. CUBE is the first T2I benchmark that evaluates cultural competence while introducing diversity as an evaluation aspect.

Benchmark	Skill	Evaluation Aspect		
		Faithfulness	Realism	Diversity
DrawBench	Spatial & Object	✓	✓	✗
CC500	Composition (color)	✓	✓	✗
T2I-CompBench	Composition	✓	✗	✗
Tifa160	Spatial	✓	✗	✗
DSG1k	Spatial	✓	✗	✗
GenEval	Object	✓	✗	✗
GenAIBench	Spatial	✓	✗	✗
CUBE	Cultural	✓	✓	✓

2 Related Work

In our discussion of related work, we focus on T2I evaluation and culture in large models.

T2I Evaluation. Inception Score (Salimans et al., 2016) and Frechet Inception Distance (Heusel et al., 2018) focus on similarity of generated images to real ones, also called realism. Metrics like DSG (Cho et al., 2024) and VQAScore (Lin et al., 2024) measure the prompt-image alignment, also called faithfulness. Other metrics such as ImageReward (Xu et al., 2023), PickScore (Kirstain et al., 2023), and HPSv2 (Wu et al., 2023) fine-tune vision-language models on human ratings to better align with human preferences. There have been recent works on bias and fairness evaluation (Feng et al., 2022; Naik & Nushi, 2023; Zhang et al., 2023; Jha et al., 2024) of T2I models. There have also been efforts to build comprehensive evaluation benchmarks aimed at tracking the progress of model capabilities over time, focusing on tasks such as realism, text faithfulness, and compositional abilities. These benchmarks, such as DrawBench (Saharia et al., 2022), CC500 (Feng et al., 2023), T2I-CompBench (Huang et al., 2023), TIFA v1.0 (Hu et al., 2023), DSG-1k (Cho et al., 2024), GenEval (Ghosh et al., 2023), and GenAIBench (Lin et al., 2024) employ diverse prompts and metrics to assess factors such as image-text coherence, perceptual quality, attribute binding, faithfulness, semantic competence, and compositionality, to list a few.

Culture in Language Technologies. NLP researchers have long argued for the need for cross-cultural awareness in language technologies (Hovy & Spruit, 2016; Hershcovich et al., 2022), and built datasets to assess cultural biases in language technologies (Jha et al., 2023; Naous et al., 2023; Seth et al., 2024). There have also been efforts to identify cultural keywords across languages (Lim et al., 2024), extract cultural commonsense knowledge (Nguyen et al., 2023), as well as to generate culture-conditioned content (Li et al., 2024b). Along those lines, CultureLLM (Li et al., 2024a) proposes generating training data using the World Value Survey for semantic data augmentation to integrate cultural differences into large language models.

Culture in Vision. Efforts to understand cultural competence in computer vision technologies are relatively more recent and limited. Basu et al. (2023) explored the geographical representation of under-specified prompts and found that most of them default to United States or India. Dig In (Hall et al., 2024) evaluates disparity in geographical diversities of household objects. SCoFT (Liu et al., 2024) enhances cultural fairness using the cross-cultural awareness Benchmark (CCUB). Recent work also shows that cultural and linguistic diversity in datasets enriches semantic understanding and helps address cultural dimensions in text-to-image models (Ye et al., 2023; Ventura et al., 2023). Proposals for more inclusive model design and dataset development have been made to address cultural stereotypes and Western-centric biases, to better represent global cultural diversity (Bianchi et al., 2023; Liu et al., 2021). Our work contributes to this line of work, where we introduce a large benchmark dataset and associated metrics to assess cultural competence along cultural awareness and cultural diversity in T2I models.

https://doi.org/10.52202/079017-0439

3 Construction of CUBE

Our benchmark aspires to enable reliable, trustworthy, and tangible measurement of text-to-image generative models for two distinct yet complementary behaviors: *cultural awareness* (i.e., the model's ability to reliably and accurately portray objects associated with a particular culture), and *cultural diversity* (i.e., the model's ability to suppress oversimplified stereotypical depiction for an underspecified input that references a specific culture). One of the core prerequisites to meaningfully evaluate these aspects of cultural competence is a broad-coverage repository of cultural artifacts to ground such an evaluation. Inspired by previous work (Jha et al., 2024; Li et al., 2024c), we focus on *geo-cultures* (realized through the lens of national identity) to build such a repository, potentially extendable to other ways of categorizing culture, such as regions, religions, races, etc. We select eight countries from different geo-cultural regions across continents and the Global South-North divide: Brazil, France, India, Italy, Japan, Nigeria, Turkey, and USA. While we acknowledge that this list of countries is necessarily incomplete, and may result in a biased global sampling, future iterations of this work could include a wider range of countries for a more comprehensive evaluation.

Additionally, we focus on distinctive artifacts, i.e., cultural aspects that reference singular real objects with clear visual elements which are commonly held as belonging to a specific country – as opposed to cultural manifestations that are not visualizable (e.g. speech accents) or multifarious (e.g. complex scenes, or unique inter-object relationships). The three artifact categories ("concepts") included here are *landmarks* (prominent and recognizable structures such as monuments and buildings, located in specific countries), *art* (clothing and regional garments or traditional regalia, performance arts, and style of painting, associated at possibly a specific time in history), and *cuisine* (specific dishes and culinary ingredients that are commonly associated with certain countries). In practice, for the art and cuisine categories, we additionally consider "country of origin" as a strong indicator of national association, acknowledging that there may be other factors.

Finally, for each country-concept combination, we aim to construct grounding "concept spaces", which leads to a collection of ∼300K cultural artifacts, which we call **CUBE-CSpace**. This is an extensive compilation of concept space instances, also intended to be used as grounding for diversity evaluation. From this, we create **CUBE-1K**: a much smaller, curated set of the 1000 artifacts across the 8 countries and 3 concepts - selected for relevance and popularity, intended to be used for testing cultural awareness. The country and concept wise split of CUBE-1K is presented in Table 7. In order to build CUBE, we adopt a Knowledge-Base (KB)-augmented LLM approach wherein we use graph-traversal on a pre-existing KB to extract a broad-coverage set of candidate cultural artifacts, followed by a self-critiquing LLM step to iteratively refine the repository.

3.1 CUBE-CSpace

We use WikiData (Vrandečić & Krötzsch, 2014) as the KB to extract cultural artifacts, as it is the world's largest publicly available knowledge base, with each entry intended to be supported by authoritative sources of information. We use the SLING framework[3] to traverse the WikiData dump of April 2024, by first manually identifying *root nodes* (see Table 8), a small seed set of manually selected nodes that represent the concept in question. For example, the node 'dish' (WikiID: Q746549) is identified as a root node for the concept 'cuisine'. We then look for child nodes that lie along the 'instance of' (P31) and 'subclass of' (P279) edges; e.g. 'Biriyani',(Q271555), a popular dish from India, is a child node of 'dish'

Algorithm 1: Cultural Artifact Extraction from Wikidata

Input: Set of root nodes R, Maximum hops H
Output: Set of cultural artifacts A

$A \leftarrow \emptyset, h \leftarrow 0$
while $h < H$ **do**
  $R_{new} \leftarrow \emptyset$
  for $r \in R$ **do**
    $C \leftarrow$ Children of r along nodes (P31) or (P279)
    for $c \in C$ **do**
      if c has property (P495) or (P17) **then**
        $A \leftarrow A \cup \{c\}$
      else
        $R_{new} \leftarrow R_{new} \cup \{c\}$
      end for
    end for
  $R \leftarrow R_{new}$
  $h \leftarrow h + 1$
end while
return A

[3]https://github.com/google/sling

Geo-culture	Concept	Cultural Artifacts
Japan	Cuisine	Ramen, Soba, Sushi, Katsu sandwich
France	Landmarks	Eiffel Tower, Mont Saint-Michel, Palace of Versailles
India	Art/Clothing	Kurta, Lehanga Choli, Dhoti, Patola Saree

Table 2: Examples of cultural artifacts collected in CUBE-CSpace

along the 'instance of' edge. The child nodes that have the 'country-of-origin' (P495) or the 'country' (P17) are extracted at the iteration. We recursively traverse the remaining nodes along the same edge classes in search of child nodes that satisfy these properties. For example 'bread' (Q7802) is a child of 'dish'; since it is a generic food item, it doesn't have the 'country-of-origin' (P495) property. However, 'Filone' (Q5449200) is a child of 'bread' and has 'country-of-origin' (P495) as Italy, which would be extracted at the step. We outline the extraction process in Algorithm 1. In practise, we iterated for H=4 hops and have detailed considerations in Appendix D.4.

Refinement. The above KB extraction process results in ∼500K collection of WikiData nodes, which is expected to have missing and inconsistent entries, owing to the noisy nature of WikiData (Kannen et al., 2023). We use GPT4-Turbo to filter out cultural artifacts that may not necessarily belong to a concept space, taking inspiration from existing self refinement (Madaan et al., 2023) and self critiquing (Lahoti et al., 2023) techniques. Once we filter out the erroneous artifacts, we prompt GPT-4 to fill out popular missing artifacts from the cultural concept, similar to the diversity expansion application in (Lahoti et al., 2023). This filtering and completion process brings down the count to ∼300K entries, which forms the **CUBE-CSpace**. Table 2 presents some examples cultural artifacts extracted by this process.

3.2 CUBE-1K

As T2I models are primarily trained on English image-text pairs (Pouget et al., 2024), we expect them to struggle with visualizing artifacts from non-English-speaking cultures. To this end, CUBE-1K consists of prompts focusing on widely recognized artifacts, reflecting a model's ability to capture mainstream cultural elements. The artifacts in CUBE-1K are a carefully curated subset of CUBE-CSpace. To ensure the inclusion of popular artifacts relevant to each country, we leverage the number of Google search results as a proxy for popularity. Specifically, we employ the Google Search API, utilizing the geolocation feature ('gl' property) to tailor search results to a user located within the target country, thus capturing local popularity. We use this popularity estimate to sample artifacts for CUBE-1K. While search results serve as a useful proxy, we acknowledge they can be noisy, potentially inflated by the presence of popular keywords. Therefore, the final collection undergoes a manual verification process (detailed in Appendix D.4) to ensure relevance of selected artifacts. CUBE-1K consists of 1000 prompts, spanning 8 countries and 3 concepts described above. Table 7 presents the distribution of artifacts across different countries in CUBE-1K. We use prompt templates designed to probe the models for cultural awareness, along with a negative prompt (Hao et al., 2023) to obtain images with desired qualities. Each prompt tests the model's ability to visualize a single artifact. The prompt templates and negative prompt are provided in Table 13.

