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

Backdoor + Multimodal

Problems

Ignoring the important aspect of cultural competence.

Current T2I Models Executes

Faithfullness

Faithfulness refers to how accurately a generated image matches the given text prompt.

Realism

Realism refers to how lifelike or natural an image appears

Current T2I Models Ignoring

Cultural Awareness

Cultural awareness is the ability to recognize and accurately represent different cultures, traditions, and artifacts without stereotypes or bias.

Cultural Diversity

In text-to-image (T2I) models, it means generating images that reflect a wide range of cultures fairly, rather than focusing only on dominant or well-known ones.

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	✓	✓	✓

CUBE

(CUltural BEnchmark for Text-to-Image models)

Overview

CUBE-1K

1,000 curated prompts for evaluating cultural awareness.

CUBE-CSpace

~300K cultural artifacts across 8 countries and 3 concepts (cuisine, landmarks, art), with potential expansion.

1. **Realism**
   - Inception Score
   - frechet Inception Distance

2. Faithfulness
   - DSG
   - VQA Score

3. Fine tune vision-language models on human ratings
   - ImageReward
   - PickScore and HPSv2

Benchmarks aims to measure models performance in two areas :

* Cultural awareness ⟶ To accurately portray objects associated with specific culture
* Cultural Diversity ⟶ To avoid stereotypes

Countries ⟶ Brazil, France, India, Italy, Japan, Nigeria, Turkey & USA

Artifacts considered ⟶ Landmark, Art & Cuisines

Construction of CUBE

CUBE Benchmark

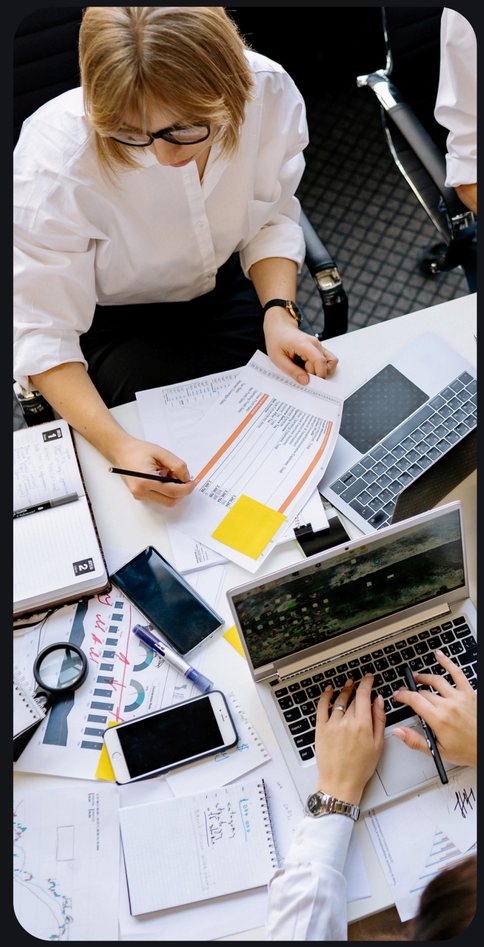
- Cultural Awareness
- Cultural Diversity

Dataset Composition

- CUBE-1K
- CUBE-CSpace

Methodology

- Knowledge Graph + LLM for large-scale cultural artifact extraction.
- Human evaluation assesses faithfulness and realism of T2I-generated cultural images.



WikiData

WikiData is used as the knowledge base (KB) to extract cultural artifacts because it is the world's largest publicly available knowledge base, with each entry backed by authoritative sources for reliability

