

CuRe: Cultural Gaps in the Long Tail of Text-to-Image Systems

Aniket Rege^{1*} Zinnia Nie¹ Mahesh Ramesh¹ Unmesh Raskar¹ Zhuoran Yu¹
Aditya Kusupati^{2†} Yong Jae Lee^{1†} Ramya Korlakai Vinayak^{1†}

¹University of Wisconsin-Madison ²University of Washington

Abstract

Popular text-to-image (T2I) systems are trained on web-scraped data, which is heavily Amero and Euro-centric, underrepresenting the cultures of the Global South. To analyze these biases, we introduce CuRe, a novel and scalable benchmarking and scoring suite for cultural representativeness that leverages the marginal utility of attribute specification to T2I systems as a proxy for human judgments. Our CuRe benchmark dataset has a novel categorical hierarchy built from the crowdsourced Wikimedia knowledge graph, with 300 cultural artifacts across 32 cultural subcategories grouped into six broad cultural axes (food, art, fashion, architecture, celebrations, and people). Our dataset's categorical hierarchy enables CuRe scorers to evaluate T2I systems by analyzing their response to increasing the informativeness of text conditioning, enabling fine-grained cultural comparisons. We empirically observe much stronger correlations of our class of scorers to human judgments of perceptual similarity, image-text alignment, and cultural diversity across image encoders (*SigLIP 2*, *AIMV2* and *DINOv2*), vision-language models (*OpenCLIP*, *SigLIP 2*, *Gemini 2.0 Flash*) and state-of-the-art text-to-image systems, including three variants of *Stable Diffusion* (1.5, XL, 3.5 Large), *FLUX.1 [dev]*, *Ideogram 2.0*, and *DALL-E 3*. The code and dataset is open-sourced and available at <https://aniketrege.github.io/cure/>.

1. Introduction

Cultural diversity brings a collective strength that can benefit all of humanity.

Robert Alan Arthur

Text-to-Image (T2I) systems [45, 48, 50, 53] are trained on web-scale data [11, 22, 55, 56], which is long-tail in nature [44]. This translates to hallucinations when generat-

*Corresponding author: aniketr@cs.wisc.edu

†Equal Advising



Figure 1. Three images of types of pottery generated by a state-of-the-art T2I system, DALL-E 3, with varying amounts of information in the prompt. The T2I system is consistently accurate at generating “ceramic diyas” (a), but hallucinates incorrect details for “jebena, from Ethiopia” (b) and “amphora of Hermonax, a type of pottery from Greece” (c). Specifying more information in the prompt *can help* (i.e. in c but not in b), but is an unreliable method to make the T2I system more culturally representative.

ing images of data in the tail, as the model has not seen enough examples during training. This training paradigm has been shown to amplify societal biases and stereotypes encoded in the training data [5, 12], including harmful and offensive content [7, 8]. We show a simple example of T2I system bias using a state-of-the-art T2I system [4] to generate images of culturally specific pottery in Fig. 1. Generating images with the text prompt “ceramic diyas” (a) gives **diverse, consistent and factual** outputs, while prompting with “jebena, from Ethiopia” (b) and “amphora of Hermonax, a type of pottery from Greece” (c) generate **diverse but low quality and inaccurate** images that do not resemble pottery (b), or miss culture-specific details of the artifact (c). Building generative models that faithfully represent the diversity of human preferences, values, and experiences across global cultures in this long tail requires a culturally-aware benchmarking and reliable bias measure-

Table 1. We tabulate existing cultural benchmarks for Text-to-Image Systems, organized by their contributions towards dataset design, quantitative metrics to measure cultural representativeness, and towards the user study. We also tabulate statistics of the CuRe dataset compared to existing cultural benchmark datasets. Here $|\mathcal{R}|$ is the number of cultural regions (countries), $|\mathcal{S}|$ the number of cultural axes (supercategories), $|\mathcal{C}|$ the number of cultural categories, and $|\mathcal{N}|$ the total number of cultural artifacts. We also note the number of T2I systems evaluated via a user study, and the number of T2I systems evaluated only quantitatively.

Work	Dataset				Quantitative Metrics				# T2I Systems				
	Crowd-sourced	Scalable	Category Hierarchy	$ \mathcal{R} $	$ \mathcal{S} $	$ \mathcal{C} $	$ \mathcal{N} $	New Metric	Img-Txt Sim	Img-Img Sim	Diversity	In User Study	Only Quant
Liu et al. [39]	✓			8	9	-	1095		✓	✓		1	1
Basu et al. [2]				27	-	10	-					2	2
Ventura et al. [65]				10	8	200	-	✓	✓	✓		3	6
Jha et al. [27]				135	-	-	-	✓	✓			1	1
Khanuja et al. [31]	✓			7	-	17	580			✓		2	2
Kannen et al. [30]	✓	✓		8	3	-	1000	✓			✓	2	2
Zhang et al. [75]				10	9	-	595	✓		✓	✓	3	3
Bayramli et al. [3]	✓	✓		10	3	-	150		✓			3	3
CuRe (Ours)	✓	✓	✓	64	6	32	300	✓	✓	✓	✓	3	6

ment, which is non-trivial. A typical method to measure bias is through large-scale user studies on crowdsourcing platforms [27, 30], leveraging human judgments for assessment. While this approach accurately reflects feedbacks from humans, it is expensive and difficult to scale. To mitigate this cost, prior works propose proxy scorers to estimate human judgments, such as computing similarity of generated images to real images with large neural encoders [31], realism metrics [6, 25], alignment of images to some desired attributes specified through text [31, 65] and cultural diversity [30]. We find that these proxy scorers empirically do not correlate well with human judgments of cultural representativeness and similarity to ground truth across popular T2I systems (Sec. 5). We also highlight the “*generative entanglement*” of existing proxy scorers: a miscalibrated estimate of T2I system quality due to overlapping training data between the scorer and T2I system (Sec. 4.2).

Research Aim

Our goal is to measure the *cultural representativeness capability* of T2I systems across global cultures, which we call CuRe.

To overcome these shortcomings, we propose CuRe, a dataset and scorer suite to accurately and efficiently benchmark the cultural representativeness of T2I systems, *i.e.* *how equipped are state-of-the-art T2I systems at accurately generating samples across global cultures that make up their long-tail training data?* We propose a novel framework for scoring cultural representativeness through the lens of **marginal utility of increasing attribute specification** (Sec. 4.1), which correlates better to human judgments than proxy scorers across T2I systems (see Tab. 2 - 4 in Sec. 5). To enable measuring this marginal utility of information, we create a benchmark dataset with a novel categori-

cal hierarchy of attributes, as seen in Fig. 2. We compare our dataset and metric design to prior work in Tab. 1 and summarize our contributions below:

- A **new dataset** constructed in a scalable fashion directly from the large crowdsourced Wikimedia knowledge graph [69] with a **novel coarse-to-fine categorical hierarchy** of 300 cultural artifacts across six cultural axes, 32 cultural categories and 64 countries (see Fig. 2).
- A novel scoring of cultural representativeness through the lens of **marginal utility of specifying more information** to the T2I system across cultural attributes (*e.g.* cultural axis, cultural category, and cultural region).
- A **large-scale user study** asking real humans to rate the perceptual similarity, cultural representativeness, offensiveness, and stereotypicalness of T2I systems, alongside detailed freeform feedback about culturally specific failures. We query workers who explicitly identify with the culture of their country of nationality, which is largely an unverified assumption made in prior works.
- A detailed analysis of how CuRe **scorers correlate to real human judgments** of cultural representativeness and factuality, which highlights the misleading takeaways of popular status quo scorers (Sec. 5).
- For the first time, we evaluate the cultural capabilities of a natively multimodal large language model (Sec. 5.7).

2. Related Work

Dataset Biases Datasets used to train generative models are known to have biases, including the Amerocentric and Eurocentric *geographical distribution bias* [14, 58] of crowd-labeled datasets such as ImageNet [52], Open Images [33] and MS COCO [37]. Several works also highlight the *data collection bias*; *e.g.*, on YFCC100m [60], the data from underrepresented countries is often taken by tourists ($\sim 47\%$ [67]), and does not capture the true local distribu-

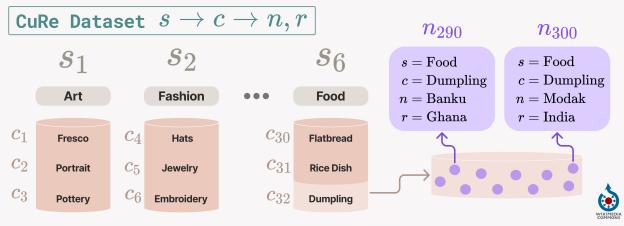


Figure 2. The CuRe dataset is constructed in a hierarchical fashion ($s \rightarrow c \rightarrow n, r$) from Wikimedia with 300 cultural artifacts $n \in \mathcal{N}$ grouped into 6 cultural axes $s \in \mathcal{S}$, 32 cultural categories $c \in \mathcal{C}$, described by a name n and associated region r . For example, $s = \text{"Food"}$, $c = \text{"dumpling"}$, $n = \text{"banku"}$, $r = \text{"Ghana"}$.

tion of objects, people, and language. These dataset biases propagate to T2I systems after training.

Text-to-Image System Biases T2I systems are predominantly built on diffusion models [50, 53], autoregressive Transformer models [17, 48, 73] or GANs [29, 54]. The pretraining datasets of these models are web-scale [56] and long-tail [44], and there have been several recent works examining the biases present within T2I systems. Several recent works analyze the biases of T2I systems from a geographical [2], gender occupation [57], cultural [5, 64] and social [40, 41] perspective. Luccioni et al. [40] provide an excellent overview of T2I system biases, from the data collection [7] and filtering [49] to model training [70].

Cultural Representativeness Metrics To measure CuRe on benchmark datasets, prior works use proxy scorers such as using deep image encoders to compute similarity of generated images to real images [31], which contain similar dataset biases as T2I systems themselves [57]. Another class of scorers use realism metrics [6, 25], which ignore culture-specific nuance. Lastly, prior works evaluate image-text alignment to carefully chosen prompts [31, 65] and cultural diversity as a proxy for representativeness [30], which often empirically do not correlate well human judgments across global cultures (see Sec. 5).

Cultural Benchmarks To create cultural benchmarks to measure T2I biases, previous works typically crowdsource data, either directly from workers on online platforms [31, 46] or with massive inter-organization efforts [51]. Some works also rely on cultural experts to create the data [39, 75]. While these methods can give high quality data, they are expensive and inscalable, as to add new data to the benchmark, new workers or experts must be hired each time. To measure cultural biases, prior works rely on querying users across global cultures for their judgment on images generated by T2I systems [2, 30, 31]. We overcome these limitations by constructing our benchmark CuRe directly from the live Wikimedia graph [69], by traversing parent nodes (cultural axes) and grouping child nodes (cultural categories) by region. This methodology is both cheap

and scalable, as new categories can be added to our benchmark on-the-fly by crawling Wikimedia.

Multimodal Language Models There has been a significant recent effort towards extending language models, which only understand text, to multimodal large language models (MLLMs) that can understand both text and images [15, 38, 61, 68, 74]. While the details of the pre-training data of state-of-the-art MLLMs is typically hidden or proprietary, they are significantly larger than previous Vision-Language Models like CLIP [47, 56] and perform very well on complex visual question answering benchmarks [1, 51]. They can thus be directly queried for culture-specific knowledge, similar to existing works that use MLLMs as a judge [9]. Even at this scale of pretraining data, we show that MLLMs still fall short at evaluating cultural representativeness (Sec. 5.7).

3. CuRe Dataset

Creating a high quality cultural benchmark is non-trivial for two primary reasons: a) it requires significant crowdsourcing efforts [30, 31, 51] or hiring domain experts [39, 75] b) it requires good “cultural coverage”, i.e. collecting cultural artifacts across a large number of cultural regions r and cultural categories. We address the first difficulty by designing a scalable dataset construction methodology that enables democratic scaling, as any cultural artifact of interest can easily be added to the benchmark by querying Wikimedia¹. We address the second difficulty by collecting cultural artifacts across 64 countries, which is higher than all existing cultural benchmarks (Tab. 1).