4 Evaluating Cultural Awareness

To assess the cultural awareness of text-to-image (T2I) models, we leverage prompts from the CUBE-1K dataset. We use traditional T2I evaluation aspects like *faithfulness* (adherence of the generated image to the input prompt) and *realism* (similarity of the generated image to a real photograph) to measure cultural awareness. Conventionally, these are measured using automated metrics like DSG (Cho et al., 2024) and FID (Heusel et al., 2018). However, these prove insufficient for capturing the complexities of *cultural* representation. Existing automated metrics are primarily trained on datasets lacking diverse cultural content and struggle to adequately assess the nuances of cultural elements. Therefore, we introduce a human annotation scheme specifically tailored to measure a model's cultural awareness along the two key dimensions: a) faithfulness and b) realism.

https://doi.org/10.52202/079017-0439

4.1 Human Annotation

In order to evaluate cultural awareness of the T2I models, we asked human annotators questions that are analogous to standard metrics used in T2I evaluation: a) *faithfulness* and b) *realism* also called *fidelity*. Each annotator was presented with the AI-generated image for an artifact and the corresponding description along with the country association, and was asked the following questions:

1. **Cultural Relevance:** Based solely on the image, does the item depicted belong to the annotator's country? (Yes/No/Maybe)
2. **Faithfulness:** If the image is from the annotator's country, how well does it match the item in the text description? (1-5 Likert scale)
3. **Realism:** How realistic does the image look, regardless of faithfulness? (1-5 Likert scale, with optional comment for scores ≤ 3)

We recruited diverse groups of raters from each of the countries we consider. Each rater pool underwent comprehensive training and was also given a "golden set" of examples, as reference. Once training was complete, the raters proceeded to annotate the 1K prompts spanning the three concepts and the eight countries outlined in Table 3. Raters were instructed to focus on both the image and text when evaluating cultural relevance, and solely the image for realism. Detailed guidelines for each criterion (Appendix E), the inter-annotator agreement (Appendix E.1), and the interface used for human annotation (Figure 4) can be found in Appendix.

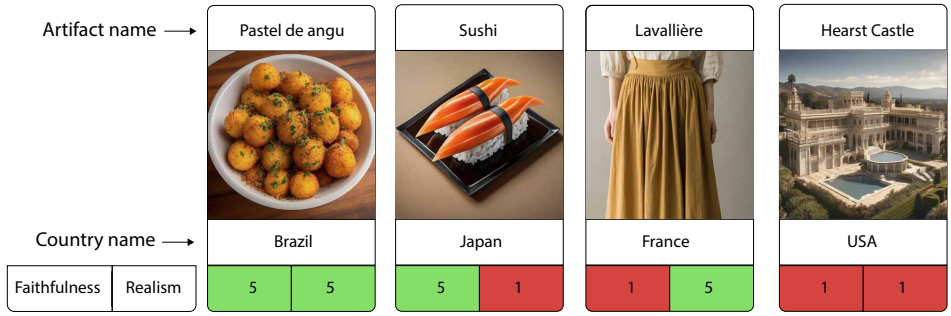


Figure 2: Examples of human evaluation results on cultural awareness for T2I models with high and low scores on faithfulness and realism. More qualitative examples are in Figures 9 and 10.

4.2 Results

Figure 2 presents examples of faithfulness and realism scores for images that were deemed culturally relevant. In 2(a), the model was prompted to generate *Pastel de angu* from Brazilian cuisine and raters gave perfect score for both faithfulness and realism. In contrast, raters gave the lowest score of 1 for both aspects in 2(d), clearly identifying that it is neither faithful nor realistic. Similarly, the image of *Sushi* (2(b)) from Japanese cuisine got an faithfulness score of 5, but realism score of 1 with an observation: "*The fish looks hard and made of glossy plastic.*". Whereas, the image of *lavallière* from France is realistic but not faithful, according to the raters.

Table 3 presents the average consensus scores (and standard deviations) for both faithfulness and realism, as rated for each model across regions and concepts. Both Imagen 2 and SDXL exhibit substantial room for improvement in both faithfulness and realism. Both models achieve relatively lower scores for countries regarded as the Global South (such as Brazil, Turkey, and Nigeria), with this disparity particularly pronounced for faithfulness. On average, in comparison to faithfulness, realism scores are lower across geo-cultures. While Imagen generally outperforms SDXL, exceptions exist, such as art faithfulness in the USA where SDXL scores higher. Table 12 shows the percentage of times our raters from each region deemed the images generated by each model to be culturally relevant (i.e., a yes answer to the first question in Annotation Guidelines in Appendix E) showing non-uniform disparities across models and cultures. This suggests that the cultures marginalized by any particular model may depend on factors such as training data, reiterating the need for such cross-cultural benchmarks.

Concept	Model	India	Japan	Italy	USA	Brazil	France	Turkey	Nigeria
Faithfulness									
Cuisine	Imagen	2.8 ± 1.9	2.4 ± 1.3	2.6 ± 1.5	3.4 ± 1.4	1.9 ± 1.5	3.1 ± 1.5	2.2 ± 1.4	2.7 ± 1.5
	SDXL	2.1 ± 1.7	1.8 ± 0.6	2.2 ± 1.1	3.7 ± 1.3	1.5 ± 1.0	2.8 ± 1.4	1.8 ± 1.1	2.1 ± 1.3
Landmarks	Imagen	3.6 ± 1.8	2.2 ± 0.9	2.6 ± 1.2	3.8 ± 0.6	2.5 ± 1.7	4.0 ± 0.9	3.6 ± 0.9	2.4 ± 0.8
	SDXL	2.7 ± 1.7	2.0 ± 0.7	2.2 ± 0.9	3.3 ± 1.3	2.5 ± 1.6	4.0 ± 0.6	3.0 ± 0.8	1.9 ± 0.7
Art	Imagen	3.5 ± 1.8	2.8 ± 0.9	4.2 ± 1.2	3.3 ± 1.2	2.9 ± 1.8	3.7 ± 1.0	2.5 ± 1.3	2.1 ± 1.4
	SDXL	3.2 ± 1.8	2.0 ± 0.8	3.0 ± 1.2	3.9 ± 1.7	2.2 ± 1.6	3.2 ± 1.0	2.1 ± 1.4	2.0 ± 1.2
Realism									
Cuisine	Imagen	4.2 ± 0.5	2.0 ± 0.8	3.2 ± 0.9	2.2 ± 0.9	4.4 ± 0.7	3.4 ± 0.9	2.4 ± 0.8	3.3 ± 0.9
	SDXL	3.6 ± 1.1	1.4 ± 0.6	2.2 ± 1.3	1.9 ± 0.9	2.1 ± 1.4	2.8 ± 1.4	2.2 ± 0.8	3.3 ± 0.9
Landmarks	Imagen	3.8 ± 1.0	1.7 ± 0.6	2.1 ± 1.4	2.4 ± 1.2	2.5 ± 1.6	3.2 ± 1.2	2.7 ± 1.0	2.9 ± 0.9
	SDXL	3.7 ± 1.0	1.5 ± 0.6	2.7 ± 1.3	2.1 ± 0.8	3.5 ± 0.9	3.9 ± 0.6	2.5 ± 0.8	3.6 ± 0.7
Art	Imagen	3.4 ± 1.4	2.3 ± 0.8	2.6 ± 1.4	1.3 ± 0.6	2.4 ± 1.5	1.9 ± 1.3	1.6 ± 0.7	2.2 ± 1.2
	SDXL	2.8 ± 1.4	1.4 ± 0.9	1.6 ± 1.1	1.4 ± 0.7	1.3 ± 0.6	2.1 ± 1.4	1.6 ± 0.8	3.0 ± 1.0

Table 3: Comparison between Imagen 2 and Stable Diffusion XL (SDXL) for Faithfulness and Realism. Reported score is the average consensus (1 to 5) and the standard deviation among 3 raters for each country. Highlighted cells indicate scores below 3 (light gray) and below 2 (dark gray).

5 Evaluating Cultural Diversity

We seek to assess the cultural diversity of T2I outputs across different seeds as a way to measure the model's intrinsic latent space cultural diversity (Xu et al., 2024). For instance, a model capable of generating a diverse array of cultural artifacts across a range of seeds demonstrates the cultural richness of its learned representations. A more detailed note on our motivation for seed variation is outlined in Appendix H. For this, we focus on *under-specified* prompts (Hutchinson et al., 2022) - prompts that elicit the generation of diverse cultural artifacts (e.g. "Image of tourist landmarks") rather than specific objects (e.g. "Image of Eiffel Tower"). We then seek to answer: *What is the geo-cultural diversity of the generated cultural artifacts for prompts that mention just a concept?*. We further study the within-culture diversity in Appendix K and perform a correlation analysis of cultural diversity with existing metrics in Appendix C.

5.1 Cultural Diversity (CD)

Existing works that focus on visual diversity use LPIPS (Zameshina et al., 2023) and Coverage (Hall et al., 2024), based on image embeddings. However, these metrics are not directly applicable in our case, as the similarity here may be attributed to color, texture, spatial orientation, and other visual aspects of the images. Measuring the cultural diversity of text-to-image (T2I) models requires a approach that accounts for both the variety of generated cultural artifacts and the quality of images generated from text prompts. To address this, we introduce **Cultural Diversity (CD)**, a new T2I evaluation component leveraging the quality-weighted Vendi Score (Nguyen & Dieng, 2024).

5.1.1 Foundation: Vendi Scores

Vendi scores are a family of interpretable diversity metrics that satisfy the axioms of ecological diversity (Dan Friedman & Dieng, 2023; Pasarkar & Dieng, 2023). Vendi score captures the "effective number" of distinct items within a collection, considering both richness (number of unique elements) and evenness (distribution of those elements), and is defined as follows

Definition 5.1 (Vendi Scores) *Let $X = (x_1, \ldots, x_N)$ be a collection of N items. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive semi-definite similarity function, such that $k(x, x) = 1$ for all $x \in \mathcal{X}$. Denote by $K \in \mathbb{R}^{N \times N}$ the kernel matrix whose i, j entry $K_{i,j} = k(x_i, x_j)$. Further denote by $\lambda_1, \lambda_2, \ldots, \lambda_N$ the eigenvalues of K and their normalized counterparts by $\overline{\lambda}_1, \ldots, \overline{\lambda}_N$ where $\overline{\lambda}_i = \lambda_i / \sum_{i=1}^{N} \lambda_i$. The Vendi score of order $q \geq 0$ is defined as the exponential of the Renyi entropy of the normalized eigenvalues of K,*

$$\mathrm{VS}_q(X; k) = \exp\left(\frac{1}{1-q} \log\left(\sum_{i=1}^{N} (\overline{\lambda}_i)^q\right)\right), \tag{1}$$

https://doi.org/10.52202/079017-0439

where we use the convention $0 * \log 0 = 0$.

The order q determines the sensitivity allocated to feature prevalence, with values of $q < 1$ being more sensitive to rarer features and $q > 1$ putting more emphasis on more common features. When $q = 1$, we recover the original Vendi score (Dan Friedman & Dieng, 2023), the exponential of the Shannon entropy of the normalized eigenvalues of K.

5.1.2 Incorporating Quality: Quality-Weighted Vendi Scores

While Vendi scores measure diversity, they treat all items equally without considering individual quality. In the context of T2I, however, it is crucial to account for the quality of the generated images conditional on text prompts. We therefore rely on *quality-weighted Vendi scores* (qVS) (Nguyen & Dieng, 2024) that extends VS to account for the quality of items in a given collection. qVS is defined as the product of the average quality of the items in the collection and their diversity,

$$\mathrm{qVS}_q(X; k, s) = \left(\frac{1}{N} \sum_{i=1}^{N} s(x_i) \right) \mathrm{VS}_q(X; k), \tag{2}$$

where $s(\cdot)$ is a function that scores the quality of the items.

In order to be able to compare different collections of images with different sizes, we normalize qVS by the size of the collection to measure cultural diversity:

$$q\overline{\mathrm{VS}}_q(X; k, s) = \left(\frac{1}{N} \sum_{i=1}^{N} s(x_i) \right) \left(\frac{\mathrm{VS}_q(X; k)}{N} \right). \tag{3}$$

We employ the HPS-v2 metric (Wu et al., 2023) as the $s(\cdot)$ function to score the quality of T2I outputs. HPS-v2, trained on 790k human preferences, provides a quality score $s \in [0, 1]$ for an image conditioned on text prompt, making it suitable proxy to measure image quality in our case. We leave exploration of other quality measures of salience of generated artifacts, for future work.