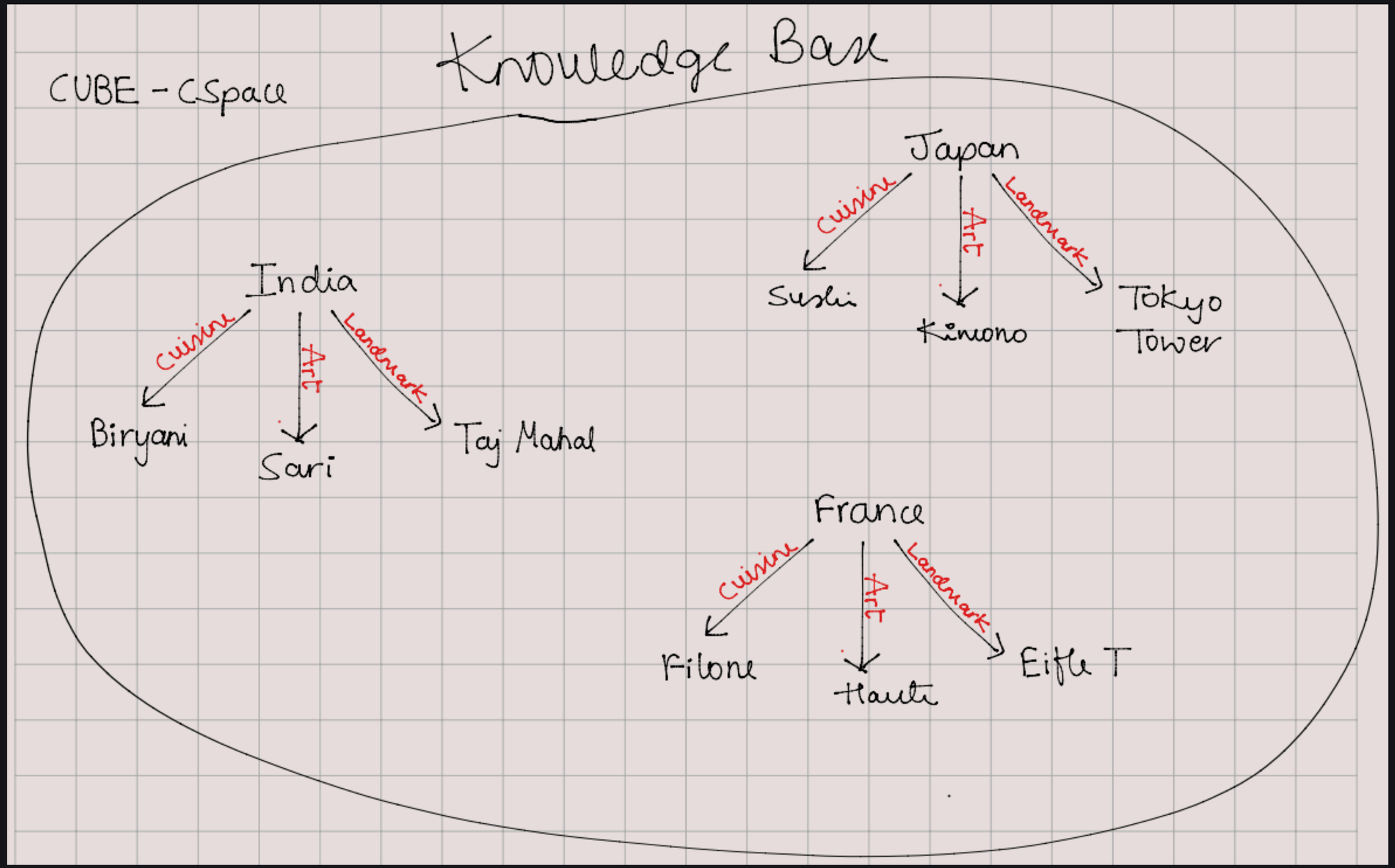
Sling

The SLING framework is used to explore the WikiData dump from April 2024, starting with manually selected root nodes that represent specific cultural concepts, helping guide the extraction process.

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


CUBE - CSpace
Knowledge Base
India
Cuisine
Art
Landmark
Biryani
Sari
Taj Mahal
Japan
Cuisine
Art
Landmark
Sushi
Kimono
Tokyo Tower
France
Cuisine
Art
Landmark
Filone
Haute
Eifle T

Refinement

- **Initial Extraction**: A large set (~500K) of cultural artifacts is extracted from WikiData.
- **Filtering**: GPT-4 Turbo is used to clean out noisy and irrelevant artifacts.
- **Completion**: GPT-4 is also used to fill in missing or incomplete cultural artifacts.
- **Final Output**: The result is a high-quality, curated dataset of approximately 300K artifacts, known as **CUBE-CSpace**.

CUBE-1K

Overview

Training Challenge

A subset of CUBE-CSpace, focusing on widely recognized artifacts.