The necessity of Categorical Hierarchy *Culture* has a sense of shared values through lived experiences and one’s surroundings (intra-culture), which differ greatly across geographies (inter-culture) [26]. An important goal towards accurately measuring cultural representativeness of T2I systems is to capture their behavior at both inter and intra-culture levels. For instance, *Cuisine* is considered an important axis of culture [13], yet it is difficult to compare how well T2I systems do at generating cuisine of, for example, the United States compared to Nigeria, as cuisine has very high *intra-class variance* (diversity). We show that aggregating bias measurement at such coarse levels can lead to misleading takeaways about T2I performance (see diversity measured by Vendi scores in Sec. 5.5). To mitigate this, for each cultural axis, we propose comparing T2I performance at a finer granularity of cultural categorization. For example, many cultures around the world have their own form of dumpling. While all these forms have the same core structure (i.e. much lower intra-class variance than cuisine as a whole), they vary greatly in their ingredients, preparation,

¹We open-source our dataset at <https://github.com/aniketrege/cure-bench>

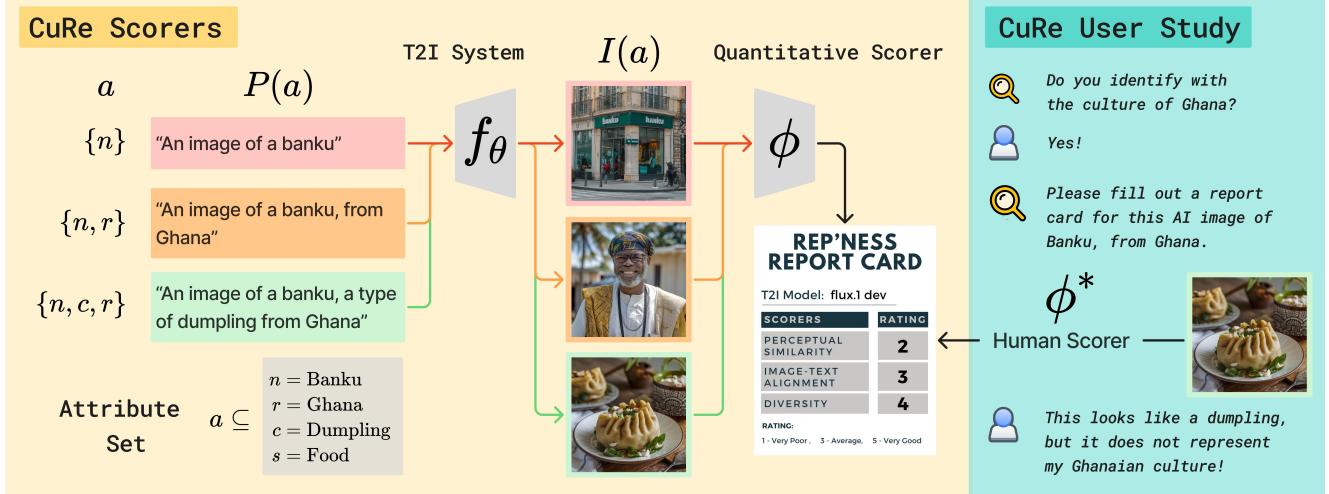


Figure 3. An overview of the scorers and user study of CuRe. Through the lens of marginal utility of attributes, we analyze images $I(a)$ generated from prompts templates $P(a)$ over attribute subsets $a \subseteq \{s, c, n, r\}$ of the 300 cultural artifacts $n \in \mathcal{N}$ in our benchmark dataset. For example, we generate an image from name and region attributes as $I(\{n, r\}) = f_\theta(P(\{n, r\}))$. This is the middle image (orange) in the figure above. These images are then scored by quantitative scorers ϕ (see Sec. 4) and user judgments from a large-scale user study ϕ^* (see Sec. 5.1). Crucially, cultural artifacts from region r are rated only by workers who identify with the culture of region r .

presentation, etc. specific to each culture. We design the CuRe dataset with a novel coarse to fine hierarchy to capture these nuances of cultural categorization.

CuRe consists of a set of 300 cultural artifacts $n \in \mathcal{N}$ organized in a categorical hierarchy of attributes: a name n (e.g. “modak”), a cultural category c (e.g. “dumpling”), a cultural supercategory s (e.g. “food”), and a cultural region of origin r (e.g. India). The CuRe dataset has six cultural axes (or supercategories) \mathcal{S} , i.e. architecture, art, celebrations, fashion, food, people, and each supercategory $s \in \mathcal{S}$ contains exactly 50 artifacts. We show an illustration of the CuRe dataset ($s \rightarrow c \rightarrow n, r$) in Fig. 2 and tabulate the $s \rightarrow c$ hierarchy of the dataset in Tab. 8 (Appendix B).

4. Measuring CuRe

We describe our proposed measure of cultural representativeness capabilities of a T2I system, CuRe, through the novel lens of marginal utility of information in Sec. 4.1. We introduce this lens into three prominent existing classes of quantitative scorers: approximating human **perceptual similarity** via cosine distance between the generated image and the ground truth (Sec. 4.2), evaluating **image-text alignment** with a desired attribute specified through text as a proxy for visual question-answering (Sec. 4.3), and measuring the **diversity** of images generated by the T2I system across global cultures (Sec. 4.4). We begin by setting up some useful notation.

Notation: To generate an image of a cultural artifact n , we must choose how to describe it to a T2I system using some subset of artifact attributes $a \subseteq \{s, c, n, r\}$. To this

end, we construct a text prompt P with a template over these attributes a . We summarize an exhaustive list of templates we use in Tab. 6 (Appendix B). We then pass this text prompt $P(a)$ to a T2I system f parametrized by weights θ , which generates an image I , i.e.

$$I(a) = f_\theta(P(a)), \text{ where } a \subseteq \{s, c, n, r\}$$

We show an example of three prompt templates and their corresponding images over three attribute subsets for banku, a type of dumpling from Ghana in Fig. 3. As seen in the figure, for this cultural artifact, the choice of a has a large impact on the generated image. Choosing $a = n$ or $a = \{n, r\}$ generates images from incorrect classes, i.e. a bank (in red) and a man (in orange). When we add category information informing the T2I system that banku is a type of dumpling, i.e. $a = \{n, c, r\}$, the image generated (in green) much more resembles real images of banku. To measure cultural representativeness capability of T2I system f_θ for region r , we need a method of scoring the quality of generated images I . Let $\phi : I \rightarrow \mathbb{R}$ be a quality scorer for CuRe (e.g. similarity of I to ground-truth images of artifact n).

Choosing a Scorer: The gold standard for quality scorer ϕ is to survey a large number of people from region r to rate T2I system performance along pre-defined rubrics such as realism and image-text alignment, typically with a 1-5 scale Likert score [36], which we denote ϕ^* (see Fig. 3). We interchangeably refer to Likert scores from the user study ϕ^* as “**gold scores**”, as they are rated by humans who identify with the culture of r . Designing surveys to collect these

human perceptual scores in this manner is non-trivial: eliciting calibrated scores is difficult and launching surveys is expensive [35]. If enough people are not queried in the survey, there is also the possibility their scores may not correlate well with the opinions of people who will use the T2I system after deployment [16]. To get around these difficulties, existing works design automated quantitative proxy scorers for cultural representativeness based on related but distinct goals like image perceptual similarity, image-text alignment, and diversity. These class of scorers have unique strengths and weaknesses, and we find they empirically do not correlate strongly to human judgments of quality (Sec. 5).

4.1. Finding a CuRe through Information

Our insight towards designing a more accurate, reliable scorer for CuRe across global cultures was motivated by a question: *how many attributes of a cultural artifact does a T2I system need to know in order to faithfully generate an image?* We illustrate our insight through an example in Fig. 3. We observe that simply specifying the name n and cultural region $\{n, r\}$ associated with banku, a type of dumpling from Ghana, is insufficient for current state-of-the-art T2I systems (Stable Diffusion 3.5 Large, FLUX.1 [dev], Ideogram 2.0) to produce an image faithful to its real-world counterpart. When also specifying the category $\{n, c, r\}$, the T2I system is able to generate an image that appears more faithful to banku. When this image is passed to existing quantitative scorers ϕ , they tend to overestimate its cultural representativeness (see examples in Fig. 6). However, when this image is shown to a real human who identifies as culturally Ghanaian (ϕ^*), they highlight that this image looks like a generic dumpling and does not reflect their cultural context of Ghanaian cuisine. Inspired by this, we propose a hypothesis complementary to these existing quantitative scorers:

Key Insight

Evaluating how the change in information *explicitly provided* to a T2I system changes its behavior reveals valuable insights into its cultural representativeness capabilities.

Unlike existing methods, this approach measures how well a T2I system internalizes cultural knowledge by analyzing the **marginal utility of each additional attribute** specified during generation. For example, with the “banku” artifact in Fig. 3, we need to marginally increase attributes specified to the T2I system via the prompt $a : n \rightarrow \{n, r\} \rightarrow \{n, c, r\}$ for culturally accurate T2I generation. We refer to this class of scorers as **Marginal Information Attribution** (MIA) scorers. For each class of proxy scorer, we show qualitative examples of the capability of MIA-

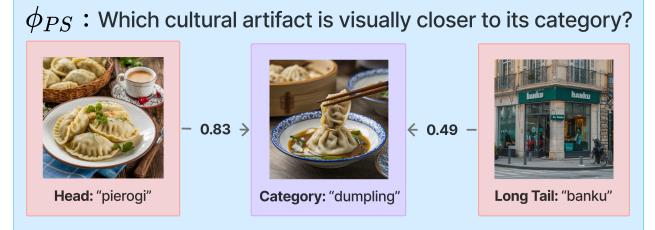


Figure 4. There is a large drop in $\phi_{PS} \in [0, 1]$, i.e. how visually similar images of cultural artifacts $I(n)$ are to images of their cultural category $I(c)$, from the *head* \rightarrow the *long tail* of the T2I system ($pierogi = 0.83 \rightarrow banku = 0.49$).

based scorers to differentiate cultural artifacts with vastly different human-rated scores for perceptual similarity and representativeness, while existing scorers are unable to do so, in Fig. 6. We demonstrate this behavior quantitatively over the entire CuRe dataset in Sec. 5. We formally define the variant of MIA scorer for each class of proxy scorer in the relevant section below.

4.2. Perceptual Similarity Scorers

The goal of perceptual similarity (PS) scorers is to compute the similarity between a generated image I and another set of images representing the same artifact n , which are typically manually-curated ground-truth (GT) images G . While these scorers correlate reasonably well with human perceptual similarity judgments, collecting appropriate ground-truth is expensive and occasionally infeasible. The gold scorer for perceptual similarity is $\phi_{PS}^*(a) = \text{Likert}(I(a), G)$, where a Likert score [36] of 1 indicates very low perceptual similarity and 5 indicates very high similarity. As getting these gold scores is inscalable and expensive, proxy quantitative scorers are used [18, 24], which we call ϕ_{GT} . The proxy scorers often ignore when two images are semantically similar in favor of spatial and textural consistency [23], which causes a discrepancy between real human perceptual similarity ratings (see Fig. 6a).

To overcome these limitations, we propose a **marginal information attribution** scorer, ϕ_{PS} , that compares T2I image features of cultural artifacts specified only by their name $a = n$ to images generated with only their categorical and region information $a = c$ or $a = \{c, r\}$. We hypothesize that if these images are similar, the T2I system has learned the artifact’s cultural association well (e.g. $a : n \rightarrow \{c, r\}$) and the artifact likely lies in the head of the T2I system’s distribution (Fig. 4 for $n = \text{"pierogi"}$). If these images are very different (Fig. 4 for $n = \text{"banku"}$), the T2I system has not learned categorical or region-specific associations well, and we hypothesize it lies in the long tail and shows poor CuRe performance. For example, while ϕ_{PS}^* and ϕ_{GT} compute similarity of T2I generations of banku $I(n)$ to ground-truth images collected from a web database

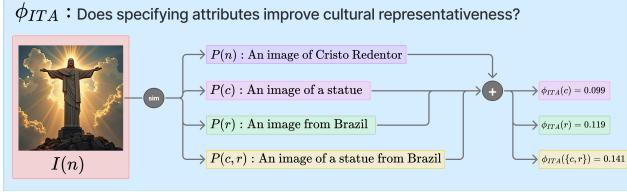


Figure 5. Computation of image-text alignment score $\phi_{ITA}(a)$ on $n = \text{"Cristo Redentor"}$ when changing attribute $n \rightarrow a$. As in Eq. (3), to compute $\phi_{ITA}(a)$, a similarity score sim is taken between image $I(n)$ and text prompt $P(a)$, and this is added to the similarity score between $I(n)$ and $P(n)$.

$G, \phi_{PS}(n)$ instead computes similarity of T2I generations of banku $I(n)$ to T2I generations of dumplings $I(c)$:

$$\phi_{GT}(a) = sim(I(a), G) \quad (1)$$

$$\phi_{PS}(a) = sim(I(a), I(c)) \quad (2)$$

Following the status quo for semantic similarity [28], sim is a cosine distance between embeddings from state-of-the-art large vision encoders (SigLIP 2 [63], AIMV2 [20], and DINOv2 [43]). We note that since ϕ_{PS} compares only generated images, it is much cheaper to compute than ϕ_{GT} and ϕ_{PS}^* , which require high-quality ground-truth data G and for the latter, also finding and hiring workers to provide ϕ^* . A drawback of this class of scorer is in the presence of ambiguity, *i.e.* if a cultural artifact has multiple meanings. For example “damper” is both a culturally Australian homemade bread, as well as a device used to suppress vibrations in mechanical systems. ϕ_{PS} cannot differentiate these easily, as it has no access to ground-truth information. We avoid these cases by discarding ambiguous prompts during our CuRe dataset design. We show the capability of ϕ_{PS} at correlating to real human perceptual similarity judgments compared to these strong baselines qualitatively in Fig. 6a and quantitatively in Tab. 2.