$q\overline{\mathrm{VS}}$ is minimized to 0 when every element has a quality score of 0, and is maximized to 1, when all elements have a perfect quality (s = 1) and are all distinct from each other ($\overline{\mathrm{VS}}$ = 1)

5.1.3 Desirable Properties of $q\overline{\mathrm{VS}}$

$q\overline{\mathrm{VS}}$ has many desiderata in the context of T2I models: it accounts for similarity and inherits several desirable features of the Vendi scores such as sensitivity to richness and evenness. It also exhibits quality-awareness, duplication scaling and offers flexibility to define kernels that capture different facets of geo-cultural diversity.

Properties. Consider the same setup as in Definition 5.1.

- ⋆ **Quality-awareness.** Denote by $\mathcal{C}_1 = (x_1, \ldots, x_M)$ and $\mathcal{C}_2 = (y_1, \ldots, y_L)$ two collections such that $\mathrm{VS}_q(\mathcal{C}_1; k) = \mathrm{VS}_q(\mathcal{C}_2; k)$. Denote by $s(\cdot)$ a function that scores the quality of an item such that $\frac{1}{M} \sum_{i=1}^{M} s(x_i) \geq \frac{1}{L} \sum_{j=1}^{L} s(y_j)$. Then
$$q\overline{\mathrm{VS}}_q(\mathcal{C}_1; k, s) \geq q\overline{\mathrm{VS}}_q(\mathcal{C}_2; k, s)\ .$$
- ⋆ **Duplication scaling.** Denote by $\mathcal{C} = (x_1, \ldots, x_N)$ a collection of N items. Define $\mathcal{C}'$ as the collection containing all elements of $\mathcal{C}$ each duplicated M times. Then
$$q\overline{\mathrm{VS}}_q(\mathcal{C}; k, s) = M \cdot q\overline{\mathrm{VS}}_q(\mathcal{C}'; k, s)\ .$$
- ⋆ **Kernel generalizability.** Let $k_1(\cdot, \cdot)$ and $k_2(\cdot, \cdot)$ represent two different positive semi-definite similarity functions. Then, given a collection $\mathcal{C} = (x_1, \ldots, x_N)$, the quantities $q\overline{\mathrm{VS}}_q(\mathcal{C}; k_1, s)$ and $q\overline{\mathrm{VS}}_q(\mathcal{C}; k_2, s)$ may capture different aspects of diversity based on the properties of k_1 and k_2.

We state above 3 core properties of $q\overline{\mathrm{VS}}$ that makes it suitable for measuring cultural diversity in T2I models: 1) prioritizes collections with higher-quality items when other factors are equal, 2) penalizes the duplication of elements, and 3) exhibits flexibility in capturing various aspects of diversity through

the selection of an appropriate similarity kernel. The proof of the quality-awareness property is immediate following the definition of $q\overline{\mathrm{VS}}$. See Appendix I for a proof of duplication scaling.

5.2 Experimental Setup

We discuss the experimental pipeline: 1) **Prompting and Seeding:** We calculate $q\overline{\mathrm{VS}}$ for 8 images per prompt, matching the typical number of output images of image-generation APIs. To account for variances in prompt wording as well as seed selections, we report scores averaged over 50 repetitions. 2) **Mapping Generated Images to cultural artifacts:** We map each image to to its most closely resembling artifact from the concept space of the domain.[4] Note that since the prompts focus on global concepts, we obtain the continent, country, and artifact name annotation for each generated image. 3) **Computing Vendi Scores:** With each generated image linked to its closest artifact, we compute the cultural diversity of the generated outputs using the metric defined in Section 5.1. We expand on the details of each of these steps in Appendix J. We details the different kernels to capture different aspects of geo-cultural diversity below.

Kernel definition. With each generated image linked to its closest cultural artifact, we now compute the *cultural diversity* (CD) of the model's output using the definition in Section 5.1. We define a general similarity kernel that allows us to analyze different aspects of geo-cultural diversity:

$$k(x_i, x_j) = w_1 \cdot k_1(x_i, x_j) + w_2 \cdot k_2(x_i, x_j) + w_3 \cdot k_3(x_i, x_j) \tag{4}$$

where $k_1(\cdot,\cdot)$, $k_2(\cdot,\cdot)$, and $k_3(\cdot,\cdot)$ are three distinct kernels measuring different aspects of similarity, and w_1, w_2, w_3 assign weights to each. We define $k_1(x_i, x_j) = 1$ if x_i and x_j have the same continent, and 0 otherwise. Similarly, $k_2(x_i, x_j) = 1$ if the two items share the same country, and 0 if not. Lastly, $k_3(x_i, x_j) = 1$ if the two items represent the same artifact, regardless of geographical origin, and 0 otherwise. To illustrate this flexibility, we present results under different kernel configurations:

- **Continent-level diversity** : $w_1 = 1, w_2 = 0, w_3 = 0$. Considers continent-level similarity.
- **Country-level diversity** : $w_1 = 0, w_2 = 1, w_3 = 0$. Considers country-level similarity.
- **Artifact-level diversity** : $w_1 = 0, w_2 = 0, w_3 = 1$. Only considers distinct artifacts.
- **Hierarchical geographical diversity** : $w_1 = 1/2, w_2 = 1/2, w_3 = 0$. This captures a hierarchical notion of diversity where both continent and country similarities are penalized equally, without explicitly considering individual artifacts.
- **Uniformly weighted diversity** : $w_1 = 1/3, w_2 = 1/3, w_3 = 1/3$.

Models. We evaluate 4 models across closed-source and open-source model types: 1) Imagen 2, 2) Stable-Diffusion-XL, 3) Playground, and 4) Realistic Vision. More details about the model usage and hyperparameters are provided in Appendix J.2.4.

5.3 Results

Results in Figure 8 reveals that when prompted with under-specified prompts mentioning general concepts (Figure 3), current T2I models tend to generate artifacts that lack comprehensive geographical representation. This finding aligns with previous observations (Basu et al., 2023), suggesting a bias towards well-represented and popular countries.

Table 4 presents the results for both the average quality score (q) and the diversity component ($q\overline{\mathrm{VS}}$) across different kernels. Playground and Imagen generally achieve the highest quality scores based on the HPS-v2 metric. As anticipated, models exhibit the lowest diversity for $w_1 = 1, w_2 = 0, w_3 = 0$, which considers only continent-level similarity, due to the limited number of continents. Conversely, $w_1 = 1, w_2 = 0, w_3 = 0$, focusing solely on artifact diversity, yields the highest scores, reflecting the wide array of potential cultural artifacts. In terms of overall performance, Imagen 2 consistently demonstrates the best $q\overline{\mathrm{VS}}$ scores across different kernels for the *Cuisine* and *Art* concepts, whereas

[4]Not all text-to-image generated images perfectly represent real-world cultural entities.

https://doi.org/10.52202/079017-0439

Table 4: Breakdown of the mean quality component (q) and mean diversity component ($q\overline{\text{VS}}$) averaged over 50 repetitions. While all models show relatively low quality scores (as per HPS-v2), Playground (PG) has best quality for *cuisine* and *art* concepts and Imagen-2 (IM) for *landmarks*. Different kernels (w_1, w_2, w_3) capture different aspects of diversity.

	Cuisine				**Landmarks**				**Art**			
	IM	**SDXL**	**PG**	**RV**	**IM**	**SDXL**	**PG**	**RV**	**IM**	**SDXL**	**PG**	**RV**
q ($\rightarrow$)	0.27	0.21	**0.29**	0.27	**0.25**	0.22	0.21	0.23	0.31	0.30	**0.34**	0.33
$\overline{\text{VS}}(w_1, w_2, w_3)$												
$\overline{\text{VS}}$ (1, 0, 0)	**0.32**	0.23	0.24	0.27	0.17	**0.27**	0.23	0.25	**0.23**	0.14	0.18	0.16
$\overline{\text{VS}}$ (0, 1, 0)	**0.59**	0.53	0.51	0.51	0.50	**0.65**	0.34	0.52	**0.42**	0.29	0.37	0.23
$\overline{\text{VS}}$ (0, 0, 1)	**0.91**	0.71	0.82	0.74	0.73	**0.84**	0.58	0.81	**0.72**	0.60	0.51	0.44
$\overline{\text{VS}}$ ($\frac{1}{2}$, $\frac{1}{2}$, 0)	**0.51**	0.44	0.41	0.38	0.42	**0.53**	0.31	0.45	**0.36**	0.24	0.31	0.22
$\overline{\text{VS}}$ ($\frac{1}{3}$, $\frac{1}{3}$, $\frac{1}{3}$)	**0.72**	0.58	0.66	0.59	0.55	**0.66**	0.45	0.52	**0.52**	0.39	0.41	0.30

Table 5: CD scores across models and concepts using various similarity kernels averaged over 50 repetitions. Imagen 2 (IM) performs best for *Cuisine* and *Art*, while SDXL performs best for *Landmarks*. Importantly, even the best scores are low, indicating significant room for improvement in the cultural diversity of T2I outputs.

	Cuisine				**Landmarks**				**Art**			
CD	**IM**	**SDXL**	**PG**	**RV**	**IM**	**SDXL**	**PG**	**RV**	**IM**	**SDXL**	**PG**	**RV**
$q\overline{\text{VS}}(1, 0, 0)$	**0.08**	0.04	0.07	0.07	0.04	**0.06**	0.04	0.05	**0.07**	0.042	0.06	0.05
$q\overline{\text{VS}}(0, 1, 0)$	**0.15**	0.11	0.14	0.13	0.12	**0.14**	0.07	0.12	**0.13**	0.08	0.12	0.07
$q\overline{\text{VS}}(0, 0, 1)$	**0.24**	0.14	0.23	0.20	**0.18**	**0.18**	0.12	**0.18**	**0.22**	0.18	0.17	0.14
$q\overline{\text{VS}}(\frac{1}{2}, \frac{1}{2}, 0)$	**0.13**	0.09	0.12	0.10	0.10	**0.11**	0.06	0.10	**0.11**	0.07	0.10	0.072
$q\overline{\text{VS}}(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	**0.19**	0.12	**0.19**	0.16	0.14	**0.15**	0.09	0.12	**0.16**	0.12	0.14	0.10

SDXL obtains the highest scores for *Landmarks* concept. Table 5 shows the cultural diversity ($q\overline{\text{VS}}$). Note that the scores across the board are still low, remaining far from the maximum score of 1. Current T2I models fall short of representing the true breadth and richness of global cultural diversity. In Appendix C, we study the correlation between the 3 metrics: faithfulness, realism and diversity for cultural prompts. As reported in Table 6, we find weak correlations between diversity-faithfulness (ρ = 0.016) and diversity-realism (ρ = 0.156) in the cultural context. This finding resonates with the Pareto fronts reported in (Astolfi et al., 2024).

6 Discussion

To the best of our knowledge, CUBE is the first large-scale cultural competence benchmark to evaluate T2I models along two crucial dimensions: cultural awareness and cultural diversity. We presented a scalable methodology with a potential to be scaled beyond the 8 countries and 3 concepts considered in this work. Furthermore, we proposed a novel T2I evaluation component: cultural diversity (CD) and measured it using the quality-aware Vendi score. From our investigations so far, one clear finding stands out: there is yet significant headroom for improvement of global cultural competence in the current generation of text-to-image models — both in terms of awareness and diversity. This seems especially true for the Global South, highlighting the urgency of the need for comprehensive and informative cultural competence testing frameworks. Our correlation analysis of the 3 metrics reveals a noteworthy trend: while faithfulness and realism exhibit a moderate positive correlation, suggesting they can be improved in tandem, cultural diversity remains weakly correlated to these metrics. By highlighting existing limitations in cultural competence of T2I models, we believe our work contributes to a critical dialogue surrounding the development of truly inclusive generative AI systems. To benefit the community, we intend to make the CUBE dataset publicly available[5], and encourage its adoption and expansion by the multimodal generative AI community.