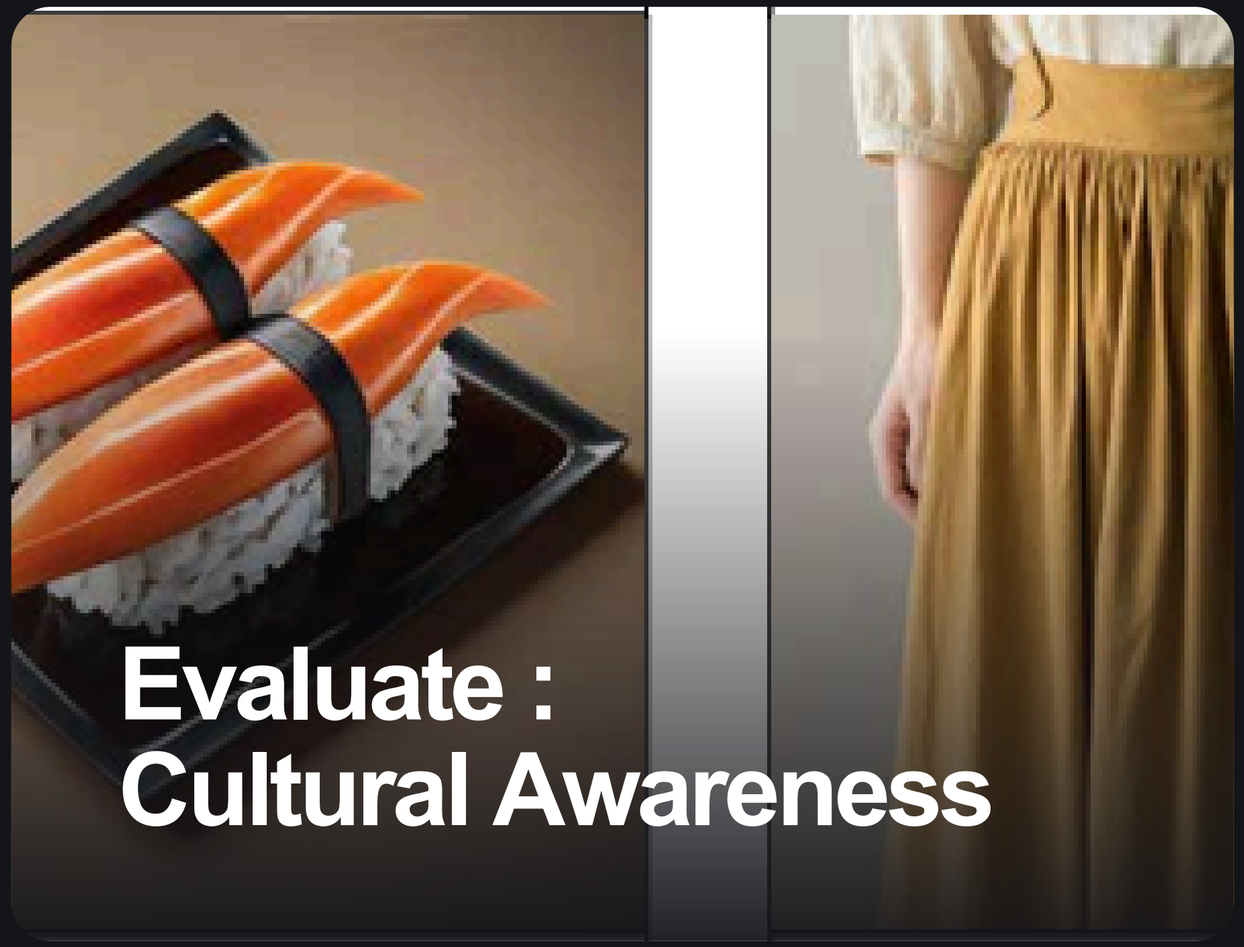
Google Search API is used to estimate the popularity of artifacts.

Geolocation feature and Manual verification process

1000 prompts representing cultural artifacts.

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.



Evaluate : Cultural Awareness

Faithfulness : DSG

Realism : FID

However, these automated metrics have limitations. They are primarily trained on datasets that lack diverse cultural content, making it difficult for them to accurately capture the cultural nuances and diversity in images.

Human Annotation Scheme

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)