4.3. Image-Text Alignment Scorers

The goal of image-text alignment (**ITA**) scorers is to compute an “alignment” or similarity between an image and a piece of text. CLIP [47] popularized using textual descriptions of ImageNet classes as zero-shot labels for image classification via unsupervised contrastive learning. In a similar vein, to evaluate cultural representativeness, previous works evaluate how “close” generated images $I(a)$ are to textual descriptions of desired attributes $P(a)$ in the aligned latent space with vision-language models (VLMs) like CLIP [47] and SigLIP [63]. We denote this class of image-text scorers by $sim(I(n), P(a))$. For example, to evaluate country-specific representativeness, Khanuja et al. [31] check CLIP similarity of $I(n)$ with $P(r) = \text{"This image is culturally relevant to [r]"}$. We note that prior works typically use CLIP trained on LAION-2B [55] as their VLM

of choice for image-text alignment, which has been shown to have an Amero and Euro-centric bias [2, 40] and is also part of the pretraining datasets of many popular T2I systems [29, 45, 50, 53, 73]. This overlap causes misleading over-estimations of quality, which we call **generative entanglement** (see Sec. 5.4 and Tab. 17 in Appendix G).

These image-text scorers assume that embeddings of images containing attribute $a = n$ are clustered close in the VLM latent space to embeddings of textual descriptions of n . In other words, they rely on the VLM’s ability to distinguish cultural relevance to different regions by seeing enough artifact-region associations $(n, c \rightarrow r)$ during training [51]. We show in Tab. 3 that VLM knowledge of this association is difficult to query explicitly as $sim(I(n), P(r))$, and indirectly querying this knowledge through the impact of changing $a : n \rightarrow r$ correlates better with human judgments, *i.e.* adding $sim(I(n), P(r))$ to $sim(I(n), P(n))$. We thus define ϕ_{ITA} as:

$$\phi_{ITA}(a) = \frac{sim(I(n), P(n)) + sim(I(n), P(a))}{2} \quad (3)$$

In contrast to baselines, when assessing an image of an artifact *e.g.* toquilla in Fig. 6b, our scorer $\phi_{ITA}(r)$ evaluates both the *visual correctness* of toquilla being a hat, *i.e.* $sim(I(n), P(n))$ and its *cultural linkage* to $r = \text{Ecuador}$, *i.e.* $sim(I(n), P(r))$.

4.4. Diversity Scorers

The goal of diversity (**DIV**) scorers is to capture the heterogeneity or intra-class variance of images generated by T2I systems [30]. In this work, we view diversity through the lens of culture, *i.e.* how culturally diverse are T2I systems, and can diversity be a predictor of cultural representativeness? We define **intra-category diversity** as the extent to which T2I images generated with underspecified prompts [2] of cultural artifacts $I(c)$ or $I(s)$ reflect culture-specific nuances of regional variants $I(n)$ or $I(\{n, r\})$. In other words, if we sample $I(c) = \text{"An image of a house"}$ [2] from a T2I system 100 times, how many of these 100 images will be $r = \text{Japanese}$ or Canadian houses? We also measure **intra-artifact diversity**, *i.e.* how diverse are multiple seeds of images of the same artifact $I(n)$, *e.g.* when sampling 100 images with $P(n) = \text{"an image of chicken biryani"}$, is there heterogeneity in the images of chicken biryani, or do they all appear visually homogeneous?

Existing works differ in the granularity at which they compute and aggregate diversity. LPIPS [76] computes an average over pair-wise dissimilarity of images across all cultural artifacts n associated with category c using deep features extracted from convolutional neural networks [34]. Similar to Perceptual Similarity scorers (Sec. 4.2), LPIPS (which uses encoders trained on ImageNet [52]) can ignore image semantics in favor of spatial, color, and textural con-

Images	Metric	ϕ	Images	Metric	ϕ	Images	Metric	ϕ
n_1	[43]	0.67	n_1	[31]	0.13	n_1	LPIPS(n)	0.72
	[63]	0.79		[65]	0.11		VS(c)	0.24
	ϕ_{CuRe}^*	0.83		ϕ_{CuRe}^*	1.00		ϕ_{CuRe}	0.93
	ϕ_{PS}^*	0.75		ϕ_{PS}^*	0.75		ϕ_{PS}	0.66
	$\phi_{PS} \downarrow$	0.49		$\phi_{ITA} \uparrow$	0.14		$\phi_{DIV} \downarrow$	0.57
	[43]	0.62		[31]	0.11		LPIPS(n)	0.70
n_2	[63]	0.71		[65]	0.09		VS(c)	0.24
	ϕ_{CuRe}^*	0.31		ϕ_{CuRe}^*	0.17		ϕ_{CuRe}	0.46
	ϕ_{PS}^*	0.44		ϕ_{PS}^*	0.15		ϕ_{PS}	0.40
	$\phi_{PS} \downarrow$	0.65		$\phi_{ITA} \uparrow$	0.01		$\phi_{DIV} \downarrow$	0.79
	AI : $I(n)$		AI : $I(n)$			AI : $I(n)$		
	Real : $G(n)$		Real : $G(n)$			Real : $G(n)$		

(a) Perceptual Similarity Scorer (ϕ_{PS}). n_1 : “Omurice”, n_2 : “Chicken Biryani”. Images were generated with SD 3.5 Large.

(b) Image-Text Alignment Scorer (ϕ_{ITA}). n_1 : “Sombrero”, n_2 : “Toquilla”. Images were generated with FLUX.1 [dev].

(c) Diversity Scorer (ϕ_{DIV}). n_1 : “Spaghetti and meatballs”, n_2 : “Saimin”. Images were generated with FLUX.1 [dev].

Figure 6. A qualitative comparisons of our proposed MIA scorers compared to baselines for three scorer classes: a) Perceptual Similarity (Sec. 4.2) ; b) Image-Text Alignment (Sec. 4.3) ; c) Diversity (Sec. 4.4). In the figure above, ϕ_{CuRe}^* and ϕ_{PS}^* are human judgments of cultural representativeness and perceptual similarity from our user study (Sec. 5.1) normalized from their original 1 - 5 Likert scale to a 0 - 1 scale. The top row of images (n_1) represent T2I generations of artifacts rated highly by humans ($\phi_{CuRe}^*, \phi_{PS}^* \rightarrow 1$), while the bottom row (n_2) represents low rated artifact generations ($\phi_{CuRe}^*, \phi_{PS}^* \rightarrow 0$). A lower divergence in perceptual similarity with marginally increasing information captured by our scorer ($\phi_{PS} \downarrow$) and a higher alignment of image features to textual queries for representativeness ($\phi_{ITA} \uparrow$) correspond to better human judgments ϕ^* . While diversity in image generation is desirable, we empirically confirm prior observations of an inverse relationship [30] between human judgments and quantitative measures of cultural diversity ($\phi_{DIV} \downarrow$). We observe that across scorer classes, our proposed scorers can differentiate between n_1 and n_2 , while baseline methods treat them nearly the same.

$$\phi_{DIV} : \text{Does changing attributes make generations more diverse?}$$

$$LPIPS \left(\begin{array}{cc} I(n) & I(n, r) \\ I(n, c) & I(n, c, r) \end{array} \right) = 0.7$$

Figure 7. Computation of diversity score ϕ_{DIV} on $n = \text{“modak”}$, a type of dumpling from India. ϕ_{DIV} computes an aggregate pairwise dissimilarity (LPIPS [76]) between multiple seeds of images generated with varying attribute specification, *i.e.* $I(n), I(n, c), I(n, r), I(n, c, r)$. In the example above, this encompasses $16c^2 = 120$ pairs.

sistency [23]. Another recent approach to scoring diversity is via Vendi Scores (VS) [21, 75], which quantify diversity by estimating the entropy of a kernel similarity matrix computed over all pairs of artifacts (a_i, a_j) belonging to same category c . While Vendi Scores capture intra-category heterogeneity, they lack any sense of each individual artifact’s image quality. Kannen et al. [30] propose quality-weighting the Vendi score (qVS) with a human preference reward model [71] to address this limitation.

We propose a modification to LPIPS that captures marginal information attribution, which we show in Fig. 7. For a given cultural artifact n , we consider a set of images generated with incrementally changing attributes, *i.e.* $\{I(n), I(\{n, c\}), I(\{n, r\}), I(\{n, c, r\})\}$. We compute LPIPS over each pair in this set and take an average, which we denote by:

$$\phi_{DIV} = LPIPS(n, \{n, c\}, \{n, r\}, \{n, c, r\}) \quad (4)$$

If $\phi_{DIV} \sim LPIPS(n)$, diversity is relatively unaffected

as we increase information specified to the T2I system from $n \rightarrow a$ (Fig. 20, Appendix H) and we hypothesize that artifact n lies in the head of the T2I system distribution. We demonstrate qualitatively (Fig. 6c) and quantitatively (Tab. 4) that LPIPS, VS, and QVS do not correlate well with human judgments of CuRe when compared to our scorer ϕ_{DIV} .

5. Experiments

In this section, we discuss user study design (Sec. 5.1) and how we measure the capability of our proposed scorers to approximate real human judgments in Sec. 5.2. We then discuss empirical setup and observations for each class of scorer in the relevant section below, *i.e.* Perceptual Similarity scorers ϕ_{PS} in Sec. 5.3, Image-Text Alignment scorers ϕ_{ITA} in Sec. 5.4, and Diversity scorers ϕ_{DIV} in Sec. 5.5. Lastly, we benchmark popular state-of-the-art T2I systems on our dataset with our proposed scorers in Sec. 5.6. Details on T2I inference and seeding used to compute all scores is provided in Appendix A.

5.1. User Study

To measure user judgments of T2I systems across global cultures, we hire workers on the crowdsourcing tool Prolific². We hire three workers per region (by country of nationality) to answer survey questions about the cultural artifacts from the CuRe dataset (Sec. 3) specific to their region. We ask each worker from region r to rate on a 1-5 Likert scale a generated image of artifact n for

²<https://www.prolific.com/>

1. Cultural representativeness ϕ_{CuRe}^* , i.e. “How likely can this image be found in your country?”
2. Perceptual similarity to ground-truth images ϕ_{PS}^* , i.e. “How similar is this AI image to these four real images?”
3. Likelihood that the image belongs to its ground-truth class ϕ_{GT}^* , i.e. “How likely is this an image of {n}”

We expect rational users to provide a similar rating for ϕ_{GT}^* and ϕ_{PS}^* - the distinction being that workers are not provided ground-truth reference images while rating ϕ_{GT}^* and must rely on their prior knowledge of artifact n . We discuss the UI, survey design choices, and survey questions asked to workers in Appendix C. We also compare our study to prior works in Tab. 9 (Appendix D). We pay workers the platform set minimum of \$8 per hour.

5.2. Scorer Correlation to Human Judgments

To validate the alignment of our proposed scorers (PS, ITA, DIV) to real human judgments, we compute a Spearman rank correlation ρ to three gold scores from the user study, i.e. ϕ_{CuRe}^* , ϕ_{PS}^* , and ϕ_{GT}^* . Spearman’s ρ is a nonparametric measure of rank correlation, i.e. how well can the relationship between quantitative proxy scorers and real human judgments be described by a monotonic function? A $\rho \rightarrow 1$ indicates a monotonically non-decreasing relationship (e.g. in Tab. 2, as our proposed ϕ_{PS} scores increase, human judgments of CuRe and perceptual similarity also increase). On the other hand, $\rho \rightarrow -1$ indicates a monotonically non-increasing relationship (e.g. in Tab. 4, as our proposed ϕ_{DIV} scores increase, human judgments of CuRe and perceptual similarity decrease). A $\rho \sim 0$ indicates a very weak correlation, i.e. the scorer contains very little predictive signal over human judgments. We show qualitatively in Fig. 6 that our proposed scorers can differentiate between images treated differently by real humans (i.e. ϕ^* are different), while baseline scorers treat these images the same. We discuss quantitative results over our entire CuRe dataset in each corresponding section below.

5.3. Perceptual Similarity

In this section, we compare our proposed PS scorers to human judgments of cultural representativeness ϕ_{CuRe}^* , perceptual similarity ϕ_{PS}^* and ground-truth label likelihood ϕ_{GT}^* , as motivated in Sec. 5.2. We tabulate a Spearman rank correlation across all artifacts n in the CuRe dataset in Tab. 2.

First, we compare how perceptually similar images of cultural artifacts $I(n)$ are to generated images of their cultural category $I(c)$, e.g. T2I generated images of n = “banku” and T2I generated images of c = “dumpling” respectively:

$$\phi_{PS}(n) = sim(I(n), I(c))$$

We find that our $\phi_{PS}(n)$ scorer is comparable to strong

Table 2. Spearman rank correlation ρ between perceptual similarity (**PS**) scorers and user judgments ϕ_{CuRe}^* , ϕ_{PS}^* , and ϕ_{GT}^* across state-of-the-art large image encoders and T2I Systems on the CuRe dataset. \uparrow indicates that $\rho \rightarrow 1$ is better, and \downarrow indicates $\rho \rightarrow -1$ is better. The highest magnitude of ρ in each column is bolded for the scorer that a) uses ground-truth images (ϕ_{GT} and Gemini) and b) does not use ground-truth images (ϕ_{PS} and $\Delta\phi_{PS}$).