[5]https://github.com/google-research-datasets/cube

7 Ethical Considerations

We built a large repository of cultural artifacts with the intended use of evaluation of T2I models. Our approach to build this resource relies partly on automated tools, including LLMs, that have been shown to exhibit various societal biases. Hence, care must be taken in interpreting the results of evaluation using this benchmark. While the CUBE benchmark enables a broad-coverage and flexible evaluation of cultural competence in T2I models, their coverage is still limited by the underlying resources it is built on — namely WikiData (the KB) and GPT-4 Turbo (the LLM). Future work should explore bridging the gaps in coverage through participatory efforts in partnership with communities of people within respective cultures. Furthermore, both CUBE-CSpace and CUBE-1K are intended to be used in evaluation pipelines, rather than training or mitigation efforts.

8 Acknowledgements

We thank Partha Talukdar, Kathy Meier-Hellstern, Caroline Pantofaru, Remi Denton, Susanna Ricco, Hansa Srinivasan, David Madras, Nitish Gupta and Sunipa Dev for their feedback and advice; Lucas Beyer and Xiaohua Zhai for insights and support on use of mSigLIP for auto-evals; Sagar Gubbi and Kartikeya Badola for helpful discussions on the human rating template; Shikhar Vashishth for discussions on the use of WikiData; Preetika Verma for assistance with the SLING framework and Dinesh Tewari and the annotation team for facilitating our human evaluation work. We are grateful to the anonymous reviewers for their insightful reviews and suggestions.

 https://doi.org/10.52202/079017-0439

References

Laura Alonso Alemany, Luciana Benotti, Hernán Maina, Lucía Gonzalez, Lautaro Martínez, Beatriz Busaniche, Alexia Halvorsen, Amanda Rojo, and Mariela Rajngewerc. Bias assessment for experts in discrimination, not in computer science. In Sunipa Dev, Vinodkumar Prabhakaran, David Adelani, Dirk Hovy, and Luciana Benotti (eds.), *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pp. 91–106, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.c3nlp-1.10. URL https://aclanthology.org/2023.c3nlp-1.10.

Pietro Astolfi, Marlene Careil, Melissa Hall, Oscar Mañas, Matthew Muckley, Jakob Verbeek, Adriana Romero Soriano, and Michal Drozdzal. Consistency-diversity-realism pareto fronts of conditional image generative models, 2024. URL https://arxiv.org/abs/2406.10429.

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

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1493–1504, 2023.

Charlotte Bird, Eddie Ungless, and Atoosa Kasirzadeh. Typology of risks of generative text-to-image models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 396–410, 2023.

Carwil Bjork-James. New maps for an inclusive wikipedia: decolonial scholarship and strategies to counter systemic bias. *New Review of Hypermedia and Multimedia*, 27(3):207–228, 2021.

Ewa S Callahan and Susan C Herring. Cultural bias in wikipedia content on famous persons. *Journal of the American society for information science and technology*, 62(10):1899–1915, 2011.

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

Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3043–3054, 2023a.

Jaemin Cho, Abhay Zala, and Mohit Bansal. Visual programming for text-to-image generation and evaluation, 2023b.

Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation, 2024.

John Joon Young Chung, Ece Kamar, and Saleema Amershi. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In *Annual Meeting of the Association for Computational Linguistics*, 2023. URL https://api.semanticscholar.org/CorpusID:259096160.

Dan Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on machine learning research*, 2023.

Sunipa Dev, Jaya Goyal, Dinesh Tewari, Shachi Dave, and Vinodkumar Prabhakaran. Building socio-culturally inclusive stereotype resources with community engagement. *Advances in Neural Information Processing Systems*, 36, 2024.

Lisa Dunlap, Alyssa Umino, Han Zhang, Jiezhi Yang, Joseph E. Gonzalez, and Trevor Darrell. Diversify your vision datasets with automatic diffusion-based augmentation, 2023.

Haiwen Feng, Timo Bolkart, Joachim Tesch, Michael J Black, and Victoria Abrevaya. Towards racially unbiased skin tone estimation via scene disambiguation. In *European Conference on Computer Vision*, pp. 72–90. Springer, 2022.

Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis, 2023.

Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment, 2023.

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, 2024.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation, 2023. URL https://arxiv.org/abs/2212.09611.

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 (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6997–7013, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.482. URL https://aclanthology.org/2022.acl-long.482.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018.

Dirk Hovy and Shannon L Spruit. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 591–598, 2016.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering, 2023.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation, 2023.

Ben Hutchinson, Jason Baldridge, and Vinodkumar Prabhakaran. Underspecification in scene description-to-depiction tasks. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1172–1184, 2022.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation, 2024.

Akshita Jha, Aida Mostafazadeh Davani, Chandan K Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. Seegull: A stereotype benchmark with broad geo-cultural coverage leveraging generative models. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.

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, 2024.

Nithish Kannen, Udit Sharma, Sumit Neelam, Dinesh Khandelwal, Shajith Ikbal, Hima Karanam, and L Subramaniam. Best of both worlds: Towards improving temporal knowledge base question answering via targeted fact extraction. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.),

 https://doi.org/10.52202/079017-0439

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 4729–4744, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.287. URL https://aclanthology.org/2023.emnlp-main.287.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation, 2023.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation, 2023.

Preethi Lahoti, Nicholas Blumm, Xiao Ma, Raghavendra Kotikalapudi, Sahitya Potluri, Qijun Tan, Hansa Srinivasan, Ben Packer, Ahmad Beirami, Alex Beutel, and Jilin Chen. Improving diversity of demographic representation in large language models via collective-critiques and self-voting. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10383–10405, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.643. URL https://aclanthology.org/2023.emnlp-main.643.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Benita Teufel, Marco Bellagente, Minguk Kang, Taesung Park, Jure Leskovec, Jun-Yan Zhu, Li Fei-Fei, Jiajun Wu, Stefano Ermon, and Percy Liang. Holistic evaluation of text-to-image models, 2023.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. Culturellm: Incorporating cultural differences into large language models. *arXiv preprint arXiv:2402.10946*, 2024a.

Huihan Li, Liwei Jiang, Nouha Dziri, Xiang Ren, and Yejin Choi. Culture-gen: Revealing global cultural perception in language models through natural language prompting. *arXiv preprint arXiv:2404.10199*, 2024b.

Huihan Li, Liwei Jiang, Jena D. Huang, Hyunwoo Kim, Sebastin Santy, Taylor Sorensen, Bill Yuchen Lin, Nouha Dziri, Xiang Ren, and Yejin Choi. Culture-gen: Revealing global cultural perception in language models through natural language prompting, 2024c.

Zheng Wei Lim, Harry Stuart, Simon De Deyne, Terry Regier, Ekaterina Vylomova, Trevor Cohn, and Charles Kemp. A computational approach to identifying cultural keywords across languages. *Cognitive Science*, 48(1):e13402, 2024.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation, 2024.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. Visually grounded reasoning across languages and cultures. *arXiv preprint arXiv:2109.13238*, 2021.

Zhixuan Liu, Peter Schaldenbrand, Beverley-Claire Okogwu, Wenxuan Peng, Youngsik Yun, Andrew Hundt, Jihie Kim, and Jean Oh. Scoft: Self-contrastive fine-tuning for equitable image generation. *arXiv preprint arXiv:2401.08053*, 2024.

Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation, 2023.

Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback, 2023.

Jabez Magomere, Shu Ishida, Tejumade Afonja, Aya Salama, Daniel Kochin, Foutse Yuehgoh, Imane Hamzaoui, Raesetje Sefala, Aisha Alaagib, Elizaveta Semenova, Lauren Crais, and Siobhan Mackenzie Hall. You are what you eat? feeding foundation models a regionally diverse food dataset of world wide dishes, 2024. URL https://arxiv.org/abs/2406.09496.

Nusrat Jahan Mim, Dipannita Nandi, Sadaf Sumyia Khan, Arundhuti Dey, and Syed Ishtiaque Ahmed. In-between visuals and visible: The impacts of text-to-image generative ai tools on digital image-making practices in the global south. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–18, 2024.

Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 786–808, 2023.

Tarek Naous, Michael J Ryan, and Wei Xu. Having beer after prayer? measuring cultural bias in large language models. *arXiv preprint arXiv:2305.14456*, 2023.

Quan Nguyen and Adji Bousso Dieng. Quality-weighted vendi scores and their application to diverse experimental design. *arXiv preprint arXiv:2405.02449*, 2024.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023*, pp. 1907–1917, 2023.

Luis Oala, Manil Maskey, Lilith Bat-Leah, Alicia Parrish, Nezihe Merve Gürel, Tzu-Sheng Kuo, Yang Liu, Rotem Dror, Danilo Brajovic, Xiaozhe Yao, et al. Dmlr: Data-centric machine learning research–past, present and future. *arXiv preprint arXiv:2311.13028*, 2023.

Amey Pasarkar and Adji Bousso Dieng. Cousins of the vendi score: A family of similarity-based diversity metrics for science and machine learning. *arXiv preprint arXiv:2310.12952*, 2023.

David Picard. Torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision, 2023.

Mao Po-Yuan, Shashank Kotyan, Tham Yik Foong, and Danilo Vasconcellos Vargas. Synthetic shifts to initial seed vector exposes the brittle nature of latent-based diffusion models, 2023.

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

Vinodkumar Prabhakaran, Rida Qadri, and Ben Hutchinson. Cultural incongruencies in artificial intelligence. *arXiv preprint arXiv:2211.13069*, 2022.

Rida Qadri, Renee Shelby, Cynthia L Bennett, and Emily Denton. Ai's regimes of representation: A community-centered study of text-to-image models in south asia. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 506–517, 2023.

Nigel Rapport and Joanna Overing. *Social and cultural anthropology: The key concepts*. Routledge, 2002.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans, 2016.

Dvir Samuel, Rami Ben-Ari, Simon Raviv, Nir Darshan, and Gal Chechik. Generating images of rare concepts using pre-trained diffusion models, 2023.

 https://doi.org/10.52202/079017-0439

Agrima Seth, Sanchit Ahuja, Kalika Bali, and Sunayana Sitaram. Dosa: A dataset of social artifacts from different indian geographical subcultures. *arXiv preprint arXiv:2403.14651*, 2024.

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

Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning, 2023.

Mor Ventura, Eyal Ben-David, Anna Korhonen, and Roi Reichart. Navigating cultural chasms: Exploring and unlocking the cultural pov of text-to-image models. *arXiv preprint arXiv:2310.01929*, 2023.

Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, sep 2014. ISSN 0001-0782. doi: 10.1145/2629489. URL https://doi.org/10.1145/2629489.

Yixin Wan, Arjun Subramonian, Anaelia Ovalle, Zongyu Lin, Ashima Suvarna, Christina Chance, Hritik Bansal, Rebecca Pattichis, and Kai-Wei Chang. Survey of bias in text-to-image generation: Definition, evaluation, and mitigation, 2024.

Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, et al. Sociotechnical safety evaluation of generative ai systems. *arXiv e-prints*, pp. arXiv–2310, 2023.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis, 2023.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation, 2023.

Katherine Xu, Lingzhi Zhang, and Jianbo Shi. Good seed makes a good crop: Discovering secret seeds in text-to-image diffusion models, 2024.