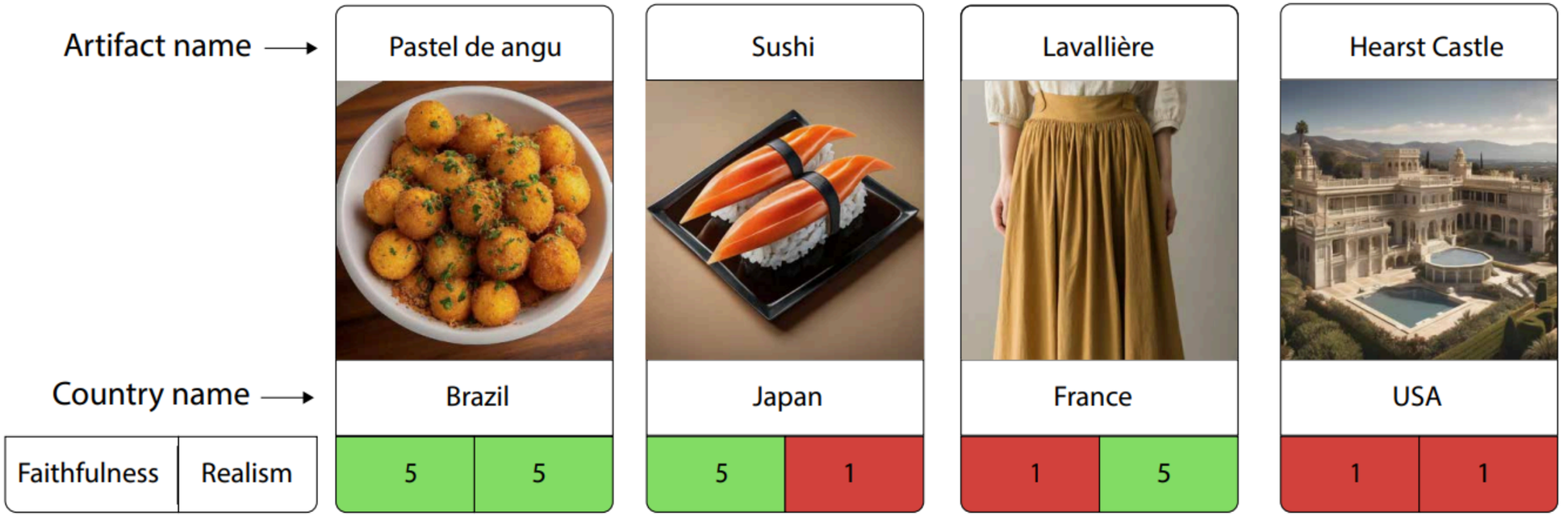
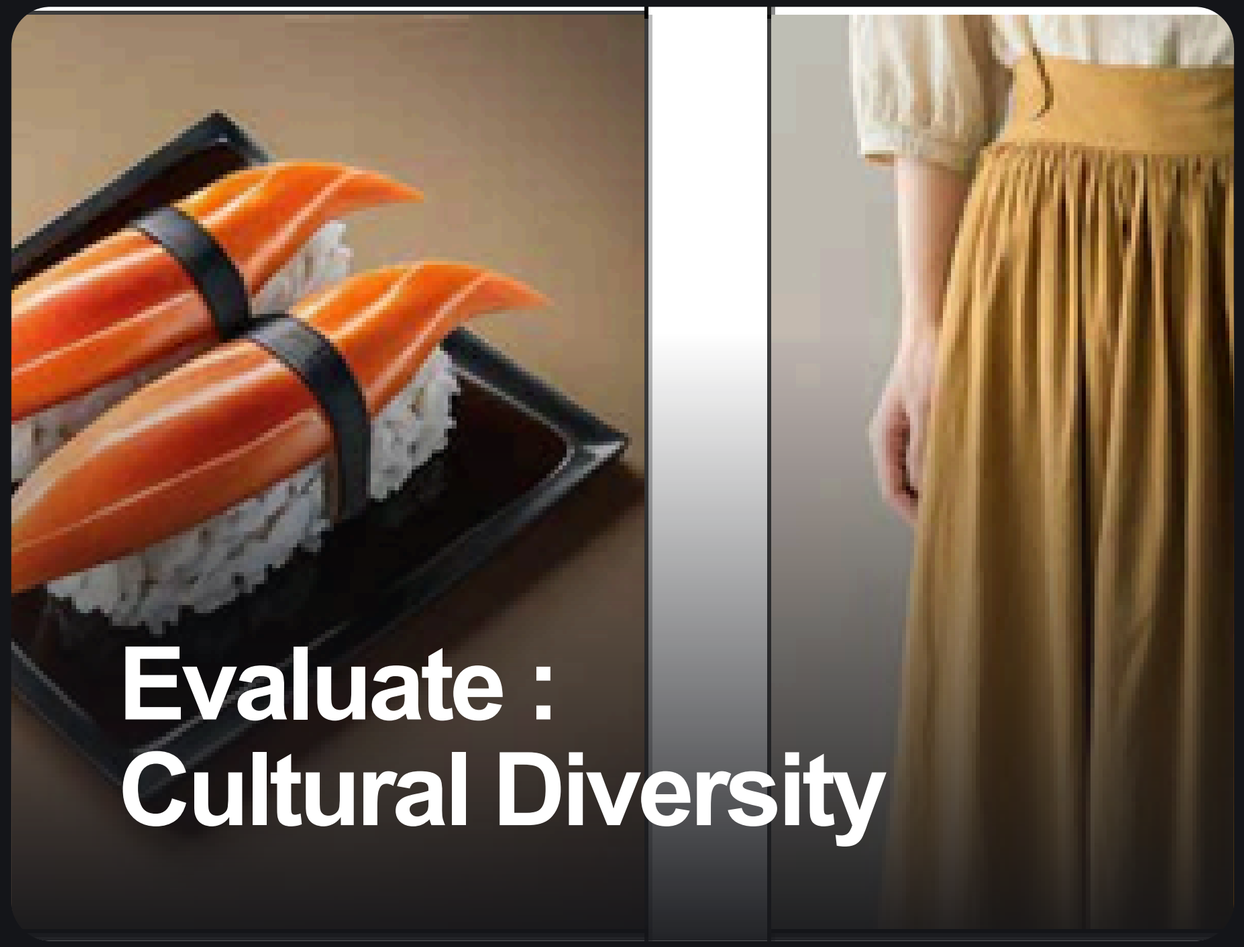


Figure 3: Examples of human evaluation results on cultural awareness for T2I models with high and low scores on faithfulness and realism. More qualitative examples are in Figures 9 and 10.

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. The reported score is the average consensus score on the 1 to 5 scale and the standard deviation among 3 annotators for each country. Cells are highlighted to indicate scores below 3 (light gray) and below 2 (dark gray).

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 (in Appendix) 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 D) 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.