Encoder	Scorer	FLUX.1 [dev]			SD 3.5 Large			SD 1.5		
		ϕ_{CuRe}^*	ϕ_{GT}^*	ϕ_{PS}^*	ϕ_{CuRe}^*	ϕ_{GT}^*	ϕ_{PS}^*	ϕ_{CuRe}^*	ϕ_{GT}^*	ϕ_{PS}^*
-	Gemini 2.0 Flash \uparrow	-	-	0.40	-	-	0.39	-	-	0.40
SigLIP 2 [63]	$\phi_{GT}(n) \uparrow$	0.25	0.36	0.44	0.27	0.33	0.45	0.25	0.39	0.51
	$\phi_{PS}(n) \uparrow$	0.18	0.25	0.32	0.22	0.29	0.38	0.18	0.27	0.30
	$\Delta\phi_{PS}(\{n, c\}) \downarrow$	-0.16	-0.27	-0.31	-0.21	-0.32	-0.37	-0.02	0.03	0.04
	$\Delta\phi_{PS}(\{n, c, r\}) \downarrow$	-0.17	-0.25	-0.30	-0.21	-0.33	-0.38	0.00	0.09	0.08
AIMV2 [20]	$\phi_{GT}(n) \uparrow$	0.20	0.28	0.39	0.24	0.32	0.44	0.20	0.35	0.45
	$\phi_{PS}(n) \uparrow$	0.08	0.17	0.25	0.06	0.14	0.24	0.09	0.11	0.12
	$\Delta\phi_{PS}(\{n, c\}) \downarrow$	-0.17	-0.30	-0.35	-0.19	-0.32	-0.31	-0.01	-0.02	0.02
	$\Delta\phi_{PS}(\{n, c, r\}) \downarrow$	-0.18	-0.28	-0.32	-0.16	-0.30	-0.33	-0.01	0.04	0.05
DINOv2 [43]	$\phi_{GT}(n) \uparrow$	0.17	0.29	0.40	0.25	0.36	0.46	0.24	0.39	0.52
	$\phi_{PS}(n) \uparrow$	0.13	0.22	0.30	0.15	0.29	0.35	0.15	0.27	0.30
	$\Delta\phi_{PS}(\{n, c\}) \downarrow$	-0.19	-0.27	-0.32	-0.21	-0.32	-0.35	-0.07	-0.05	-0.01
	$\Delta\phi_{PS}(\{n, c, r\}) \downarrow$	-0.19	-0.25	-0.31	-0.20	-0.31	-0.34	-0.08	0.00	-0.02

baseline $\phi_{GT}(n) = sim(I(n), G)$ in Spearman’s ρ with the gold scores ϕ_{GT}^* across all T2I systems and image encoders, despite using no ground-truth information (Tab. 2).

Next, recall that we claim that the relative importance or marginal utility of increasing attribute specification $a \subseteq \{n, c, r, s\}$ to the T2I system can provide valuable signal towards its cultural representativeness capabilities (Sec. 4.1). To more directly measure this marginal utility, we evaluate how much the *change* in attribute specification ($n \rightarrow \{n, c\}$ and $n \rightarrow \{n, c, r\}$) affects perceptual similarity by computing a divergence Δ of $\phi_{PS}(a)$ from $\phi_{PS}(n)$ evaluated at attribute subsets $a = \{n, c\}$ and $a = \{n, c, r\}$:

$$\begin{aligned}\Delta\phi_{PS}(\{n, c\}) &= 0.5 + \phi_{PS}(\{n, c\}) - \phi_{PS}(n) \\ \Delta\phi_{PS}(\{n, c, r\}) &= 0.5 + \phi_{PS}(\{n, c, r\}) - \phi_{PS}(n)\end{aligned}$$

We add a 0.5 scaling constant to our divergence scorers to bring them to a similar scale as ϕ_{GT} and ϕ_{PS} .

On FLUX.1 [dev] and Stable Diffusion 3.5 Large, our divergence scorers $\Delta\phi_{PS}(\{n, c\})$ and $\Delta\phi_{PS}(\{n, c, r\})$ match or outperform $\phi_{GT}(n)$ in Spearman’s ρ across gold scores with AIMV2 and DINOv2 encoders. We note that rank correlations of divergence scorers with gold scores are always negative, since a low divergence with marginally increasing information indicates high perceptual similarity (denoted by \downarrow in Table 2). Overall, for all image encoders, all quantitative PS scorers ϕ_{GT} , ϕ_{PS} and $\Delta\phi_{PS}$ correlate weaker with gold CuRe scores ϕ_{CuRe}^* than gold ground truth likelihood scores ϕ_{GT}^* and gold perceptual similarity scores ϕ_{PS}^* , which they directly attempt to approximate.

Lastly, we also evaluate Gemini 2.0 Flash [15], a strong natively multimodal large language model (see Sec. 5.7). We query Gemini to score perceptual similarity between $I(n)$ and ground truth images G on a 1 to 5 scale, similar to the setup for $\phi_{GT}(n)$ and ϕ_{PS}^* from the user study. We

Table 3. Spearman rank correlation ρ between image-text alignment (**ITA**) scorers and user judgments ϕ_{CuRe}^* , ϕ_{PS}^* , and ϕ_{GT}^* across T2I systems on the CuRe dataset. All scorers except Gemini and human preference scorers (first block below) compute a cosine distance with SigLIP 2 embeddings. The highest magnitude of ρ in each column is bolded, and $\rho \rightarrow 1$ is the optimum scorer.

Scorer	FLUX.1 [dev]			SD 3.5 Large			SD 1.5		
	ϕ_{CuRe}^*	ϕ_{GT}^*	ϕ_{PS}^*	ϕ_{CuRe}^*	ϕ_{GT}^*	ϕ_{PS}^*	ϕ_{CuRe}^*	ϕ_{GT}^*	ϕ_{PS}^*
PickScore [32]	0.20	0.29	0.34	0.23	0.37	0.40	0.23	0.38	0.45
Imagereward [72]	0.19	0.21	0.26	0.16	0.24	0.31	0.23	0.30	0.35
HPS v2 [71]	0.23	0.29	0.33	0.18	0.35	0.37	0.24	0.39	0.43
Gemini 2.0 Flash	0.23	0.41	-	0.27	0.37	-	0.17	0.38	-
Khanuja et al. [31]	0.13	0.08	0.11	0.05	0.04	0.00	0.06	-0.02	-0.04
Ventura et al. [65]	0.19	0.15	0.14	0.10	0.07	0.05	0.11	0.02	0.01
o3-mini [42]	0.17	0.13	0.14	0.06	0.03	0.00	0.06	-0.01	-0.03
$\text{sim}(I(n), P(n))$	0.24	0.35	0.38	0.18	0.31	0.35	0.22	0.34	0.44
$\text{sim}(I(n), P(c))$	0.20	0.32	0.34	0.17	0.30	0.35	0.19	0.34	0.37
$\text{sim}(I(n), P(r))$	0.20	0.11	0.12	0.09	0.03	0.02	0.07	-0.04	-0.04
$\text{sim}(I(n), P(\{c, r\}))$	0.25	0.35	0.38	0.22	0.31	0.36	0.24	0.34	0.38
$\phi_{ITA}(c)$	0.24	0.37	0.40	0.20	0.34	0.40	0.22	0.37	0.45
$\phi_{ITA}(r)$	0.28	0.33	0.35	0.20	0.27	0.29	0.23	0.29	0.39
$\phi_{ITA}(\{c, r\})$	0.27	0.38	0.42	0.23	0.34	0.39	0.24	0.36	0.44

observe that Gemini correlates almost identically well with ϕ_{PS}^* across T2I systems, and slightly lags behind $\phi_{GT}(n)$ in Spearman’s ρ .

We highlight that while $\phi_{GT}(n)$ and Gemini in general show the highest correlation with human perceptual judgments, our *proposed scorers are capable of nearly matching them in rank correlation with real human judgments while using no ground-truth information*, and are thus much cheaper to compute. We note that no quantitative scorer surpasses $\rho = 0.51$, indicating that using cosine similarity with dense vector embeddings from large vision encoders still fall short at accurately approximating human judgements, even with large state-of-the-art encoders [20, 43, 63].

Discussion: Failure Mode of our Scorer: While our PS scorers approximate gold scores well for FLUX.1 [dev] and SD 3.5 Large, divergence scorers $\Delta\phi_{PS}$ correlate very weakly with gold scores for SD 1.5, an older lower-resolution T2I system trained on much less data. In contrast, the baseline scorer $\phi_{GT}(n)$ maintains similar correlation across all T2I systems. This suggests that while our scorer, while not requiring ground truth images, relies more on the capacity of the T2I system to learn artifact - region associations than baselines. In other words, it is more sensitive to the coverage of the pretraining data, or the quality of the base model (e.g. Image Arena Quality ELO in Tab. 5).

5.4. Image-Text Alignment

Similar to PS scorers, we evaluate all quantitative ITA scorers via a Spearman rank correlation with gold scorers (ϕ_{CuRe}^* , ϕ_{GT}^* and ϕ_{PS}^*) in Tab. 3. Recall that to compute ϕ_{ITA} in Eq. (3), we require a similarity measure between images and prompts $\text{sim}(I, P)$. To evaluate the **generative entanglement** between T2I system and scorer, we use a suite of vision-language models (VLMs) with different pre-

training datasets as our similarity measure: four versions of OpenCLIP (trained on LAION-2B [56], Datacomp-1B [22], DataFN-5B [19], and OpenAI WIT [47]) and SigLIP 2 [63]). We begin by discussing results only with SigLIP 2, the current state-of-the art VLM.

First, we compare our scorer to three baseline scorers: two prior works evaluating cultural relevance [31] and cultural identity [65], and the best performing prompt $P(r)$ suggested by o3-mini [42]. We also evaluate alignment of image $I(n)$ to text prompts with marginally increased attribution specification, i.e. $P(n)$, $P(c)$, $P(r)$, and $P(\{c, r\})$.

Our evaluation reveals that comparing $I(n)$ to $P(n)$, which specifies only the artifact’s name in the prompt (e.g., “Qingming festival”), consistently has a stronger positive rank correlation with human judgments than all methods that compare to only $P(r)$ (baseline scorers [31, 42, 65] and $\text{sim}(I(n), P(r))$ across all T2I systems (Tab. 3 block 2). This supports our claim in Sec. 4.3 that explicitly querying VLMs for alignment to prompts describing the image by name $P(n)$ is more effective at assessing region-specific cultural relevance than directly querying for cultural relevance with region-specific prompts $P(r)$. Replacing $P(n)$ with category-specific prompts $P(c)$ (e.g. “spring festival” for an image of Qingming festival), slightly reduces rank correlation with all gold scores across all T2I systems, which suggests that state-of-the-art VLMs learn $n \rightarrow c$ associations reasonably well. Lastly, using both category and region specific prompts $P(\{c, r\})$ tends to match and occasionally slightly outperform name-specific prompts $P(n)$ across T2I systems, indicating that switching to categorical and regional information ($n \rightarrow \{c, r\}$) is important to capture culture-specific human judgments.

Next, we compare our scorers to state-of-the art human preference reward model scorers [10] trained on a million human-labeled pairwise preferences over T2I system generations, i.e. PickScore [32], Imagereward [72] and HPS v2 [71]. Lastly, similar to PS scorers, we also evaluate Gemini 2.0 Flash (details in Sec. 5.7).

Recall that our scorers ϕ_{ITA} directly measure the marginal utility of increasing attributes specified to the T2I system in two parts, i.e. the T2I system’s ability to a) generate images that match textual descriptions of the artifact ($\text{sim}(I(n), P(n))$), b) capture an artifacts categorical ($\text{sim}(I(n), P(c))$) and regional associations ($\text{sim}(I(n), P(r))$). We observe that $\phi_{ITA}(\{c, r\})$ has higher positive rank correlation than all baselines (Tab. 3 block 2) across T2I systems, and is slightly outperformed by preference reward models trained on hundreds of thousands of human-labeled preferences and Gemini on Stable Diffusion T2I systems for ϕ_{PS}^* and ϕ_{GT}^* ($\phi_{ITA}(\{c, r\}) \geq \text{PickScore} > \text{HPS v2} > \text{Imagereward}$). Our scorers outperform all methods including Gemini for FLUX.1 [dev].

Remark: Generative Entanglement. So far we have discussed ITA scorer results only with SigLIP 2 as a base VLM, the current state-of-the-art for image-text similarity (Tab. 3). For a more rounded examination, we also evaluate the impact of changing the base VLM as a similarity measure in the context of *generative entanglement*. In Tab. 17 (Appendix G), we show that with FLUX.1 [dev], all baseline methods have a high variance in Spearman’s ρ to human judgments of perceptual similarity (ϕ_{PS}^*) with respect to the choice of similarity measure (OpenCLIP variant or SigLIP 2). In contrast, our scorers have lower variance across VLMs. As these VLMs differ primarily in their pre-training data, this result indicates that our scorers ϕ_{ITA} are much more robust to the choice of pretraining data and are less entangled with the T2I system.

5.5. Diversity

We next show a Spearman’s rank correlation between all quantitative scorers ϕ_{DIV} , LPIPS(n) and VS(c) and all gold scorers ($\phi_{CuRe}^*(a)$, $\phi_{GT}^*(a)$, and $\phi_{PS}^*(a)$) in Tab. 4. Consistent with our observations with PS and ITA scorers, we show that ϕ_{DIV} correlates more to human judgments than baselines across all T2I systems. This difference is more pronounced for modern, larger, high quality T2I systems (FLUX.1 [dev] and SD 3.5 Large) compared to a smaller, low quality T2I system (SD 1.5). See remark on Image Arena ELO remark in Sec. 5.6 for a discussion on overall T2I system quality.