Andre Ye, Sebastin Santy, Jena D Hwang, Amy X Zhang, and Ranjay Krishna. Cultural and linguistic diversity improves visual representations. *arXiv preprint arXiv:2310.14356*, 2023.

Mariia Zameshina, Olivier Teytaud, and Laurent Najman. Diverse diffusion: Enhancing image diversity in text-to-image generation, 2023.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. Alignscore: Evaluating factual consistency with a unified alignment function, 2023.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023.

Cheng Zhang, Xuanbai Chen, Siqi Chai, Chen Henry Wu, Dmitry Lagun, Thabo Beeler, and Fernando De la Torre. Iti-gen: Inclusive text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3969–3980, 2023.

A Contributions

This paper was the result of close collaboration and teamwork. Nithish worked on the ideation of the dataset extraction and metrics, and implemented the end-to-end pipelines for dataset extraction and evaluation. Arif was part of the explorations and contributed to data cleaning, image generation, and the quality evaluation pipeline under the guidance of Pushpak. Marco participated in the design discussions for datasets and metrics and owned the data analysis of human annotation results. Utsav and Vinod kept us honest on the cultural dimension of this work. Adji contributed to defining how to measure cultural diversity, including how to use qVS to measure it and how to design the similarity kernel. Shachi oversaw the entire project and provided guidance and mentorship to Nithish and Arif. Everyone contributed to paper writing.

B Limitations

Design of this benchmark required many challenging decisions that needed careful thought. The increased coverage of our benchmark as a result of largely automated, scalable approaches - built on curated data sources and existing model capabilities still comes with its own limitations. For instance, existing structured knowledge bases such as WikiData are known to have inherent cultural biases reflecting disparities in global distribution of knowledge production (Callahan & Herring, 2011; Bjork-James, 2021). Therefore, it is important to note that our approach of expanding coverage using existing knowledge bases should be complemented with community based and participatory approaches for richer socio-cultural representation (Alonso Alemany et al., 2023; Dev et al., 2024). Notably, the World Wide Dishes effort (Magomere et al., 2024) builds a dataset of images representing dishes from around the world through a community-led effort that complements our dataset that relies on existing knowledge bases.[6]

Furthermore, even with significant filtering and completion, we expect our data to be noisier than other methods. We have not yet explored the potential utility of our curation method, nor of the dataset itself, to empower other research on evaluating or improving cultural competence in generative models in general. Many aspects of the curation process for this benchmark are automated, due to the large scale of the problem — e.g., we rely on image similarity scores and LLM-based selection flows to ground generated images to a specific cultural artifact. This approach risks ingraining biases that may already exist within these tools into the benchmark construction itself. The process of mapping artifacts to country/continent may also introduce biases since the annotator VLM itself may not be aware of several cultural artifacts around the world. On the other hand, human annotation to measure faithfulness and realism is also quite challenging and subjective, as annotators may not be aware of the multitude of representations of their own culture. Even beyond cultural knowledge, different cultures may also have different standards for realism of their images — which could result in mis-calibrated results obtained from human annotations, making it hard to compare across different cultures (e.g., see Table 14).

We acknowledge that our results are susceptible to such errors stemming from both the subjective nature of human annotations (for faithfulness/realism), as well as issues in VLM annotations (for diversity). Nevertheless, the evaluation methods and frameworks we introduce in this work hold significant relevance, as we move towards training multicultural models with globally diverse datasets (Pouget et al., 2024) in our path to equitable representation in generative AI. Finally, we use a narrow definition of culture, defined in terms of geo-political boundaries such as countries and continents in our kernel definition for diversity. However, culture is a more complex concept — countries are rarely monolithic in terms of cultures, and cultural scopes may often transcend geo-political boundaries. Future work could investigate applying our metric to other finer-grained definitions of cultural groups.

C Correlation Analysis: Cultural Diversity-Faithfulness-Realism

In this section, we investigate the correlations between our three key metrics: faithfulness, realism (Table 3), and diversity (Figure 6) across different geo-cultures, focusing on the Imagen model. Our analysis reveals a positive correlation between faithfulness and realism ($\rho = 0.400$), as shown in Table 6 for cultural prompts. This suggests that images judged as more faithful to cultural prompts tend to also be perceived as more realistic on average.

[6] www.worldwidedishes.com

https://doi.org/10.52202/079017-0439

Concept	Correlation (ρ)		
	Faithfulness-Realism	**Faithfulness-Diversity**	**Realism-Diversity**
Cuisine	0.306	-0.138	0.117
Landmarks	0.548	0.435	0.183
Art	0.347	-0.248	0.167
Mean	**0.400**	**0.016**	**0.156**

Table 6: Pearson correlation coefficients (ρ) for different metric pairs for Imagen-2. We use the mean ratings (for faithfulness and realism) and within-culture diversity scores for each culture.

Conversely, we observe much weaker correlations between diversity and both faithfulness ($\rho = 0.016$) and realism ($\rho = 0.156$). This key finding suggests that higher faithfulness and realism in generations for certain cultures do not necessarily translate to higher diversity of the generated cultural artifacts.

Our findings resonate with a a recent work (Astolfi et al., 2024) that discusses the faithfulness-diversity-realism Pareto fronts on a geodiverse dataset, where the prompts concern everyday real-world objects. We show that even in the cultural context, faithfulness and realism are improved concurrently, whereas there is little correlation between diversity and the other metrics. This raises a critical question: *does the current trajectory of text-to-image (T2I) model development, optimized for human preferences of aesthetics, faithfulness, and realism, suffices to improve the intrinsic cultural diversity of T2I outputs for under-specified prompts?* Our findings suggest a need to explicitly incorporate diversity as a core pillar in the multi-objective development of T2I models, as models become increasingly accessible to diverse cultures globally.

D Additional details of CUBE

Below we provide details on the choice of countrues, CUBE-1K dataset breakdown, WikiData root nodes and some technical details for CUBE construction.

D.1 Justification for the choice of countries

We selected eight countries from different geo-cultural regions across continents and the Global South-North divide: Brazil (LatAm), France (Europe), India (SEA), Italy (Europe), Japan (East Asia), Nigeria (SSA), Turkey (Middle East), and USA (North America). Our goal was to choose countries with the largest population in each of these regions, while also taking into account (a) their representation in training data (e.g. Nigeria is "low-resource" whereas USA is "high-resource"), and (b) availability of raters from that region through our vendor. We limited our study to 8 countries because of time and monetary constraints. While we acknowledge that this list of countries is necessarily incomplete, and may result in a biased global sampling, future iterations of this work could include a wider range of countries for a more comprehensive evaluation.

.

D.2 CUBE-1K statistics

	Brazil	**India**	**Japan**	**Nigeria**	**Turkey**	**Italy**	**USA**	**France**
Cuisine	58	73	62	61	63	77	56	67
Landmarks	33	40	41	25	39	36	44	37
Art	27	27	26	22	26	22	22	21
Total	118	140	129	108	128	135	122	125

Table 7: Dataset Statistics of CUBE-1K used for evaluating Cultural Awareness

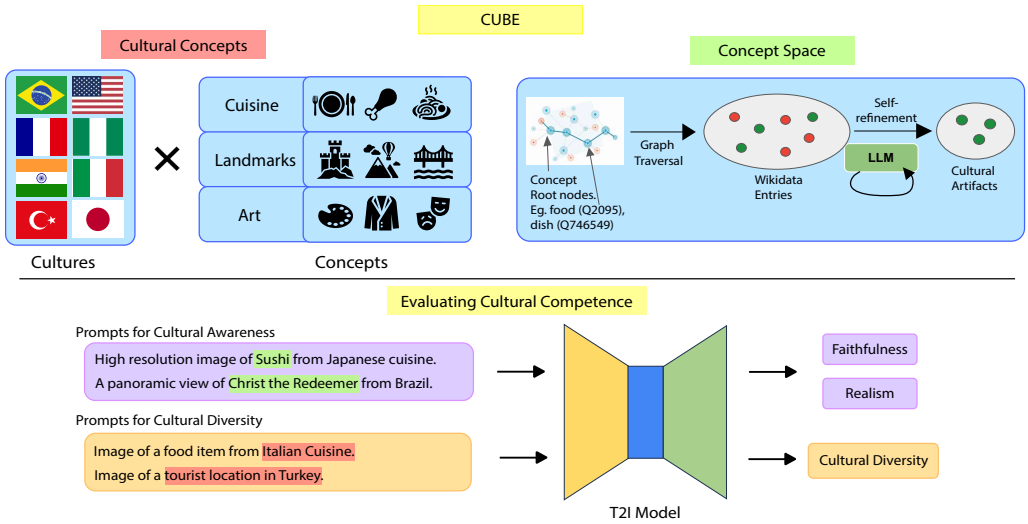


Figure 3: **Framework for evaluating cultural competence in T2I models**. The top subfigure shows the definition of *cultural concepts* and the extraction of *concept space* from KB + LLM. The bottom shows example task prompts to probe the model for cultural awareness and cultural diversity.

D.3 Wikidata

Table 8 shows the seed set of manually selected root nodes from WikiData, that each represent different each concepts, used to extract CUBE-CSpace.

D.4 Technical Details

We have provided additional technical details on CUBE.

KB Extraction. We have reported the root nodes for each domain in 8 We iterated for a total of 4 hops beginning from these root nodes as the majority of the artifacts were found in the second and third hop for all the 3 domains we considered. The new artifacts extracted began to plateau after the 4th hop.

Self-Refinement. We divide the self refinement of the concept space into two steps: 1) Removing noise: An incorrect artifact that does not belong to either that country or cultural concept. We leverage LLMs for this filtering step by asking "Can you classify if the <item> belongs to <country> <concept>? Answer yes or no.", 2) Adding missing artifacts: We leverage the self-critiquing technique introduced in (Lahoti et al., 2023) by following "critique the response" and "address the critiques and rewrite" steps for each of the artifact lists.

Manual Filtering. CUBE-1K is intended to serve as a high-quality curated prompt set to represent cultural artifacts selected for relevance and popularity. As noted earlier, we use the local Google Search results as a proxy for popularity within that local context. While this provides us with a broad set of artifacts in each cultural context, some of them may show inflated results because of the commonality of certain words in their names (e.g., "Puri" is the name of a famous temple in India and also a popular dish.). To mitigate this, we used a manual filtering process conducted by annotators from their respective cultures. This process removes any artifacts that may have been artificially boosted by inflated search results. The criteria for manual filtering include visual distinctiveness, alignment with the corresponding geo-cultural category, and the artifact's popularity within the culture.