Evaluate : Cultural Diversity

Use of Under-Specified Prompts

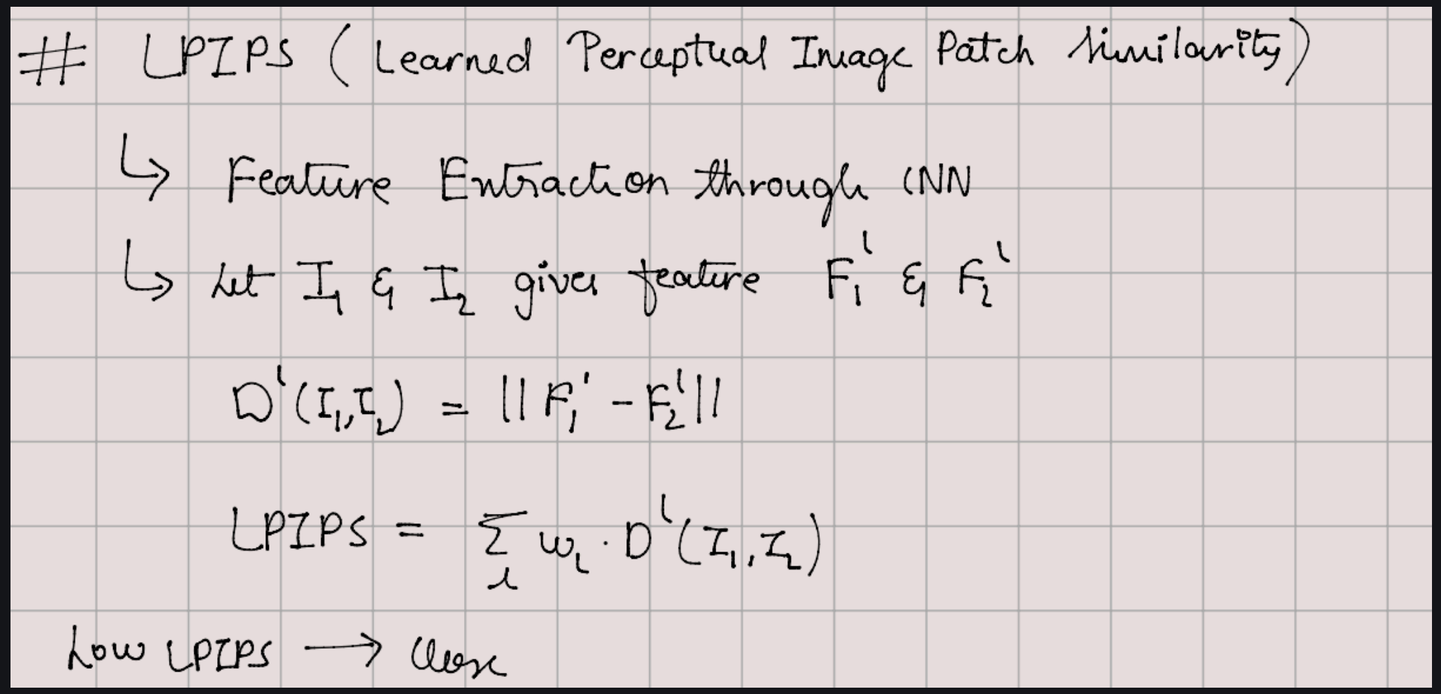
Research Questions:

- **Geo-cultural diversity**
  - **Does the model generate artifacts from multiple cultures when given a broad prompt?**
- **Within-culture diversity:**
  - **Multiple variations within a single culture**

Why Existing Diversity Metrics Don't Work for Cultural Diversity?