First, we consider LPIPS(n), *i.e.* a pairwise dissimilarity between seeds of a single cultural artifact n (**intra-artifact**), averaged over all artifacts in the CuRe dataset $n \in \mathcal{N}$. LPIPS(n) shows a weak negative correlation with gold scores, but outperforms Vendi Score [21] based scorers VS(c) and qVS(c) substantially, which are nearly uncorrelated ($\rho \sim 0$). As VS(c) is computed as an aggregate at a coarse category level (**intra-category**) on static assignments over cultural artifacts and has no sense of image quality (see Appendix H for details on VS computation), it shows the weakest correlation with human judgments for strong T2I systems (SD 3.5 Large and FLUX.1 [dev]). The recent quality-weighted Vendi Score [30] fails to resolve these limitations, as it simply scales VS(c) by a single scalar quality measure q averaged over items (HPSv2 [71]), leaving the rank correlation with human judgments unchanged.

In contrast, our MIA-based scorer ϕ_{DIV} achieves stronger negative correlations with human judgments than baselines for all T2I systems, especially those of higher quality (SD 3.5 Large and FLUX.1 [dev]). In other words, if by mixing images generated by the T2I system $I(n)$ with images generated with more information specification in the prompt $I(\{n, c\})$, $I(\{n, r\})$, and $I(\{n, c, r\})$, we observe a drop in diversity, human judgments of CuRe and perceptual similarity improve. In this case, the cultural artifact n is ho-

Table 4. Spearman rank correlation ρ between diversity (**DIV**) scorers and user judgments ϕ_{CuRe}^* , ϕ_{PS}^* , and ϕ_{GT}^* across T2I systems on the CuRe dataset. The highest magnitude of ρ in each column is bolded, and $\rho \rightarrow -1$ is the optimum scorer.

Scorer	FLUX.1 [dev]			SD 3.5 Large			SD 1.5		
	ϕ_{CuRe}^*	ϕ_{GT}^*	ϕ_{PS}^*	ϕ_{CuRe}^*	ϕ_{GT}^*	ϕ_{PS}^*	ϕ_{CuRe}^*	ϕ_{GT}^*	ϕ_{PS}^*
LPIPS(n)	-0.11	-0.06	-0.16	-0.13	-0.06	-0.09	-0.04	-0.03	-0.02
VS(c)	-0.02	-0.01	-0.01	0.03	0.02	0.05	0.01	-0.02	-0.02
qVS(c)	-0.02	-0.01	-0.01	0.03	0.02	0.05	0.01	-0.02	-0.02
ϕ_{DIV}	-0.20	-0.23	-0.30	-0.22	-0.23	-0.29	-0.07	-0.12	-0.11

mogeneous across marginally increasing attribute specification and likely lies in the head of the T2I system’s distribution. This aligns with our hypotheses in Sec. 4.4 and indicates the potential of marginal utility of information specification as a measure of diverse cultural representativeness.

Remark: Factuality-Diversity Tradeoff. We note that while T2I systems being culturally diverse is a desired outcome of training, a high diversity across all quantitative scorers is negatively correlated to gold scores ϕ_{CuRe}^* and ϕ_{PS}^* . This indicates there is a factuality-diversity trade-off with existing state-of-the-art T2I systems, which agrees with conclusions from prior work [30, 66].

5.6. Benchmark Results

We evaluate several popular state-of-the-art T2I systems on the CuRe benchmark dataset with our three scorer classes sorted by decreasing Image Arena T2I Quality ELO³, an overall measure of T2I quality, in Tab. 5. We note that due to inbuilt safety filters, DALL-E 3 and Stable Diffusion 1.5 refuse to generate 17% and 1.5% of images respectively (detailed breakdown in Appendix A.2). For all scorers, we use the variant which had the highest negative Spearman’s ρ with gold scores on average.

For perceptual similarity, we use our $\Delta\phi_{PS}(\{n, c\})$ scorer (Sec. 5.3). As this scorer is a divergence, a value closer to zero is better (\downarrow). As our scorer requires generating multiple seeds (80) of $I(c)$, we omit Ideogram 2.0 due to API compute constraints. SDXL and DALL-E 3 perform best for SigLIP 2 (SL2) cosine distance, while DALL-E 3 and FLUX.1 [dev] slightly edge out SD 1.5 and SDXL with DINOv2 (DN2). AIMV2 (AM2) has lower separability between T2I systems, with SD 3.5 Large and 1.5 slightly edging out the other T2I systems. We would also like to caveat that as our PS scorers show poor rank correlation to SD 1.5 (see discussion in Sec. 5.3), PS scores on SD 1.5 are likely to be overestimates.

For image-text alignment, we use $\phi_{ITA}(\{c, r\})$ (Sec. 5.4). Across VLM backbones, the Stable Diffusion class of T2I systems perform quite strongly ($XL \geq 3.5 > 1.5$) compared to FLUX.1 [dev], Ideogram 2.0 and

³<https://artificialanalysis.ai/text-to-image#quality>, scores taken on June 10, 2025

Table 5. We evaluate several state-of-the-art T2I systems on our CuRe benchmark across our three scorer classes: perceptual similarity (Sec. 4.2), image-text alignment (Sec. 4.3), and diversity (Sec. 4.4), and tabulate the mean and standard deviation ($\mu \pm \sigma$) across all 300 artifacts in the CuRe dataset. The best entry of each column is bolded, and the next best if the T2I system is DALL-E 3 or SD 1.5*, which have moderate refusal rates due to safety filters (see Appendix A.2). The last row shows a Spearman rank correlation ρ between the scorer and the Image Arena Quality ELO.

T2I System	ELO \uparrow	$\phi_{\text{CuRe}}^* \uparrow$	$\phi_{PS}^* \uparrow$	$\phi_{GT}^* \uparrow$	$\phi_{PS} \downarrow$			$\phi_{ITA} \uparrow$			
		SL2	DN2	AV2	SL2	L2B	WIT	ALX			
FLUX.1 [dev]	1045	2.814 \pm 1.424	2.157 \pm 1.141	2.251 \pm 1.326	0.561 \pm 0.110	0.575 \pm 0.137	0.523 \pm 0.049	0.094 \pm 0.051	0.218 \pm 0.059	0.209 \pm 0.039	0.708 \pm 0.078
Ideogram 2.0	1043	-	-	-	-	-	-	0.096 \pm 0.052	0.214 \pm 0.067	0.195 \pm 0.050	0.693 \pm 0.072
SD 3.5 Large	1028	2.986 \pm 1.439	2.396 \pm 1.228	2.534 \pm 1.393	0.567 \pm 0.107	0.604 \pm 0.166	0.532 \pm 0.056	0.115 \pm 0.047	0.251 \pm 0.053	0.225 \pm 0.036	0.670 \pm 0.082
DALL-E 3*	922	-	-	-	0.562 \pm 0.103	0.579 \pm 0.143	0.525 \pm 0.055	0.105 \pm 0.051	0.219 \pm 0.062	0.222 \pm 0.041	0.789 \pm 0.043
SDXL	840	-	-	-	0.557 \pm 0.100	0.579 \pm 0.151	0.520 \pm 0.049	0.113 \pm 0.051	0.255 \pm 0.056	0.230 \pm 0.039	0.753 \pm 0.042
SD 1.5*	587	2.724 \pm 1.412	2.094 \pm 1.159	2.175 \pm 1.291	0.559 \pm 0.104	0.576 \pm 0.142	0.519* \pm 0.041	0.107 \pm 0.050	0.240 \pm 0.055	0.229 \pm 0.035	0.755 \pm 0.057
p with ELO	1.00	-	-	-	0.564	-0.100	0.564	-0.600	-0.657	-0.829	-0.600

DALL-E 3, and this is especially pronounced for CLIP trained on LAION-2B. As Stable Diffusion systems were trained on LAION-2B, we suspect their strong performance on LAION is because of this overlap with the training set of our scorer, which was also trained on LAION-2B (**generative entanglement**, see Sec. 4.2 for details).

Lastly, for ϕ_{DIV} , which is an LPIPS score computed over a mixture of seeds across prompt styles, DALL-E 3 outperforms SDXL and SD 1.5, which are substantially ahead of the rest. It is noteworthy that DALL-E 3 had a 17% refusal rate across all seeds of images on the CuRe dataset, which likely caused a slight inflation in its diversity score (as LPIPS is computed across all pairs, generating fewer seeds will drop diversity less). Across all our scorers, Stable Diffusion XL performs the most consistently, followed by DALL-E 3.

Remark: Relation to Image Arena ELO To compare to crowdsourced user judgments of overall T2I system quality, we compute a Spearman rank correlation of our scorers with Image Arena ELO scores. Image Arena pits two T2I systems in a “battle”, where a user observes images generated by each system with a fixed prompt and picks their preferred image. Image Arena has collected 100K+ crowdsourced pairwise preferences to compute ELO for 50+ popular T2I systems. To begin with, we highlight that *Image Arena ELO does not perfectly correlate with human judgments from our user study*: on a 1-5 Likert Scale, the most culturally representative (ϕ_{CuRe}^*) T2I system (SD 3.5) is only 0.2 points (5%) better than the least culturally representative T2I system (SD 1.5), while they are separated by nearly 500 ELO points, *i.e.* a 93% chance that SD 3.5 will beat SD 1.5 in a head-to-head pairwise comparison. For perceptual similarity (ϕ_{GT}^*, ϕ_{PS}^*), the point difference only increases to 0.36 (*i.e.* SD 3.5 is 9% better than SD 1.5).

Our PS divergence scorer $\Delta\phi_{PS}$ shows weak negative rank correlation to Image Arena ELO with DINOv2 embeddings ($\rho = -0.100$), but a moderately strong positive correlation with SigLIP 2 and AIMV2 embeddings ($\rho = 0.564$). This suggests that the degree to which perceptual similarity of T2I system is indicative of overall T2I system quality

evaluated via pairwise preferences may be sensitive to the choice of embedding backbone. With two popular backbones (SigLIP 2 and AIMV2), Spearman’s ρ suggests an inverse relationship between a T2I system’s ability to generate images closely matching ground truth images and overall human judgments of quality (“is x better or y ?”), which may be dominated by aesthetic appeal or other factors [4].

Our ITA scorer has a strong negative correlation with Image Arena ELO, especially with OpenAI CLIP as a VLM backbone (WIT, $\rho = -0.829$). In other words, with increasing T2I system quality, our scorer predicts lower image feature alignment to the text prompt. We suspect that this seemingly counter-intuitive result stems from a key difference in evaluation setup: our scorer uses “underspecified prompts” [2] $I(n)$, *i.e.* the T2I generations only have the attribute name specified, whereas Image Arena ELO scores are based on detailed, descriptive prompts. In such underspecified cases, prior work has shown T2I systems with higher quality (ELO) have more “creativity” and tend to prioritize aesthetic visual appeal over actuality [4]. We also find from our user study that higher quality T2I systems rely on cultural stereotypes for generation (Appendix D.4), which negatively impacts user judgments of quality.

Lastly, as high quality T2I systems like FLUX.1 [dev] and SD 3.5 Large have seen orders of magnitude more data during training, they can generate more diverse renditions of underspecified prompts (which we also empirically observe through high ϕ_{DIV} scores for T2I systems with low ELO). As diversity is often at odds with factuality [66], this phenomenon ends up hurting the alignment of higher quality T2I systems to simple text prompts, *i.e.* $P(c, r)$.

5.7 Multimodal LLM as a Judge

To our knowledge for the first time, we perform a preliminary, exploratory analysis of using a multimodal large language model (MLLM) to approximate human judgments (gold scores) of cultural representativeness (ϕ_{CuRe}^*) and image perceptual similarity (ϕ_{PS}^*, ϕ_{GT}^*).

Specifically, we query Gemini 2.0 Flash [15] with the same set of questions asked to real humans in our user study

(details in Appendix I), and output a score from 1 (low) to 5 (high), *i.e.* the same scale as the Likert scores from our user study (Sec. 5.1). We also query Gemini for a textual justification of its scores to inspect its reasoning. We follow the same evaluation setup as PS, ITA and DIV scorers (Sec. 5.2) and compute a Spearman rank correlation between Gemini scores and gold scores from our user study. For PS scorers Tab. 2, we observe that Gemini slightly lags behind ϕ_{GT} , and our scorers approximate both these strong baselines well without access to any ground-truth images. For ITA scorers (Tab. 3), we observe that Gemini 2.0 Flash nearly matches (Flux.1 [dev] and SD 1.5) and occasionally outperforms (SD 3.5) our proposed ϕ_{ITA} scorers in Spearman’s ρ , suggesting a pretraining dataset has wider cultural coverage of artifact - region associations compared to smaller VLMs like OpenCLIP and SigLIP 2. Lastly, we examine Gemini’s textual justifications for its scores to inspect the reliability of its reasoning (see Fig. 22 in Appendix I). We find that Gemini tends to overlook culture-specific details, even when attribute details may be correct (*e.g.* textures, shapes and patterns of an object that are not culturally accurate). This issue is especially prevalent in regions from the Global South (*e.g.* Nigeria, Ghana, Iran), where Gemini hallucinates culture-specific details (*e.g.* “jollof rice” from Ghana, “Femi Kuti” from Nigeria and “sami headwear” from Finland) and mistakenly highlights accuracy to the wrong culture (*e.g.* Chinese instead of Korean for “Chuseok” and Egyptian instead of Iranian in “Takht-e Fulad”).