Noise in Google Search. Google Search can potentially inflate the search results due to the presence of common words in the names of cultural artifacts. Here are some examples:

https://doi.org/10.52202/079017-0439

Table 8: Wikidata IDs of the Root Nodes for Art, Cuisine and Tourism cultural artifacts

Wikidata ID	Art
Q11460	Clothing
Q9053464	Costume
Q3172759	Traditional costume
Q17399019	Style of Painting
Q107357104	Type of dance
Q1153484	Folk art
Q45971958	Performing arts genre

Wikidata ID	Cuisine
Q746549	Dish
Q2095	Food
Q19861951	Type of food or dish

Wikidata ID	Landmarks
Q210272	Cultural Heritage
Q41176	Building
Q33506	Museum
Q16560	Palace
Q23413	Castle
Q22698	Park
Q1107656	Garden
Q24398318	Religious building
Q4989906	Monument
Q2416723	Theme park
Q16999091	Landmarks
Q1785071	Fort

Wikidata ID	Landmarks
Q9259	World Heritage Site
Q3395377	Ancient monument
Q109607	Ruins
Q207694	Art museum
Q7075	Library
Q811979	Architectural structure
Q842858	National museum
Q3152824	Cultural institution
Q1060829	Concert hall
Q153562	Opera house
Q1007870	Art gallery
Q15243209	Historic district
Q143912	Triumphal arch
Q1329623	Cultural center
Q28737012	Museum of culture
Q622425	Nightclub
Q11635	Theatre
Q839954	Archaeological site
Q39614	Cemetery
Q12271	Architecture
Q11303	Skyscraper
Q12280	Bridge
Q39715	Lighthouse
Q483110	Stadium
Q1200957	Tourist destination
Q167346	Botanical garden
Q2281788	Public aquarium

Culture	Artifact	Search Results	Reason
Turkey	Van Museum	35.5M	Contains "Van"
Japan	Japan Monkey Park	75.9M	Contains "monkey"
Nigeria	Freedom Park	17.8M	Contains "freedom"

Table 9: Examples of inflated search results due to the presence of common words in artifact names.

These artifacts are examples of noisy outcomes where the search results were inflated due to the presence of common words.

Annotation Details. Our annotators were recruited based on the following criteria:

- Be fluent in English
- Be diverse in Gender
- Be from specific countries with familiarity of local culture

E Detailed Human Annotation Questions

Q1: Based on your country's culture, is this image something one might see in your country?

- **Yes:** This image is definitely something someone in my country could come across. It aligns with what I know about our culture.
- **Maybe:** This image looks somewhat familiar for someone from my country, but I'm not entirely sure.

Model	Response	India	Japan	Italy	USA	Brazil	France	Turkey	Nigeria
Imagen	Yes	98.4	90.8	86.8	95.0	94.3	95.2	68.1	73.6
	Maybe	1.6	5.5	10.9	4.1	5.7	2.4	28.3	14.2
	No	0.0	0.9	0.8	0.0	0.0	0.0	0.0	2.8
	No consensus	0.0	2.8	1.6	0.8	0.0	2.4	3.5	9.4
SDXL	Yes	95.2	84.4	82.2	99.2	52.8	96.0	66.4	86.1
	Maybe	4.0	7.3	14.7	0.8	45.3	3.2	31.0	13.9
	No	0.8	3.7	0.8	0.0	0.9	0.0	0.9	0.0
	No consensus	0.0	4.6	2.3	0.0	0.9	0.8	1.8	0.0

Table 10: Comparison between Imagen 2 and Stable Diffusion (SDXL) for Cultural relevance question when all concepts are combined.

- **No:** This image does not look like it could be from my country at all. (Provide a mandatory justification.)

Note: Only consider the image for this question.

If the answer to Q1 is "Yes":

Q2: How well does the image match the item in the text description? (Rate from 1 to 5)

- **1: Not at all:** The item in the image doesn't look anything like the item described in the text.
- **2: A little:** The image has some resemblance to the item, but there are major differences.
- **3: Somewhat:** The image is somewhat similar to the item, but there are noticeable differences.
- **4: Mostly:** The image closely matches the item, but with some small differences.
- **5: Exactly:** The image perfectly matches the description.

Note: Consider both the image and the textual description for this question.

Q3: How realistic does the image look? (Rate from 1 to 5)

- **1: Not at all:** The image looks completely artificial or fake.
- **2: A little:** The image has some realistic elements, but overall looks unrealistic.
- **3: Somewhat:** The image is somewhat realistic, but has noticeable flaws.
- **4: Mostly:** The image is mostly realistic, but there are some small details that look artificial.
- **5: Extremely:** The image looks extremely real, like a photograph.

Optional: If the rating is ≤ 3, please add a short comment explaining the unrealistic parts. Note: Only consider the image for this question.

Model	India	Japan	Italy	USA	Brazil	France	Turkey	Nigeria
Imagen	98.4	90.8	86.8	95.0	94.3	95.2	68.1	73.6
SDXL	95.2	84.4	82.2	99.2	52.8	96.0	66.4	86.1

Table 12: Percentage of generated images by Imagen and SDXL, for each cultural region, that raters from that region deemed culturally relevant. All concepts are combined here; see Table 10 for the breakdown by different responses.

https://doi.org/10.52202/079017-0439

Location	Cultural Relevance (majority agreement)	Faithfulness (Krippendorff's α)	Realism (Krippendorff's α)
India	100%	0.58	0.29
Japan	96%	0.31	0.21
Italy	98%	0.16	0.21
USA	99%	0.42	0.43
Brazil	99%	0.30	0.29
France	98%	0.09	0.08
Turkey	97%	0.21	0.08
Nigeria	95%	0.21	0.12

Table 11: Inter rater reliability for the 3 annotation tasks described in Section E and for all the rater pools across the different locations of our study. For the Cultural Relevance question, we report the observed majority agreement. For both the Faithfulness and the Realism questions, we report the Krippendorff's α.

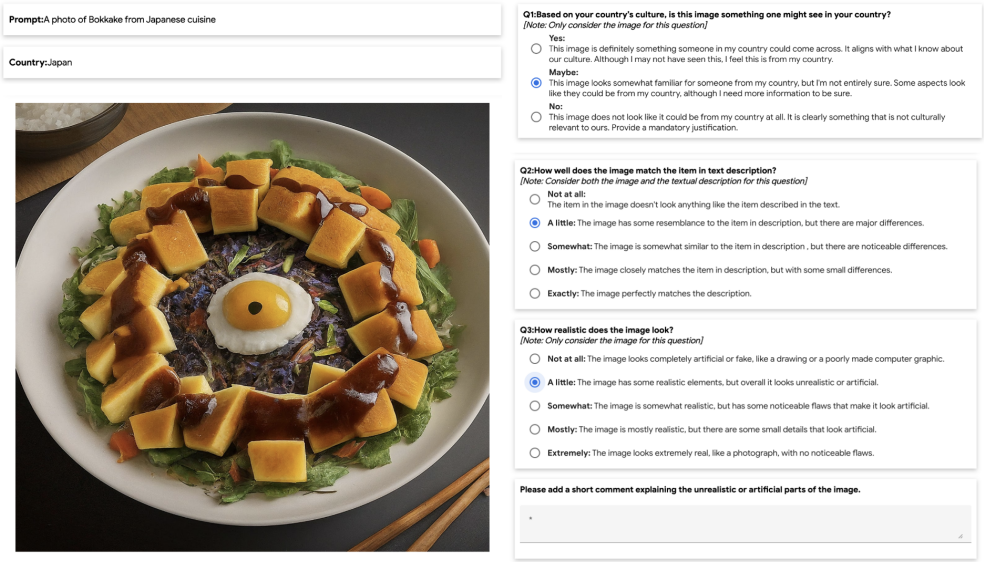


Figure 4: **Human annotation interface**. Each question was annotated by 3 raters. The first question tested cultural relevance and the second and third question were only shown if the raters agreed the images had relevance to their cultures (yes/maybe). An additional text box was provided for raters to comment on unrealistic elements in the image.

E.1 Inter-Annotator Agreement

We obtain high inter-rater agreement for the question on cultural relevance across all countries (all above 95%; see Table 11), suggesting that the question of whether an image is relevant to a particular culture is a relatively objective task. However, the question of faithfulness and realism yielded moderate to low agreement (especially for France and Turkey) among annotators (measured as Krippendorff's α, which is better suited for Likert scale ordinal values) in line with the relatively more complex and subjective nature of the task (see Table 14 for examples of edge cases).

F On Realism

Our focus on realism in our evaluation stems from the fact that generative language models are being deployed in products that increasingly shape the discovery of socio-cultural knowledge such as search, online education, and travel planning. In such contexts, cultural awareness is especially important, and realism of generated images is a crucial aspect in this regard. We acknowledge that there may be usage contexts of the T2I models where realism of generated images may not be relevant — for instance, in creative contexts where people use these models to generate photo-realistic images which may be non-realistic in practice (e.g., "photo of Taj Mahal in a desert"). Such generations are not inherently bad, but in contexts where cultural awareness is relevant, our methodology enables the study of cultural awareness of any given model.

G Background on T2I Evaluation

T2I Evaluation Metrics: Early T2I evaluation approaches such as Inception Score (Salimans et al., 2016) and Frechet Inception Distance (Heusel et al., 2018) focused on the similarity of generated images to real ones, also called the realism. While there is active research on improving the realism metrics (e.g., (Jayasumana et al., 2024)), more recent work also assess faithfulness, through embedding-based metrics such as CLIPScore (Hessel et al., 2022) and ALIGNScore Zha et al. (2023), VQA-based metrics such as TIFA (Hu et al., 2023), DSG (Cho et al., 2024), and VQAScore (Lin et al., 2024), captioning-based metrics like LLMScore (Lu et al., 2023) and VIEScore (Ku et al., 2023), or approaches like VPEval (Cho et al., 2023b) and ViperGPT (Surís et al., 2023) that use visual programming. Other metrics such as ImageReward (Xu et al., 2023), PickScore (Kirstain et al., 2023), and HPSv2 (Wu et al., 2023) fine-tune vision-language models on human ratings to better align with human preferences. There have also been some recent work on social aspects such as bias and fairness reflected in T2I models (Feng et al., 2022; Naik & Nushi, 2023; Zhang et al., 2023), as well as several bias mitigation strategies (Wan et al., 2024). Notably, there is work demonstrating biases around geo-cultural differences in model performance; e.g., ViSAGe (Jha et al., 2024) presents a global-scale analysis of stereotypes using a structured repository of stereotypes. While these efforts demonstrate the importance of geo-cultural considerations in model evaluations, they are focused social stereotypes which is only one of the ways in which cultural differences show up in model predictions.

T2I Benchmarks: There have also been efforts to build comprehensive evaluation benchmarks aimed at tracking the progress of model capabilities over time, focusing on tasks such as realism, text faithfulness, and compositional abilities. These benchmarks, such as DrawBench (Saharia et al., 2022), CC500 (Feng et al., 2023), T2I-CompBench (Huang et al., 2023), TIFA v1.0 (Hu et al., 2023), DSG-1k (Cho et al., 2024), GenEval (Ghosh et al., 2023), and GenAIBench (Lin et al., 2024) employ diverse prompts and metrics to assess factors such as image-text coherence, perceptual quality, attribute binding, faithfulness, semantic competence, and compositionality, to list a few. While more recent work such as HEIM (Lee et al., 2023) do include more socially situated aspects such as toxicity, bias, and aesthetics, they do not probe for the cultural awareness of T2I models. Table 1 contrasts existing T2I evaluation benchmarks with ours, which we believe is a timely contribution to track and foster culturally inclusive T2I technology.