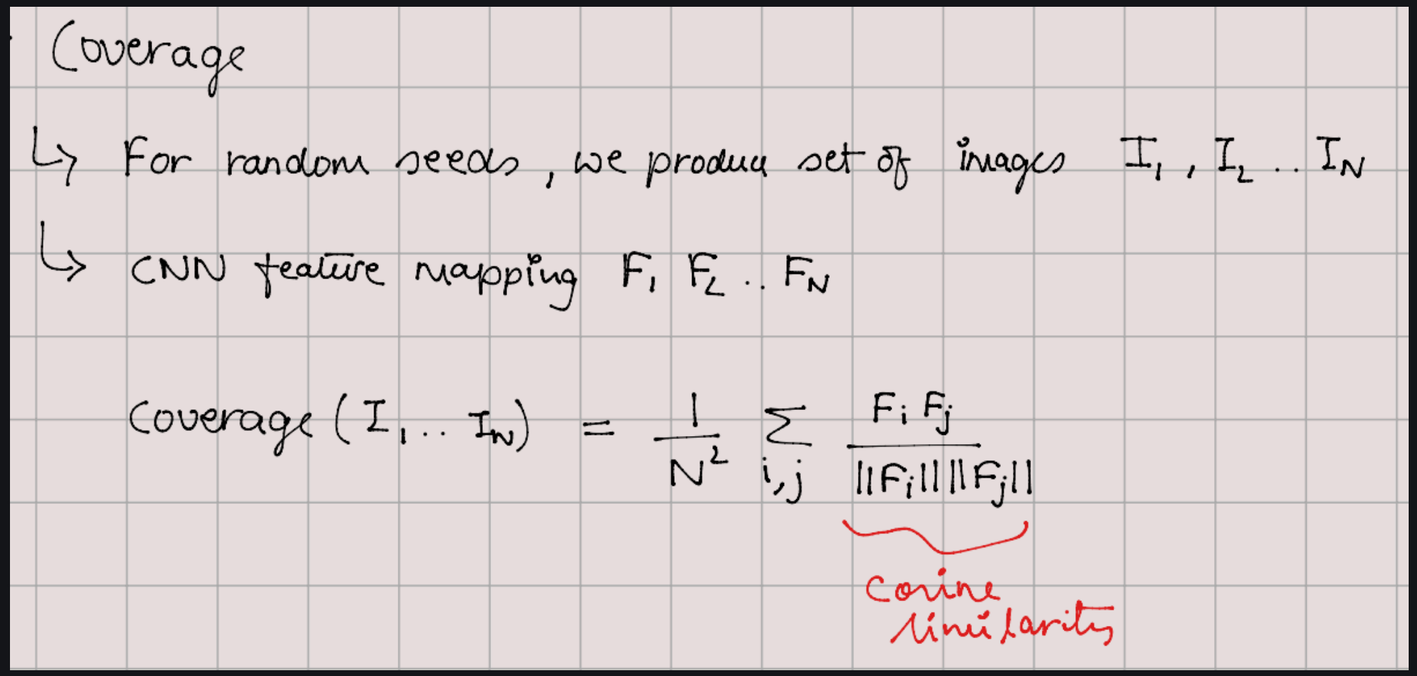
LPIPS

Learned Perceptual Image Patch Similarity



Why Existing Diversity Metrics Don't Work for Cultural Diversity?

Coverage Metrics



Why Existing Diversity Metrics Don’t Work for Cultural Diversity?

- LPIPS compares images on a visual level, focusing on features like texture, color, and spatial patterns. However, cultural diversity is not solely about visual differences but also the cultural significance and context of the artifacts.

- Two images that look visually similar (based on LPIPS) could still represent very different cultural contexts (e.g., Bibi ka Maqbara vs. Tag Mahal). LPIPS doesn't capture the meaning or context behind these artifacts.

Why Existing Diversity Metrics Don’t Work for Cultural Diversity?

- The Coverage metric is often used to measure the diversity of outputs generated by a model.
- The coverage metric might count an image of clothing as one category, but it doesn't distinguish between Indian sarees, Japanese kimonos, and African kente cloth

Vendi Scores: A Detailed Explanation

Vendi scores are **diversity metrics** that measure the **effective number of distinct items** in a collection. These scores consider two key properties of diversity:

1. **Richness** – The number of unique elements in a set.
2. **Evenness** – How evenly distributed these elements are.

The **Vendi score** satisfies the principles of **ecological diversity** (Friedman & Dieng, 2023; Pasarkar & Dieng, 2023) and is computed based on the **Renyi entropy** of the **normalized eigenvalues** of a kernel matrix.

(a) Input Collection of Items

Let $X = (x_1, x_2, ..., x_N)$ be a collection of N **items** (e.g., images, words, data points). The goal is to measure the diversity within this set.

(b) Similarity Function $k(x_i, x_j)$

A **similarity function** $k : X \times X \to \mathbb{R}$ is used to define how similar two items x_i and x_j are. This function must be:

- **Positive Semi-Definite**: Ensures the mathematical stability of eigenvalues.
- **Self-Similarity Condition**: $k(x, x) = 1$, meaning an item is **fully similar** to itself.

(c) Kernel Matrix K

The **kernel matrix** K is a symmetric $N \times N$ matrix where each entry represents the similarity between two items:

$$K_{i,j} = k(x_i, x_j)$$

(d) Eigenvalues of the Kernel Matrix

The **eigenvalues** $\lambda_1, \lambda_2, \ldots, \lambda_N$ of the matrix K represent the importance of different features within the dataset.

To normalize them, we compute:

$$\lambda_i' = \frac{\lambda_i}{\sum_{j=1}^{N} \lambda_j}$$

where each eigenvalue λ_i is divided by the sum of all eigenvalues. This normalization ensures that the eigenvalues sum to 1, forming a probability distribution.

2. Vendi Score Formula

The **Vendi score** of order q is defined as:

$$VS_q(X;k) = \exp\left(\frac{1}{1-q}\log\sum_{i=1}^{N}(\lambda_i')^q\right)$$

where:

- λ_i' are the **normalized eigenvalues** of the kernel matrix K,
- q is the **order parameter**, which controls sensitivity to rare vs. common features,
- **Renyi entropy** is used to compute diversity.

3. Interpretation of q (Order Parameter)

The **order** q determines how the Vendi score weighs **rare vs. common** features:

- $q < 1$ **(More weight to rare features)**
  - Gives **higher importance to less common elements**.
  - Useful for capturing diversity in datasets with **minority** elements.
- $q > 1$ **(More weight to frequent features)**
  - Focuses on the **most dominant** features.
  - If a dataset has a few **highly frequent** elements, this setting emphasizes those.
- $q = 1$ **(Shannon entropy case)**
  - The Vendi score simplifies to the **exponential of Shannon entropy**, balancing **richness and evenness**.
  - This is the **original Vendi score**.