In summary, these initial findings suggest that while a state-of-the-art multimodal LLM can show promise at evaluating cultural representativeness, it **requires access to ground truth images for reliability** (which none of our scorers require), and **hallucinates incorrect culture-specific details**, often in regions from the Global South.

6. Conclusion

In summary, we propose CuRe, a novel benchmarking and scoring suite for cultural representativeness that leverages the marginal utility of attribute specification to text-to-image systems as a proxy for human judgments. Our CuRe dataset has a novel categorical hierarchy that enables benchmarking T2I systems in this manner. We empirically observe much stronger correlations to user judgments across three quantitative scoring classes, *i.e.* perceptual similarity (PS, Sec. 5.3), image-text alignment (ITA, Sec. 5.4), and cultural diversity (DIV, Sec. 5.5).

We highlight that our PS scorer approximates strong baseline performance while using no ground-truth information, indicating strong potential for efficient and democratic benchmarking of T2I systems. Our ITA and DIV scorers (Tab. 4) outperform or match all baselines, including a strong multimodal large language model (Gemini 2.0 Flash, Sec. 5.7). Finally, we highlight the generative entanglement

issue of prior benchmarks and scorers, *i.e.* a miscalibrated estimation of human judgments caused by an overlap of T2I system and quantitative scorer pretraining data (Tab. 3).

Limitations and Future Work While we employ a categorical hierarchy of attributes, similar to prior work [2, 31], CuRe also uses geography as a proxy for culture (each artifact has a single associated country of origin). In the future, we would like to extend CuRe to a more holistic sense of culture, such as religion and spoken language. For perceptual similarity scorers, we find that overall T2I system quality determines the ability of our scorer ϕ_{PS} to match human judgments of cultural representativeness. When using a smaller and lower quality system (Stable Diffusion 1.5), our scorers were less correlated with human judgments than baseline scorers (Tab. 2), while this was not the case for strong systems (Flux.1 [dev], Stable Diffusion 3.5 Large). This scorer also suffers in the case of ambiguous prompts (*i.e.* artifact name n referring to multiple distinct artifacts), as it has no access to ground-truth image information to distinguish between them. We design our CuRe dataset to filter out such ambiguities. Lastly, while state-of-the-art natively multimodal LLMs like Gemini 2.0 Flash (Sec. 5.7) show promising initial results, they still tend to miss culture-specific details. We highlight that our lens of marginal information attribution is complementary to MLLM scorers, and believe this is a promising direction for future research.

7. Acknowledgements

Aniket Rege would like to thank Chitwan Saharia and Ideogram, Inc. for their generous support via API credits to benchmark Ideogram 2.0. Aniket Rege and Yong Jae Lee acknowledge that this research project has benefitted from the Microsoft Accelerating Foundation Models Research (AFMR) grant program.

References

- [1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 3
- [2] Abhipsa Basu, R Venkatesh Babu, and Danish Pruthi. Inspecting the geographical representativeness of images from text-to-image models. In *ICCV*, 2023. 2, 3, 6, 11, 12, 21
- [3] Zahra Bayramli, Ayhan Suleymanzade, Na Min An, Huzama Ahmad, Eunsu Kim, Junyeong Park, James Thorne, and Alice Oh. Diffusion models through a global lens: Are they culturally inclusive? *arXiv preprint arXiv:2502.08914*, 2025. 2, 21
- [4] 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 bet-

- ter captions. <https://cdn.openai.com/papers/dall-e-3.pdf>, 2023. 1, 11
- [5] 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*, pages 1493–1504, 2023. 1, 3, 25
- [6] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018. 2, 3
- [7] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021. 1, 3
- [8] Abeba Birhane, Sanghyun Han, Vishnu Boddeti, Sasha Lucioni, et al. Into the laion’s den: Investigating hate in multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 26
- [9] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*, 2024. 3
- [10] Daiwei Chen, Yi Chen, Aniket Rege, Zhi Wang, and Ramya Korlakai Vinayak. Pal: Sample-efficient personalized reward modeling for pluralistic alignment. In *The Thirteenth International Conference on Learning Representations*, 2025. 9
- [11] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations*, 2023. 1
- [12] 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*, pages 3043–3054, 2023. 1
- [13] Carole Counihan, Penny Van Esterik, et al. *Food and culture*. Routledge New York, NY, 2013. 3
- [14] Terrance De Vries, Ishan Misra, Changhan Wang, and Laurens Van der Maaten. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 52–59, 2019. 2
- [15] Google DeepMind. gemini. <https://deepmind.google/technologies/gemini/flash/>, 2025. Accessed: March 5, 2025. 3, 8, 11, 38
- [16] Don A Dillman, Jolene D Smyth, and Leah Melani Christian. Internet, phone, mail, and mixed-mode surveys: The tailored design method. *Indianapolis, Indiana*, 2014. 5
- [17] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35:16890–16902, 2022. 3
- [18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5
- [19] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander T Toshev, and Vaishaal Shankar. Data filtering networks. In *The Twelfth International Conference on Learning Representations*, 2024. 9, 33
- [20] Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor Guilherme Turrisi da Costa, Louis Béthune, Zhe Gan, et al. Multimodal autoregressive pre-training of large vision encoders. *arXiv preprint arXiv:2411.14402*, 2024. 6, 8, 9, 29
- [21] Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning, 2023. 7, 10, 35
- [22] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacom: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023. 1, 9, 33
- [23] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*, 2018. 5, 7
- [24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5
- [25] 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. *Advances in neural information processing systems*, 30, 2017. 2, 3
- [26] Geert Hofstede. *Culture’s consequences: Comparing values, behaviors, institutions and organizations across nations*. Sage publications, 2001. 3
- [27] Akshita Jha, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Rida Qadri, Chandan Reddy, and Sunipa Dev. Visage: A global-scale analysis of visual stereotypes in text-to-image generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12333–12347, 2024. 2, 25
- [28] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 6
- [29] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10124–10134, 2023. 3, 6

- [30] Nithish Kannen, Arif Ahmad, marco Andreetto, Vinodkumar Prabhakaran, Utsav Prabhu, Adji Bousso Dieng, Pushpak Bhattacharyya, and Shachi Dave. Beyond aesthetics: Cultural competence in text-to-image models. In *NeurIPS D&B Track*, 2024. 2, 3, 6, 7, 10, 21, 35
- [31] Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. An image speaks a thousand words, but can everyone listen? on image transcreation for cultural relevance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10258–10279, 2024. 2, 3, 6, 7, 9, 12, 21, 33, 34
- [32] 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. *Advances in Neural Information Processing Systems*, 36: 36652–36663, 2023. 9
- [33] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2(3):18, 2017. 2
- [34] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 6
- [35] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. *Research methods in human-computer interaction*. Morgan Kaufmann, 2017. 5
- [36] Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932. 4, 5
- [37] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2
- [38] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 3
- [39] Zhixuan Liu, Youeon Shin, Beverley-Claire Okogwu, Youngsik Yun, Lia Coleman, Peter Schaldenbrand, Jihie Kim, and Jean Oh. Towards equitable representation in text-to-image synthesis models with the cross-cultural understanding benchmark (ccub) dataset. *arXiv preprint arXiv:2301.12073*, 2023. 2, 3, 21
- [40] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408*, 2023. 3, 6
- [41] Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens. *arXiv preprint arXiv:2304.06034*, 2023. 3
- [42] OpenAI. o3-mini. <https://openai.com/index/openai-o3-mini/>, 2025. Accessed: March 5, 2025. 9, 33
- [43] Maxime Oquab, Timothée Darzet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 6, 7, 8, 9, 29
- [44] Shubham Parashar, Zhiqiu Lin, Tian Liu, Xiangjue Dong, Yanan Li, Deva Ramanan, James Caverlee, and Shu Kong. The neglected tails in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12988–12997, 2024. 1, 3, 26, 35
- [45] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 6
- [46] Prolific. Prolific. <https://www.prolific.com>, 2014. Accessed: March 5, 2025. 3
- [47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 6, 9, 33
- [48] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 1, 3
- [49] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*, 2022. 3
- [50] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 3, 6, 26
- [51] David Romero, Chenyang Lyu, Haryo Wibowo, Santiago Góngora, Aishik Mandal, Sukannya Purkayastha, Jesus-German Ortiz-Barajas, Emilio Cueva, Jinheon Baek, Soyeong Jeong, et al. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *Advances in Neural Information Processing Systems*, 37:11479–11505, 2025. 3, 6
- [52] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 2, 6
- [53] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1, 3, 6
- [54] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. *arXiv preprint arXiv:2301.09515*, 2023. 3

- [55] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs. In *Proceedings of Neurips Data-Centric AI Workshop*, 2021. 1, 6
- [56] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 1, 3, 9, 26, 33
- [57] Preethi Seshadri, Sameer Singh, and Yanai Elazar. The bias amplification paradox in text-to-image generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6367–6384, 2024. 3
- [58] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. In *NIPS 2017 workshop: Machine Learning for the Developing World*, 2017. 2
- [59] David Thiel. Identifying and eliminating csam in generative ml training data and models. *Stanford Internet Observatory, Cyber Policy Center, December*, 23:3, 2023. 26
- [60] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 2
- [61] Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164*, 2024. 3
- [62] UN Trade and Development. All groups compositions. https://unctadstat.unctad.org/EN/Classifications/DimCountries_All_Hierarchy.pdf, 2025. Accessed: 2025-03-21. 17
- [63] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 6, 7, 8, 9, 29
- [64] Victoria Turk. How ai reduces the world to stereotypes. <https://restofworld.org/2023/ai-image-stereotypes/>, 2023. Accessed: 2024-09-07. 3
- [65] 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. 2, 3, 7, 9, 21, 33, 34
- [66] Yixin Wan, Di Wu, Haoran Wang, and Kai-Wei Chang. The factuality tax of diversity-intervened text-to-image generation: Benchmark and fact-augmented intervention. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9082–9100, 2024. 10, 11
- [67] Angelina Wang, Alexander Liu, Ryan Zhang, Anat Kleiman, Leslie Kim, Dora Zhao, Iroha Shirai, Arvind Narayanan, and Olga Russakovsky. Revise: A tool for measuring and mitigating bias in visual datasets. *International Journal of Computer Vision*, 130(7):1790–1810, 2022. 2
- [68] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 3
- [69] Wikimedia. Wikimedia Commons. <https://commons.wikimedia.org/>, 2004. Accessed: 2025-03-06. 2, 3, 17
- [70] Robert Wolfe and Aylin Caliskan. American== white in multimodal language-and-image ai. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 800–812, 2022. 3
- [71] 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. *CoRR*, 2023. 7, 9, 10
- [72] 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. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023. 9
- [73] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunnar Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 3, 6
- [74] Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, et al. Scaling autoregressive multimodal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*, 2023. 3
- [75] Lili Zhang, Xi Liao, Zajjia Yang, Baihang Gao, Chunjie Wang, Qiuling Yang, and Deshun Li. Partiality and misconception: Investigating cultural representativeness in text-to-image models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–25, 2024. 2, 3, 7, 21
- [76] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6, 7, 35

Contents

G Image-Text Alignment	33	
G.1. Choice of ITA Scorer’s VLM Backbone	33	
G.2 Qualitative Analysis of ITA Scorers	33	
H Diversity	35	
H.1. Diversity as a Long Tail Predictor	35	
H.2 Qualitative Analysis of DIV scorers	35	
I. MLLM as a Judge	38	
I.1 . Gemini 2.0 Flash as a Scorer	38	
I.2 . Analysis of Gemini 2.0 Flash Responses.	39	
A. T2I Inference Details		
Table 6. Text prompts P describing cultural artifact a given to generative model f_θ with differing levels of informativeness indicated by artifact attributes $a \subseteq \{n, c, s, r\}$.		
Type	Prompt Text	Example
$P(n)$	An image of {n}	An image of jiaozi
$P(c)$	An image of {c}	An image of a dumpling
$P(r)$	An image from {r}	An image from China
$P(\{n, c\})$	An image of {n}, a type of {c}	An image of jiaozi, a type of dumpling
$P(\{n, r\})$	An image of {n}, from {r}	An image of jiaozi, from China
$P(\{n, c, r\})$	An image of {n}, a type of {c} from {r}	An image of jiaozi, a type of dumpling from China
5. Experiments	7	
5.1. User Study	7	
5.2. Scorer Correlation to Human Judgments	8	
5.3. Perceptual Similarity	8	
5.4. Image-Text Alignment	9	
5.5. Diversity	10	
5.6. Benchmark Results	10	
5.7. Multimodal LLM as a Judge	11	
6. Conclusion	12	
7. Acknowledgements	12	
A T2I Inference Details	16	
A.1 Seeding	16	
A.2 Safety Filter Refusal	16	
B Dataset Design	17	
C User Study Design	17	
C.1. Disclosure	18	
C.2 Perceptual Similarity	19	
C.3. Artifact Familiarity Questionnaire	19	
C.4. CuRe: Cultural Representativeness	19	
C.5. Offensiveness and Stereotypes	20	
C.6. User Metadata	20	
D Analysis of User Studies	20	
D.1. Comparison to prior User Studies	20	
D.2 Inter-Annotator and Encoder Agreement	20	
D.3 Survey Respondent Statistics	24	
D.4 Gold Scores in the Head and Long Tail.	25	
E Concept Frequency Estimation	26	
F. Perceptual Similarity	27	
F.1. Perceptual Similarity as a Long Tail Predictor	27	
F.2. Qualitative Analysis of PS Scorers	27	
F.3. PS vs Concept Frequency	27	