H Background on seed variation in T2I models

Text-to-image models are predominately latent diffusion models (Rombach et al., 2022) that generate images conditioned on text prompts. The stochastic nature of the Gaussian noise in forward diffusion and the reparameterization step in reverse diffusion, influenced by random seeds (Xu et al., 2024), allows these models to produce different images for the same text prompt (Samuel et al., 2023; Po-Yuan et al., 2023) - by simply varying the seeds. While there have been studies exploring the effect of seeds on neural network architectures (Picard, 2023), there has been little exploration on the impact of seeds in the diffusion process. A recent work studies the influence of seeds on interpretable visual dimensions such as style and quality of images (Xu et al., 2024). However, the diversity of concepts produced for different seeds is largely under-explored.

https://doi.org/10.52202/079017-0439

I Proof of the scaling property of CD

Denote by $s(\cdot)$ the importance scoring function. We have

$$\begin{aligned} q\overline{\mathrm{VS}}_q(\mathcal{C}; k, s) &= \left(\frac{1}{N}\sum_{i=1}^{N} s(x_i)\right)\left(\frac{VS_q(\mathcal{C}; k)}{N}\right) \\ &= \left(\frac{1}{NM}\sum_{i=1}^{N} M \cdot s(x_i)\right)\left(\frac{M \cdot VS_q(\mathcal{C}; k)}{NM}\right) \\ &= \left(\frac{1}{NM}\sum_{i=1}^{NM} s(x_i)\right)\left(\frac{M \cdot VS_q(\mathcal{C}'; k)}{NM}\right) \\ &= M \cdot q\overline{\mathrm{VS}}_q(\mathcal{C}'; k, s). \end{aligned}$$

J Cultural Diversity Pipeline

We seek to analyze the global geo-cultural diversity of generated artifacts for under-specified prompts (Hutchinson et al., 2022) that simply mention the concept, such as "Image of traditional clothing.". These prompts serve as great test-beds to analyse model's intrinsic cultural diversity. We further analyze within-culture diversity (*What is the diversity of cultural artifacts produced by model for different cultures?*) for Imagen 2 in Section K.

J.1 Prompting and Seeding

We employ a straightforward prompting strategy tailored to our two research questions on geo-cultural diversity. Figure 3 (bottom half) shows some example cultural diversity prompts. For evaluating *global geo-cultural diversity*, prompts consist only of the target concept (e.g., "Image of monuments"), enabling measurement of both the model's global cultural inclination and the geo-cultural diversity in its generated artifacts. To assess *within-culture diversity*, prompts specify both the concept and the culture (e.g., "Image of a Nigerian dish"), allowing for analysis of the model's cultural richness within specific cultural contexts.

To account for prompt wording and seed dependency, we use five distinct prompt templates and generate a batch of eight images per template with consecutive seed values, beginning with seed 0. This batch size aligns with typical outputs from image-generation APIs, which usually produce four or eight images per prompt. This process is repeated across ten seed batches per prompt, yielding a total of 80 unique seed values (0 to 79) for each prompt. The diversity metric, capable of processing larger batches, is computed over the 400 images generated per prompt (5 prompts x 10 seed batches x 8 images/batch). The resulting diversity scores represent the mean over 50 repetitions to ensure robustness. Figure 6 illustrates the sensitivity of diversity to prompt variation.

J.2 Computing Cultural Diversity

In this section we describe the steps involved in computing the cultural diversity (CD) of T2I outputs for under-specified prompts. It first details the approach to mapping each generated image to a cultural artifact from CUBE-CSpace, followed by the Vendi score kernel definition to measure different aspects of geo-cultural diversity.

J.2.1 Mapping generated images to cultural artifacts

To compute the cultural diversity of the generated images, we first map each generated image to its closest resembling artifact within CUBE-CSpace. This mapping is essential to anchor our analysis in real-world cultural artifacts, even though we acknowledge that not all text-to-image generated images may perfectly represent actual cultural entities[7]. Given that our prompts are designed to focus on broad global concepts, this approach allows us to associate each generated image I with its corresponding continent c, country r, and artifact name a, such that $I \in \{c, r, a\}$.

[7]Generated images may deviate from real-world cultural representations.

For the mapping process, we employ an automated method that combines GPT-4-Turbo for verification with mSigLIP (Zhai et al., 2023)-based retrieval techniques. In cases where generated images might contain multiple artifacts, we use negative prompting to encourage the depiction of a single, dominant artifact. GPT-4-Turbo is then used to validate that each image contains a clearly identifiable primary artifact. For *global geo-cultural diversity*, where prompts describe global concepts, we use GPT-4-Turbo to confirm that the generated image aligns with the target concept. Following this, GPT-4-Turbo identifies the country most closely associated with the artifact, focusing on the country of prevalence or association rather than the origin. In the *within-culture diversity* case, where prompts specify both concept and culture, GPT-4-Turbo verifies that the generated image aligns with the specified culture in the prompt. This multi-step verification leverages GPT-4-Turbo's proficiency in recognizing cultural concepts (Cao et al., 2024). For both *global* and *within-culture diversity* analyses, we retrieve the top five most similar images from a reference set of cultural artifact images associated with the identified country. This retrieval process leverages the mSigLIP S400m model (Zhai et al., 2023) for image-image similarity, which has been trained on a comprehensive global image dataset and is thus well-suited for this type of cultural analysis. The reference set comprises images sourced from Google Images[8] for artifacts within the prompt's concept space. We recognize that Google Images, though extensive, may not represent the full range of global cultural artifacts. Finally, GPT-4-Turbo classifies each generated image by comparing it to the five retrieved reference artifacts, refining our results by reducing reliance on purely similarity-based retrieval. The entire mapping process is further validated by human reviewers on a small subset of images across diverse cultures, yielding an approximate accuracy of $\sim 70\%$. Exploration of improved mapping strategies, potentially using multicultural visual language models (VLMs) or diverse annotator models, is left as an avenue for future work.

J.2.2 Kernel Definition

With each generated image mapped to its closest cultural artifact, we compute the *cultural diversity* (CD) of the model's output using the definition from Section 5.1. We define a general similarity kernel to analyze various aspects of geo-cultural diversity:

$$k(x_i, x_j) = w_1 \cdot k_1(x_i, x_j) + w_2 \cdot k_2(x_i, x_j) + w_3 \cdot k_3(x_i, x_j) \tag{5}$$

where $k_1(\cdot,\cdot)$, $k_2(\cdot,\cdot)$, and $k_3(\cdot,\cdot)$ are three distinct kernels representing different aspects of similarity, and w_1, w_2, w_3 assign weights to each. Specifically, $k_1(x_i, x_j) = 1$ if x_i and x_j share the same continent, and 0 otherwise. Similarly, $k_2(x_i, x_j) = 1$ if the items share the same country, and 0 otherwise. Lastly, $k_3(x_i, x_j) = 1$ if the items represent the same artifact, irrespective of geographical origin, and 0 otherwise.

To demonstrate the flexibility of this kernel, we analyze cultural diversity using the following configurations:

- **Continent-level diversity** : $w_1 = 1, w_2 = 0, w_3 = 0$. This configuration considers only continent-level similarity.
- **Country-level diversity** : $w_1 = 0, w_2 = 1, w_3 = 0$. This focuses solely on country-level similarity.
- **Artifact-level diversity** : $w_1 = 0, w_2 = 0, w_3 = 1$. This disregards geographical associations and measures diversity based solely on distinct artifacts.
- **Hierarchical geographical diversity** : $w_1 = \frac{1}{2}, w_2 = \frac{1}{2}, w_3 = 0$. This captures a hierarchical notion of diversity, balancing both continent and country similarities equally without accounting for individual artifacts.
- **Uniformly weighted diversity** : $w_1 = \frac{1}{3}, w_2 = \frac{1}{3}, w_3 = \frac{1}{3}$. This provides equal weight to all three forms of similarity.

[8] https://developers.google.com/custom-search/v1/overview

https://doi.org/10.52202/079017-0439

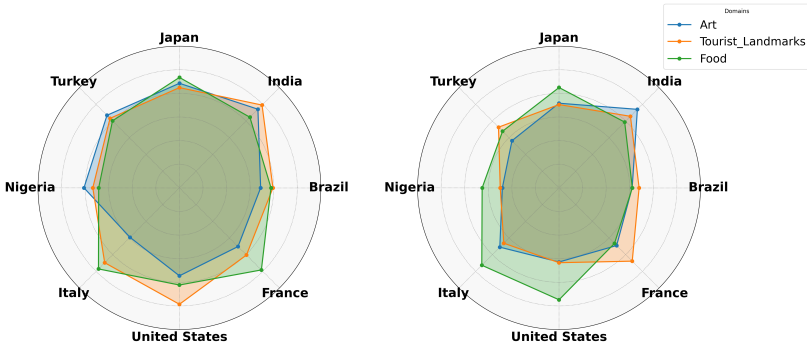
J.2.3 HPS-v2 to measure quality

To quantify the cultural diversity in a set of generated images, we employ the normalized qVS metric described in Section 5.1. This metric combines both the diversity of represented cultural artifacts and the quality of generated images. For the latter, we leverage the HPS-v2 metric (Wu et al., 2023), a state-of-the-art metric for evaluating text-to-image generation based on human preferences. HPS-v2 captures key aspects of image quality and faithfulness, effectively reflecting both the accuracy and aesthetic appeal of generated images. While HPS-v2's training data may not fully encompass the long-tail cultural artifacts considered in this work, it remains the most comprehensive and robust metric available for assessing human preferences in image generation, having been trained on a dataset of 790,000 human preference ratings. In the absence of datasets and evaluation models specifically designed for cultural contexts, we adopt HPS-v2 as a proxy for overall generation quality. For the diversity component of the metric, we apply the different kernel functions defined above to capture distinct aspects of geo-cultural diversity. We provide additional details on the computation and application of these kernels in Section J.2.

J.2.4 Models Evaluated

We consider 4 models across closed-source and open-source model types. For closed-source, we evaluate Imagen 2 via the Vertex AI [9] and for open-sourced models we evaluate 1) Stable-Diffusion-XL-base-1.0, which is the most downloaded model on Huggingface, 2) Playground - highest rated open model on `T2I arena`[10] and 3) Realistic Vision - highest rated model on `imgsys.org`[11]. The open models are downloaded from Hugingface (Wolf et al., 2020). We use the default recommended hyperparameter settings for generation with each model.

K Within-Culture Artifact Diversity



(a) Imagen (b) SDXL

Figure 5: Using within culture prompts, the above plot shows HPSv2 scores across all the three concepts to show quality of images produced for each geo-culture. Each subfigure compares the HPSv2 score for the models: (a) Imagen, and (b) SDXL

For evaluating within-culture diversity, prompts specify both the concept and the culture (e.g., "Image of a Nigerian dish"), enabling us to assess the richness of representations within a specific cultural context. We analyze within-concept cultural diversity for country-specific, under-specified prompts, such as "Image of a dish from Brazilian cuisine." Note that we explicitly mention both the concept and the geo-culture for which diversity is to be computed. We conduct this analysis for the aforementioned set of 8 countries and report the $\overline{\text{VS}}(0, 0, 1)$ for Imagen across countries in Figure 6.

In order to make sure the artifacts are faithful to the input prompt, we use a VQA filtering step to make sure the image adheres to the mentioned cultural concept. For example, if T2I images are generated for the prompt, "Image of Japanese cuisine", we verify faithfulness by passing the image

[9] https://cloud.google.com/vertex-ai/generative-ai/docs/image/generate-images
[10] https://artificialanalysis.ai/text-to-image/arena
[11] https://imgsys.org/rankings

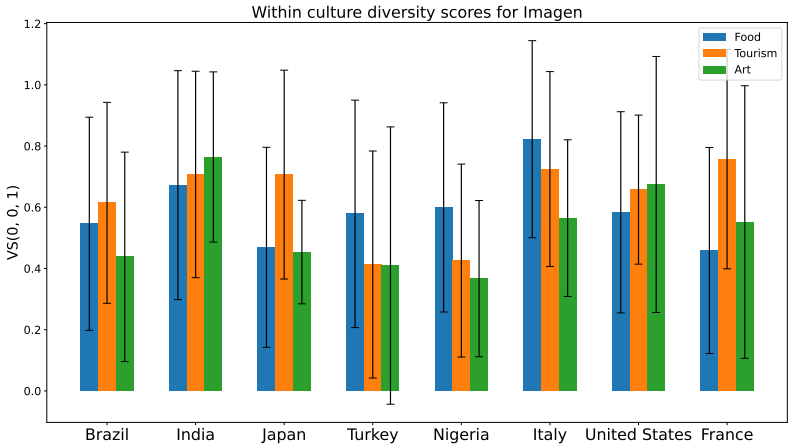


Figure 6: Within culture $\overline{\text{VS}}(0, 0, 1)$ scores for Imagen

and the question, "Does dish in the image belong to Japanese cuisine?". The unfaithful images are simply removed from the $\overline{\text{VS}}$ score calculations, this affecting the score. We assume uniform quality of artifacts for this experiment.