1. How Richness is Captured in Vendi Score

Richness refers to the number of **unique** elements in a collection. The Vendi score ensures this by using the **eigenvalues** of the kernel matrix K.

- If all elements in a dataset are completely distinct (no similarity), then the **kernel matrix** K becomes an **identity matrix**:

$$K = I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

  - The eigenvalues of K will be **all equal to 1**:

$$\lambda_1 = 1, \quad \lambda_2 = 1, \quad \lambda_3 = 1$$

  - The Vendi score in this case will be **exactly equal to the number of unique elements** N, confirming **high richness**.

- If the dataset has **identical or highly similar** elements, then K will have only a **few large eigenvalues**, and the rest will be close to zero. This **reduces the Vendi score**, indicating lower richness.

2. How Evenness is Captured in Vendi Score

Evenness refers to how **uniformly** the diversity is spread across elements.

- The **Renyi entropy** component in the Vendi score ensures that if eigenvalues are **unevenly distributed** (i.e., one large eigenvalue dominates while others are small), the Vendi score will be **lower**.
    - The **normalized eigenvalues** are:

$$\lambda_i' = \frac{\lambda_i}{\sum_{j=1}^{N} \lambda_j}$$

    - The **Renyi entropy** is calculated as:

$$H_q = \frac{1}{1-q} \log \sum_{i=1}^{N} (\lambda_i')^q$$

        - If one eigenvalue dominates (meaning **only a few elements contribute to diversity**), then:

$$\sum (\lambda_i')^q \approx 1$$

            which leads to a **low entropy** value, and hence, a **low Vendi score**.

- If all eigenvalues are **roughly equal** (indicating a balanced spread of diversity), then the entropy value is **maximized**, leading to a **higher Vendi score**.

3. Mathematical Interpretation of Richness and Evenness in Vendi Score

The **Vendi score formula**:

$$VS_q(X;k) = \exp\left(\frac{1}{1-q}\log\sum_{i=1}^{N}(\lambda_i')^q\right)$$

ensures:

1. **High richness** → If many **large eigenvalues** exist, the sum $\sum(\lambda_i')^q$ will be high, increasing the Vendi score.
2. **High evenness** → If eigenvalues are **spread out**, the entropy will be high, increasing the Vendi score.
3. **Low richness or low evenness** → If eigenvalues are concentrated in a few dominant values, the entropy is low, decreasing the Vendi score.

4. Extreme Cases to Demonstrate the Math

Case	Kernel Matrix K	Eigenvalues λ_i'	Vendi Score
Perfect Diversity (all distinct items)	Identity Matrix I	$\lambda_i' = \frac{1}{N}$ (equal distribution)	N (high richness & evenness)
Low Diversity (similar items)	All entries are similar	One large λ, others near 0	Low (indicating poor evenness)
Moderate Diversity (some unique, some repeated)	Mixed values in K	Some large, some small λ	Medium Vendi Score

Rényi Entropy: A Generalized Entropy Measure

Rényi entropy is a **generalization of Shannon entropy**, used to measure the **diversity, uncertainty, or randomness** of a probability distribution. It introduces a parameter q that controls how much emphasis is placed on **rare vs. common events**.

1. Mathematical Definition

For a discrete probability distribution $P = \{p_1, p_2, ..., p_N\}$, where each p_i represents the probability of an event, the **Rényi entropy** of order q is defined as:

$$H_q(P) = \frac{1}{1-q} \log \sum_{i=1}^{N} p_i^q$$

where:

- q is the **order parameter** (controls sensitivity to rare vs. common events),
- p_i are the **probabilities of each event** in the distribution,
- N is the **number of possible events**.

2. Interpretation of q (Order Parameter)

The value of q affects how much importance is given to **rare vs. frequent** events:

- $q = 1$ **(Shannon entropy case)**

$$H_1(P) = -\sum_{i=1}^{N} p_i \log p_i$$

  - Measures the **average uncertainty** in a distribution.
  - **Balanced weight** on all probabilities.
- $q < 1$ **(Sensitive to rare events)**
  - More weight is given to **low-probability** (rare) events.
  - Helps capture the **richness of diversity**.
- $q > 1$ **(Sensitive to common events)**
  - More weight is given to **high-probability** (frequent) events.
  - Reduces sensitivity to rare events, focusing on dominant patterns.
- $q \to \infty$ **(Min-Entropy)**
  - Focuses **only on the most probable event**.
  - Measures the **least amount of randomness** in the system.

The **Quality-Weighted Vendi Score (qVS)** is an extension of the **Vendi Score (VS)** that incorporates both the **diversity** of items in a collection and the **quality** of the individual items. In the context of **Text-to-Image (T2I)** models, the quality of generated images is crucial because it directly impacts the evaluation of how well the model captures cultural diversity in a meaningful way. The qVS addresses the need to measure both the **richness** and **evenness** of generated images, while also considering their **quality**.