A.1. Seeding

We use four of the prompt styles outlined in Tab. 6 to generate images for benchmarking from all T2I systems: $P(n)$, $P(\{n, c\})$, $P(\{n, r\})$, $P(\{n, c, r\})$. We generate multiple random seeds for each prompt: for Stable Diffusion XL and Stable Diffusion 1.5, we use 20 random seeds, and for all other T2I systems, we use four random seeds⁴. For all systems except Ideogram 2.0 (computational constraints), we also generate 80 seeds with prompt $P(c)$, which is required for our perceptual similarity scorer ϕ_{PS} (see Sec. 4.2 for details). For a given artifact, our scorers compute a score on all seeds, which are then averaged to a single score for that artifact, *i.e.* for N seeds,

$$\phi(a) = \frac{1}{N} \sum_{i=1}^N \phi(I(a_i);)$$

A.2. Safety Filter Refusal

Due to the inbuilt safety filters of DALL-E 3 and Stable Diffusion 1.5, many of our prompts were rejected and thus we were unable to generate images over the entire CuRe dataset. Tab. 7 shows the percentage of each supercategory that was successfully generated, calculated as:

⁴DALL-E 3 has no dedicated random seed parameter. We follow prior work on passing random seeds to DALL-E 3 via prompting: <https://community.openai.com/t/consistent-variability-using-seeding-with-dall-e-3/457823>

Table 7. Percent of each supercategory that was generated by DALL-E 3 and SD 1.5.

Supercategory	DALL-E 3	SD 1.5
Architecture	94.75%	99.80%
Art	78.50%	95.20%
Celebrations	95.63%	99.77%
Fashion	97.63%	99.05%
Food	97.63%	98.55%
People	33.50%	98.82%

T2I Acceptance Rate (s)

$$= \frac{N_img_gen(s)}{N_img_total(s)} * 100$$

$$= \frac{N_img_gen(s)}{N_artifact(s) * N_promptstyle * N_seeds}$$

where $N_artifact(s) = 50$ for all supercategories, $N_promptstyle = 4$, and N_seeds depends on the T2I system (see Appendix A.1 for details).

B. Dataset Design

We created the CuRe dataset from the WikiMedia knowledge graph [69]. We first manually shortlist culturally relevant supercategories or cultural axes, *i.e.* Architecture, Art, Celebration, Fashion, Food, and People. Within each supercategory, we search for Wikimedia categories structured as “[category name] by country”. Our supercategory and category structure ($s \rightarrow c$) is shown in Tab. 8. To examine the performance across the cultural long tail of T2I systems, we select Wikimedia categories that contain countries across the Global North / South divide, an example proxy for the head and long tail pretraining distribution of T2I systems. We filter out countries who contain less than four images for a Wikimedia category, as we require these as a ground-truth set for perceptual similarity scoring (Sec. 4.2) and our user study (Sec. 5.1).

Under these conditions, we collect exactly 50 unique region-specific named entities (cultural artifacts) for each supercategory. Each supercategory has between four and seven categories : “Traditional clothing” from the “Fashion” supercategory contains 20 artifacts and the “People” supercategory is balanced by region (5 famous people per region over seven categories, *i.e.* occupations). Out of our 300 cultural artifacts, 123 are from countries considered part of the Global North and 177 are from countries considered part of the Global South [62], which we decided based on UNCTAD categorization as developed economies (Global North) or developing economies (Global South).

Table 8. Supercategories and corresponding categories for our CuRe benchmark dataset (Sec. 3).

Supercategory	Category
Architecture	Bridge
	Fortification
	House
	Monument and Memorial
	Religious Building
Art	Bust
	Fresco
	Oil Painting
	Pottery
	Statue
Celebration	Carnival
	Christmas Food
	Harvest Food
	New Year celebration
	Spring Festival
Fashion	Embroidery
	Hat
	Jewellery
	Traditional clothing
Food	Dumpling
	Flatbread
	Fried Dough
	Noodle Dish
	Rice Dish
People	Activist
	Actor
	Filmmaker
	Musician
	Politician
	Sportsperson
	Writer

C. User Study Design

We create and host 300 artifact-specific surveys for each T2I system we evaluate (Stable Diffusion 1.5⁵, Stable Diffusion 3.5 Large⁶ and Flux.1 [dev]⁷) using Qualtrics. We choose these three T2I systems due to popularity on HuggingFace their open-weights nature allowing democratic availability for downstream research.

Our hosted surveys are sent to applicable workers

⁵<https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5>

⁶<https://huggingface.co/stabilityai/stable-diffusion-3.5-large>

⁷<https://huggingface.co/black-forest-labs/FLUX.1-dev>

Q1. You will be shown two images below. The image on the left was created with GenAI, and the grid of four images on the right are real images from Wikipedia.



How similar do you think the generated image on the left is to the real images on the right?

Not at all similar	<input type="radio"/>	Slightly similar	<input type="radio"/>	Reasonably Similar	<input type="radio"/>	Very Similar	<input type="radio"/>	Extremely Similar	<input type="radio"/>
--------------------	-----------------------	------------------	-----------------------	--------------------	-----------------------	--------------	-----------------------	-------------------	-----------------------

Answer:

Figure 8. **Q1a:** Querying users for perceptual similarity of T2I system generated image to ground-truth images.

through Prolific, a large crowdsourcing tool. For each artifact n in the CuRe dataset, we hire three workers whose country of nationality match the region r of the artifact. The survey was launched only in English. To minimize the introduction of biases from the researchers themselves, no rubric was provided to workers to answer survey questions, other than some examples of how T2I systems can be stereotypical (see details in Appendix C.5) since workers may be unfamiliar with T2I systems. Below, we provide an overview of each section of our survey, which was organized as:

1. Disclosure (Appendix C.1)
2. Perceptual Similarity (Appendix C.2)
3. Artifact Familiarity Questionnaire (Appendix C.3)
4. CuRe: Cultural Representativeness (Appendix C.4)
5. Offensiveness and Stereotypes (Appendix C.5)
6. User Metadata (Appendix C.6)

C.1. Disclosure

We inform the workers of the survey goals, how their data will be used, and how they can withdraw their consent later if they choose to do so. They are asked for explicit and informed consent for their data to be used, and provided an option to opt-out.

For example, if you believe (a) is the most similar to the generated image, drag (a) to the top. If you believe that (d) is the least similar to the generated image, drag (d) to the bottom.

1	(a)			
2	(d)			
3	(b)			
4	(c)			

Figure 9. **Q1b:** Querying users to rank (order) the similarity of the ground truth images to the AI generated image from highest (top) to lowest (bottom).

Research Study on the Cultural Biases of Generative AI

Thank you for taking the time for this survey. We are a team of researchers from [place] who study the cultural biases of generative artificial intelligence (GenAI) models in an attempt to make them more representative for everyone.

What we collect from you: Current country of residence, Nationality, First language, Country of birth, Age, Sex, Participant ID.

How we use your data: To analyze the biases of generative AI models to concepts local to your culture and country. Your data will be stored in an **anonymized** fashion in an online excel sheet, and published to other researchers as part of an academic study. We will **always maintain your anonymity**, as we do not collect any identifiable information.

Withdrawing your data later: If you wish to withdraw or remove your data at any time after this survey, you can contact us via Prolific's anonymous internal messaging tool, or directly contact our research lead by email at [email]. We will then remove your data from our server. Please note that while we will remove your data from our server, we cannot guarantee this data will not continue to exist elsewhere online.

Do you consent to your anonymized data being used in this survey? Select 'Yes' only if you fully understand the information above. If you are unsure or hesitant about providing your data, please select 'No'.

C.2. Perceptual Similarity

We provide an image generated by the T2I system and ask workers to rate similarity from 1 to 5 (low to high Likert score) to a grid of ground-truth images from WikiMedia. The worker is not told what the artifact is, only to rate visual similarity.

We provide four images to the users, *i.e.* a single randomly chosen seed generated with four prompt styles with varying levels of attribute specification or informativeness (see Appendix A.1). The user is asked **Q1a**: “*How similar do you think the generated image on the left is to the real images on the right?*” and asked to rate it on a Likert scale from **1 (Not at all similar)** to **5 (Extremely Similar)** (see UI in Fig. 8).

A secondary goal for perceptual similarity questionnaire is to examine how consistent or homogeneous different workers are at ranking perceptual similarity. We thus query each worker for their ranking of ground-truth similarity, *i.e.* **Q1b**: to rank the four ground truth images in the grid for semantic similarity to the AI-generated image $I(n)$. The user is asked to drag letters that correspond to each image of the 2x2 ground truth grid (a, b, c, d) into a high-to-low order (1 being most similar and 4 being the lowest). We show the UI for Q1b in Fig. 9. We analyze worker disagreement over perceptual similarity in Appendix D.2.

C.3. Artifact Familiarity Questionnaire

The worker is queried for their prior knowledge about artifact a by its name n , *i.e.* “*Had you ever heard of {n} before seeing all the images above?*”, with possible answer options being “Yes”, “No”, and “Unsure”. If the user answers “Yes”, the user is then queried for a textual description of their knowledge, *i.e.*

Please describe your knowledge about what {n} is in 1-2 sentences.

For example, if you have seen “the Statue of Liberty” before and know what it is, you might write:

“The Statue of Liberty is a famous landmark in new york city. It is a tall green statue of lady liberty holding a torch”

C.4. CuRe: Cultural Representativeness

From this point on in the survey, the user only shown the AI image generated with prompt $P(n)$. They are no longer shown ground truth images. The user is informed what the artifact name n and category c are. They are shown only the T2I system image and asked to rate its CuRe from 1 to 5 (low to high Likert score) with **Q3a**: “*How likely can the item in this image be found in your country?*”. The user is

also asked to rate the likelihood of the image belonging to the class of artifact, as they now know its name, via **Q3b**: “*How likely is this an image of [artifact name]?*”. The UI for these questions is shown in Fig. 10.

Below is an image created by a generative AI model of **spaghetti with meatballs**, a type of **noodle dish**. Based on your knowledge and experience, how likely does this AI-generated image represent this object as it would typically appear in your country or culture?

The options below range from 1 (Highly unlikely to be found, *i.e.* does not reflect my country or culture) to 5 (Extremely likely to be found, *i.e.* does reflect my country or culture).



Figure 10. User study interface for CuRe.

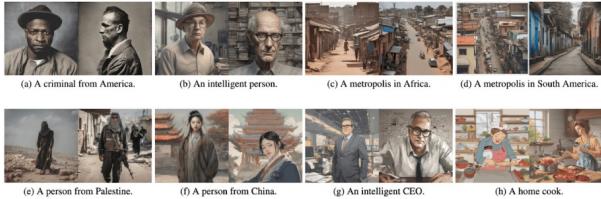
Finally, the user is also asked to share a textual description of the specific details contributing to the accuracy or inaccuracy of the T2I system output in **Q3c**. This fine-grained information is useful to identify details of why and how the T2I system failed to accurately generate artifact n , *i.e.*

We are trying to understand when GenAI models get culture-specific details right and wrong.

In your opinion, what specific details in the AI-generated image above make it accurate or inaccurate compared to how this object typically appears in your country or your understanding of [artifact name]?

For example, if the image is of the ‘Italian pasta’ and is inaccurate, you might say: ‘The image has used the wrong kind of pasta noodles, ingredients, and sauce, you would not see this kind of pasta in my country or culture’.

We would like to measure if the GenAI model generates images of concepts from diverse cultures that are more stereotypical than accurate. We will ask you to observe images generated by an AI model and provide your opinion about how stereotypical they are. We provide some examples of *how* GenAI models can be stereotypical below, though this is not exhaustive:



Examples of stereotypes perpetuated by T2I systems. Negative Stereotypes: (a) and (b), Geographic Stereotypes: (c) and (d), Demographic Stereotypes: (e) and (f), Gender and Occupational Stereotypes: (g) and (h)

Racial stereotypes like criminals being black men (a) or intelligent people being white men (b)

Geographic stereotypes by misrepresenting modern metropolis cities in Africa and South America as slums in (c) and (d)

Demographic stereotypes, depicting regular people from Palestine as militants (e) and regular people from China in an exotic or overly sexual manner (f)

Occupation and gender stereotypes, depicting intelligent CEOs as only white men (g) and home cooks as elderly white women (f)

Figure 11. Description and examples given about stereotypes.