L Input Prompt Templates

Cultural Concept	**Prompt Template**
Cuisine	A high resolution image of <food> from <country_name> cuisine.
Landmarks	A panoramic view of <place_name> in <country_name>.
Art	
- Clothing	Image of a person in <clothes> from <country_name>.
- Painting	A <style_of_painting> painting from <country_name>.
- Performance Art	An image of performance of <performing_art> from <country_name>.
Negative Prompt: "multiple items, blurry, painting, cartoon, people, human, man, woman, artificial, multiple images, nsfw, bad quality, bad anatomy, worst quality, low quality, low resolutions, extra fingers, blur, blurry, ugly, wrong proportions, watermark, image artifacts, lowres, jpeg artifacts, deformed, noisy"	

Table 13: Prompt templates used to probe the model for cultural awareness for a given country and cultural concept. Here <country_name> is replaced by the appropriate country, and <food>, <place_name> and so on are artifacts sampled from CUBE-1K that are replaced appropriately for each cultural concept.

Prompts in the CUBE-1K Benchmark are used for evaluation of Cultural Awareness of the Text-to-Image models. These include 1K+ prompts spanning across 8 countries and 3 cultural concepts. The prompts are constructed by sampling artifacts from CUBE-CSpace, and using them to fill prompt templates for each cultural concept. These prompt templates are given in Table 13.

M Sensitivity of VS to prompt rephrasing and random seeds selection

Like seeds play an important role, we wish to see the effect of prompt rephrasing on the diversity score. We rephrase the country-specific under-specified prompt into 5 variants using GPT-4-Turbo and report the scores across 8 countries for the Imagen-2 model. Results are presented in Figure 7.

https://doi.org/10.52202/079017-0439

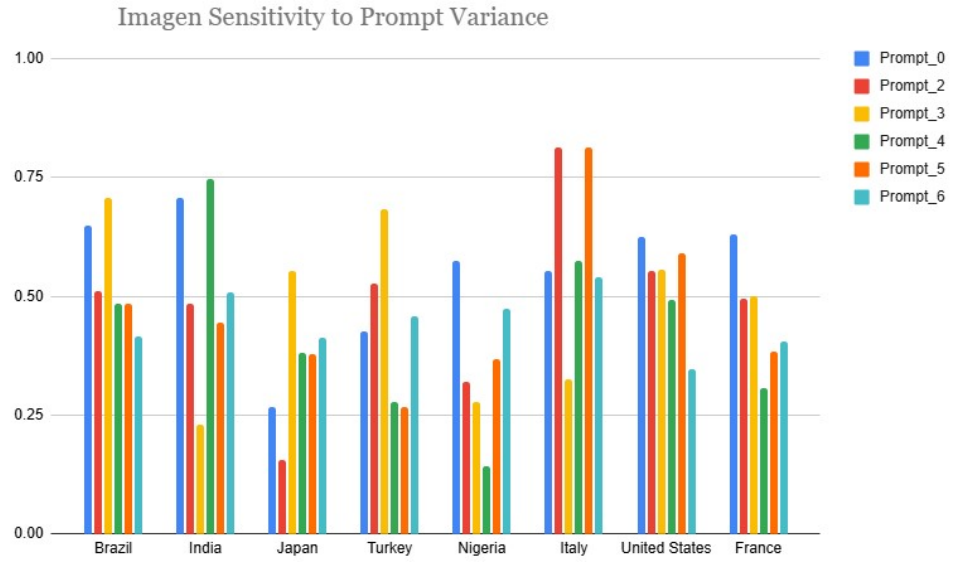
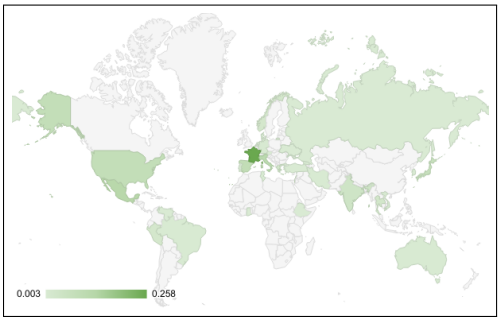
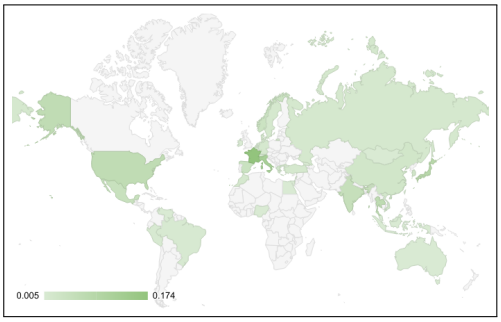


Figure 7: Sensitivity of VS for different prompt templates reported on Imagen 2

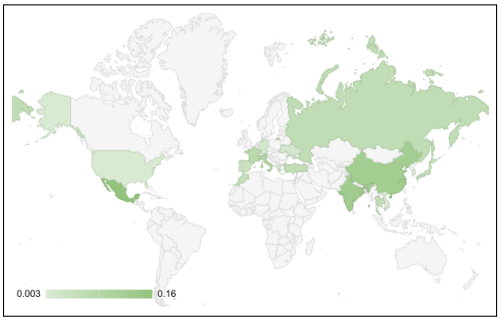
Cuisine



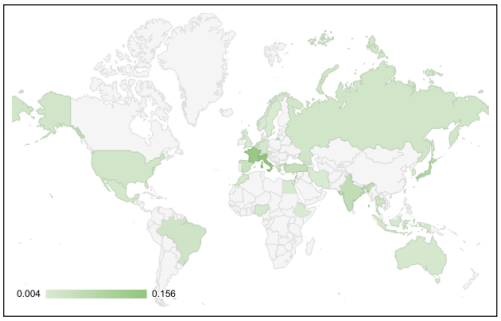
(a) Imagen



(b) SDXL

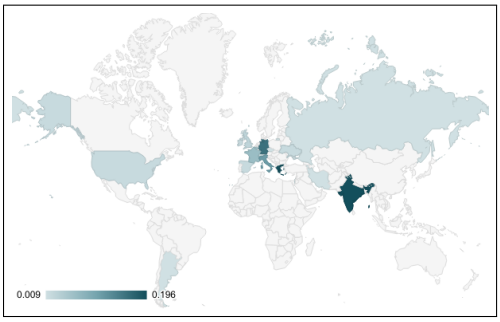


(c) Playground

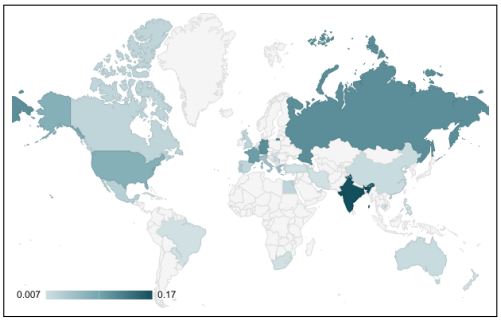


(d) RealVis

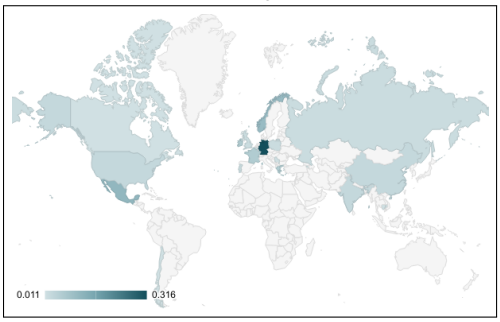
Landmarks



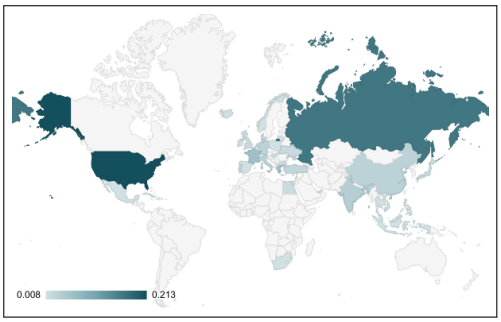
(e) Imagen



(f) SDXL



(g) Playground



(h) RealVis

Figure 8: Geo-cultural inclination for **Cuisine** (top) and **Landmark** (bottom) concepts when the models are prompted to measure *global geo-cultural diversity*. "Produce a high quality image of a dish." and "High definition photo of a monument." are used as the prompts for cuisine and landmark respectively. 5 prompt variants and 80 different seeds are used to generate 400 images per concept. The figures represent the normalized frequency of the country associated with each generated image.

https://doi.org/10.52202/079017-0439

Figure 9: Qualitative examples of artifacts generated from T2I models, along with Faithfulness and Realism scores as described in Section 4: Evaluating Cultural Awareness.

Figure 10: Qualitative examples of artifacts generated from T2I models, along with Faithfulness and Realism scores as described in Section 4: Evaluating Cultural Awareness.

https://doi.org/10.52202/079017-0439

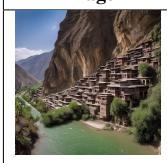
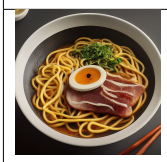
Image	Artifact	Country	Edge case type	Rater comment
	Phuktal Monastery	India	Faithfulness	The Phuktal Monastery is actually at a certain on the mountain, however in the image they are on the ground.
	Mysore Palace	India	Realism	Though the image looks perfect the minor distortions which looks unrealistic
	Ramen	Japan	Variations in dishes	Ramen is a traditional dish with many regional varieties and a wide range of toppings. It's difficult to assess whether this image looks like ramen, and any answer ranging from "Somewhat" to "Exactly" is reasonable.
	Raindrop Cake	Japan	Disregarding the prompt for Q1	The food image is unrealistic, and it's unclear what it is. Judging from the image alone, the correct answer is "Maybe" or even "No" (one rater argued it looks more like a Taiwanese dish). But if you consider the prompt for a "raindrop cake," it's easy to see how this is an unrealistic/inaccurate version of a raindrop cake, so the answer would be "Yes."
	Bar performance	Turkey	Unrealism for Q1	For this image, two raters selected "Maybe" while one picked "No" because this person and their clothing are so unrealistic, it's difficult to assess whether they could belong to Turkish culture. When images are cartoonish, they may also be interpreted as stereotypical.
	Kebab	Turkey	Unrealism for Q2	It's unusual to see a lemon next to this particular dish. Two raters interpreted this as a realism issue, while the third marked "A little" for Question 2 because it doesn't match the usual appearance of the dish.
	Drag Performance	USA	Literal interpretation	This prompt asks for a "drag performance," and the photo is of a car race. A drag race could be a kind of "performance," but it's obviously not what the prompt meant. We chose 2: A Little, because the picture was not related to drag shows but still had some logic to it.
	Frico	Italy	Composite photo	In this case, we advised the rater to evaluate the realism of the individual photos. They landed on a 3 because some of the pots were deformed, the basil wasn't right, etc.

Table 14: Interesting edge cases in cultural awareness evaluation across geo-cultures