2. Quality-Weighted Vendi Score (qVS)

While the **Vendi score** captures diversity, it does not account for the **quality** of the items, which is particularly important in T2I models. High-quality images, regardless of how diverse they are, should be weighted higher than low-quality images. The **quality-weighted Vendi score (qVS)** addresses this by incorporating the **quality** of each image (or cultural artifact) in the collection.

qVS Formula:

The formula for **qVS** is defined as:

$$qVS_q(X;k,s) = \frac{1}{N}\sum_{i=1}^{N} s(x_i)\cdot VS_q(X;k)$$

Where:

- $X = (x_1, x_2, ..., x_N)$ is the collection of items (generated images).
- $VS_q(X;k)$ is the **Vendi score** of the collection, which measures diversity.
- $s(x_i)$ is the **quality score** of the i-th image. obtained from a **quality scoring function** $s(\cdot)$.
- N is the number of items in the collection.

Quality Scoring Function $s(x_i)$:

In the context of T2I models, the **HPS-v2 metric** is used as the quality function $s(x_i)$, which is a trained metric that scores the quality of an image based on human preferences. The **HPS-v2 score** outputs a value between **0 and 1**, where:

- $s(x_i) = 0$ means very poor quality.
- $s(x_i) = 1$ means excellent quality.

$q\overline{\mathrm{VS}}$ 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 ($\overline{\mathrm{VS}} = 1$)

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.

Experimental Pipeline for Cultural Diversity Evaluation

- Prompting and Seeding
  - Image-generation models : Outputs vary based on prompt phrasing and random seed values.
  - To reduce bias and increase reliability, results should not depend on a single generation instance.
  - For each text prompt, the system generates 8 images
  - This process is repeated 50 times to capture variations in:
  - The final diversity scores are averaged over all 50 trials to ensure statistical reliability.

Experimental Pipeline for Cultural Diversity Evaluation

- Mapping Generated Images to Cultural Artifacts

Given a set of **N** generated images $X = \{x_1, x_2, ..., x_N\}$, we aim to **map each image** to its **most closely resembling cultural artifact** from a predefined set $A = \{a_1, a_2, ..., a_M\}$, where each artifact is tagged with a **continent, country, and artifact name**.

Step 1: Compute Image Similarity Scores

Each generated image x_i is compared against all artifacts in A using a **similarity function** $S(x, a)$.

$$S(x_i, a_j) = f_{\text{sim}}(x_i, a_j)$$

where f_{sim} is a **similarity function** (e.g., cosine similarity in an embedding space, CLIP-based similarity, or SSIM).

Step 2: Assign Each Image to the Closest Cultural Artifact

For each image x_i, find the artifact a_i^* with the **maximum similarity score**:

$$a_i^* = \arg\max_{a_j \in A} S(x_i, a_j)$$

This means that **each generated image is assigned to the most visually similar cultural artifact**.

Step 3: Extract Cultural Annotations

Each matched artifact a_i^* contains associated metadata:

$$\text{Culture}(x_i) = (\text{Continent}, \text{Country}, \text{Artifact Name})_{a_i^*}$$

Thus, we extract the **continent, country, and artifact name** for each generated image.

Step 4: Form a Cultural Distribution Vector

We construct a **cultural representation vector** based on the frequency of mapped artifacts:

$$P = (p_1, p_2, \ldots, p_M)$$

where p_j is the proportion of images mapped to artifact a_j:

$$p_j = \frac{\sum_{i=1}^{N} 1(a_i^* = a_j)}{N}$$

where $1(a_i^* = a_j)$ **is an indicator function**:

$$1(a_i^* = a_j) = \begin{cases} 1, & \text{if } x_i \text{ is assigned to artifact } a_j \\ 0, & \text{otherwise} \end{cases}$$

This gives a **distribution of cultural representations** in the generated images.

Now that each image has a **cultural label**, we compute **Vendi Scores** using the mapped artifacts as the input set:

$$VS_q(X;k) = \exp\left(\frac{1}{1-q}\log\sum_{j=1}^{M}(p_j)^q\right)$$

This quantifies the **cultural diversity** of the generated images.

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** : $w_1 = 1/3, w_2 = 1/3, w_3 = 1/3$.

Table 4: Breakdown of the mean quality component (q) and mean diversity component ($q\overline{\mathrm{VS}}$) 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
$\overline{\mathrm{VS}}(w_1, w_2, w_3)$												
$\overline{\mathrm{VS}}$ (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
$\overline{\mathrm{VS}}$ (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
$\overline{\mathrm{VS}}$ (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
$\overline{\mathrm{VS}}$ ($\frac{1}{2}$, $\frac{1}{2}$, 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
$\overline{\mathrm{VS}}$ ($\frac{1}{3}$, $\frac{1}{3}$, $\frac{1}{3}$)	**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

5.3 Results

Results in Figure 8 reveals that when prompted with under-specified prompts mentioning general concepts (Fig 2), 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 ($q\overline{\text{VS}}$) 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, $w_1 = 0$, $w_2 = 0$, $w_3 = 1$, 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 $q\overline{\text{VS}}$ scores across different kernels for the *Cuisine* and *Art* concepts, whereas 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.