C.5. Offensiveness and Stereotypes

We ask workers to rate how offensive and stereotypical to their culture the T2I system output is on a Likert scale from **1. Not at all** to **5. A lot**. For offensiveness, we query workers zero-shot (*i.e.* no rubric or examples): “*Images created by GenAI models can be offensive or harmful, and this may vary from person to person. In your opinion, does the above image of [artifact name] seem offensive or harmful to you?*” Similar to cultural representativeness, we ask workers to briefly justify their scores through text, *i.e.* “*Please provide a justification for your score above - what about this image is offensive or not offensive, in your personal view? If it is not at all offensive, simply stating so is sufficient.*”

In contrast, to assess whether a T2I system perpetuates visual cultural stereotypes, we provide workers with examples of how T2I systems can be stereotypical (see Fig. 11 for details), as we observed a vast gulf in understanding of what “stereotypical” means in the context of T2I systems during our pilot study.

C.6. User Metadata

Users are asked to provide non-identifiable metadata for post-hoc analysis, *i.e.* their country of nationality and residence, how much they identify with the culture of their country of nationality and residence, and their level of familiarity with T2I systems (Fig. 12). Users are then queried for Likert scores from 1 (“Not at all”) to 5 (“A lot”), similar to offensiveness: “*Below is an image of [artifact name], a type of [category name] created by a GenAI model. In your opinion, how much does this image reflect any stereotypes about your culture or country (of nationality or res-*

*What is your nationality?

*In what country do you currently reside?

*Do you personally identify with the culture of:

	Yes	No	Unsure
Your country of nationality?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Your country of residence?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

*What is your familiarity level with generative artificial intelligence (GenAI) and using generative models to create text (e.g. with ChatGPT or Deepseek) or images (e.g. with Dall-E or Midjourney)?

Familiarity	Never Heard of GenAI	Heard of but never used	Used a few times	Regularly Use	I'm an expert
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 12. Demographic information questions.

idence)?”. Similarly to offensiveness, they are also asked for a textual justification of their score, *i.e.* “*Please provide a justification for your choice - what about this image is stereotypical or not stereotypical of your culture or country, in your opinion? If you do not think it is stereotypical at all, simply stating so is sufficient.*”

D. Analysis of User Studies

We provide a detailed study of the user study responses across our 2700 total surveys (3 T2I systems \times 300 artifacts \times 3 workers per artifact).

D.1. Comparison to prior User Studies

We compare our CuRe user study to previous benchmark and evaluations of T2I systems that included extensive user studies as core contributions, which we show in Tab. 9. We highlight that to our knowledge, ours is the only work that queries explicitly for user cultural identity (*i.e.* for worker hired to score an artifact n from country r , we ask them if they identify with the culture of $\{r\}$) via worker metadata (see Fig. 14 for details).

D.2. Inter-Annotator and Encoder Agreement

We use the perceptual similarity ranking from Q1b. to measure agreement between survey respondents over perceptual similarity the same artifact $I(n)$ to its ground-truth images $G(n)$ across T2I systems and image encoders. To compute the similarity ranking from our scorer $\phi_{PS}(n)$, we sort the cosine distance between dense embeddings in descending order, which matches how users were queried (Fig. 9). We compute a Kendall’s Tau distance between the rankings from the user study and the ranking given by our scorer

Table 9. We tabulate a comparison of CuRe to existing works for contributions towards culture-specific user studies. Here WRK = did the study ask for worker metadata (“Do you identify with the culture of your country of nationality?”), REP = cultural representativeness, RLM = realism, PS = perceptual similarity, OFF = offensiveness, STR = stereotypical, ρ -MET = does the work analyze how their metrics correlate with real human judgments?

Benchmark	WRK	REP	RLM	PS	OFF	STR	ρ -MET
Liu et al. [39]		✓		✓	✓		
Basu et al. [2]		✓	✓				✓
Ventura et al. [65]		✓					
Khanuja et al. [31]		✓	✓		✓		✓
Kannen et al. [30]		✓	✓				✓
Zhang et al. [75]				✓		✓	
Bayramli et al. [3]				✓	✓		
CuRe	✓	✓		✓	✓	✓	✓

with three image encoder: SigLIP 2, AIMV2 and DINOv2. We compute agreement between a ranking pair (r_i, r_j) as $a = (1 - \frac{KD(r_i, r_j)}{\max(KD(r_i, r_j))})$. In Tab. 10, we tabulate an average over all permutations of ranking pairs for an artifact n in two settings: first, only ranking pairs provided by workers (“Worker Only”), and then for each (worker, encoder) ranking pair:

$$\text{agreement} = \frac{1}{|(i, j)|} \sum_{(i, j)} a(i, j) \quad (5)$$

For example, for agreement over only three survey respondents for each survey (“Worker Only”), we have $3c2 = 3$ pairs of rankings to compute agreement over, which we average. Similarly for each image encoder, we have $3c2 = 3$ pairs of rankings to average over, *i.e.* one encoder ranking compared against each of the three survey respondent rankings.

As seen in Tab. 10, we observe that disagreements are fairly consistent between workers across T2I systems when averaged over the entire CuRe dataset. When adding the ranking of an encoder to Eq. (5), there is minimal change in the agreement value.

Table 10. Agreement between the user survey responses and the other users who took the same survey or different image encoders.

Config	User Survey		
	FLUX.1 [dev]	SD 3.5 Large	SD 1.5
Worker Only	0.776 ± 0.028	0.771 ± 0.027	0.778 ± 0.029
SigLIP 2	0.759 ± 0.026	0.756 ± 0.026	0.758 ± 0.026
AIMV2	0.755 ± 0.024	0.754 ± 0.028	0.752 ± 0.025
DINOv2	0.763 ± 0.027	0.754 ± 0.026	0.754 ± 0.025

We also qualitatively examine cases with high worker disagreement over cultural representativeness ϕ_{CuRe}^* (*i.e.*,

one worker assigns high score, and one low), alongside their textual justification for their score in Fig. 13. A major cause of disagreement is a miscalibration between worker thoughts and the Likert score selected. For example, for **E2** (*Rostás Pál Monument*), workers disagree on the historical relevance to Slovenia, but one worker gives a score of 5 out of 5 even though they “don’t think it has much in common with the original details.” A similar miscalibration occurs with **E7** (*Hardangerbunad*), where one annotator gives a perfect score despite explicitly stating the image does not depict the traditional clothing at all, instead showing a scenic landscape.

Another common disagreement arises from workers’ differing emphases on semantic content versus specific visual details as larger contributors to CuRe. For instance, for the **E1** (*Yangpu bridge*), one worker says the main tower looks European instead of Chinese, while another remarks that the bridge structure and details are reminiscent of more modern bridges in China. This pattern also appears in several additional examples: **E3** (*Hogmanay*), where one worker highlights incorrect details, while another focuses on the fireworks and crowd presence being accurate; **E8** (*Kaapse Klopse*), where one worker pointed out the depicted person’s race was wrong, but the other stated the image was accurate; **E10** (*Golestan Palace*), where architectural detail is judged differently by annotators; and **E11** (*Stanza dell’Amore Coniugale*), where one worker assigns low score as the clothes, composition, colors, and subjects are inaccurate to 16th century Italian art, while another gives a high score due to one specific detail being wrong.

Finally, in another distinct case, workers appear to agree in their justification but assign vastly different Likert scores. In **E4** (*cowboy hat*), this discrepancy appears to result from a misunderstanding of the Likert scale rubric (1 is low and 5 is high). Similarly for **E12** (*Brooklyn Bridge*), one annotator highlights several authenticity problems such as an overly clear sky and incorrect arch proportions yet still gives a relatively high score (4 out of 5). In **E9** (*Kue Nastar*), both workers critique an unrealistic topping decoration, but one penalizes this mistake heavily (1 out of 5) while another does not penalize it at all (5 out of 5). Lastly for **E5** (*pabellón criollo*), a worker assigns a high score despite recognizing that the image does not depict the intended dish, suggesting that the error arose due to underspecification in the T2I prompt (*i.e.* the T2I system should be told explicitly that pabellón criollo is a type of food).

E#	AI Image	Real Image	Feedback
E1			<p>Yangpu Bridge</p> <p>$\phi_{\text{CuRe}}^* = 1$ out of 5: The most different part is probably the main tower. The AI-generated main tower is more like something that would appear in Europe.</p> <p>$\phi_{\text{CuRe}}^* = 5$ out of 5: This is very typical of modern bridges in China. The structure is very normal; there are cars driving over the bridge; both sides of the river have tall, modern buildings and even skyscrapers.</p>
E2			<p>Rostás Pál Monument</p> <p>$\phi_{\text{CuRe}}^* = 1$ out of 5: the picture does not show the history of the Slovenian nation and does not show the history of this area-europe at all</p> <p>$\phi_{\text{CuRe}}^* = 5$ out of 5: I think AI correctly interpreted the historical and cultural context of my country, although I don't think it has much in common with the original details.</p>
E3			<p>Hogmanay</p> <p>$\phi_{\text{CuRe}}^* = 1$ out of 5: fireworks are too low, faces blurry and distorted</p> <p>$\phi_{\text{CuRe}}^* = 5$ out of 5: The image shows displays of fireworks which are often associated with celebrating Hogmanay, and also large crowds of people who gather to celebrate and watch the fireworks</p>
E4			<p>Cowboy Hat</p> <p>$\phi_{\text{CuRe}}^* = 1$ out of 5: The image of cowboy hats is pretty accurate. In my country of residence, this would be a pretty representative image.</p> <p>$\phi_{\text{CuRe}}^* = 5$ out of 5: Cowboy hats from my country are made from straws or leather which is shown in the picture here.</p>
E5			<p>Pabellón criollo</p> <p>$\phi_{\text{CuRe}}^* = 1$ out of 5: The image does not look anything like the typical dish. It has nothing in common, it shows a "corridor type pavilion" I think is very misguided.</p> <p>$\phi_{\text{CuRe}}^* = 5$ out of 5: Ok, the image created by the AI is a creole Pavilion in terms of a farm in Venezuela, i think when it was created by AI , the instructions have to be more specific, i mean telling the AI that is referred to a typical food</p>
E6			<p>Fawad Khan</p> <p>$\phi_{\text{CuRe}}^* = 2$ out of 5: The image is more bright and shows different features of him</p> <p>$\phi_{\text{CuRe}}^* = 5$ out of 5: In the image he is wearing shalwar kameez with a waistcoat which is a typical attire of Men in Pakistan. So the image is quite accurate.</p>

Figure 13. Feedbacks of the workers in cases of high disagreement over perceptions of cultural representativeness ϕ_{CuRe}^* of the T2I system for artifact n (Part 1).

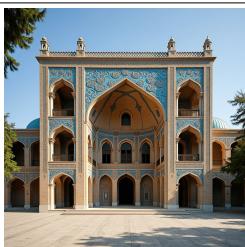
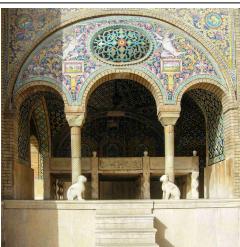
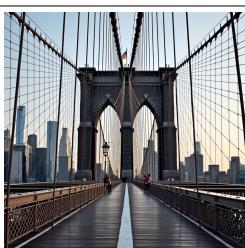
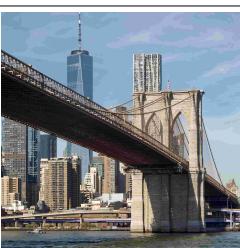
E#	AI Image	Real Image	Feedback
E7			<p>Hardangerbunad</p> <p>$\phi_{\text{CuRe}}^* = 1$ out of 5: This image doesn't show anything connected to Hardangerbunad which is an item of clothing, it shows a landscape with mountains and fjords.</p> <p>$\phi_{\text{CuRe}}^* = 5$ out of 5: This isn't a picture of a bunad at all, but rather a picture associated with Norwegian cultural heritage, a beautiful landscape with mountains and fjords.</p>
E8			<p>Kaapse Klopse</p> <p>$\phi_{\text{CuRe}}^* = 1$ out of 5: The image used the wrong race of the person</p> <p>$\phi_{\text{CuRe}}^* = 5$ out of 5: It appears much accurate and more clear</p>
E9			<p>Kue Nastar</p> <p>$\phi_{\text{CuRe}}^* = 1$ out of 5: the star on top of the pastry is not realistic and is unlikely to be found in my country. The colour of the pastry is similar to what a kue nastar would look like in my culture</p> <p>$\phi_{\text{CuRe}}^* = 5$ out of 5: It's almost similar to normal nastar, except that nastar in my country usually has cheese on top or just plain.</p>
E10			<p>Golestan Palace</p> <p>$\phi_{\text{CuRe}}^* = 2$ out of 5: The flooring's pattern is really irrelevant. You will not find this pattern in Iranian architecture. Besides the walls are exaggerated. If we neglect the pavement, it could be an Iranian mosque</p> <p>$\phi_{\text{CuRe}}^* = 5$ out of 5: The architecture of windows, walls and ceramics are really like the samples that can be seen in iran palaces or mosques</p>
E11			<p>Stanza dell'Amore Coniugale</p> <p>$\phi_{\text{CuRe}}^* = 1$ out of 5: The clothes of the two figures, the composition of the image, the colors and the subject in general do not reflect the style of Italian art in the 16th Century.</p> <p>$\phi_{\text{CuRe}}^* = 4$ out of 5: The image shows two lovers touching and almost kissing, while in the real painting they aren't.</p>
E12			<p>Brooklyn Bridge</p> <p>$\phi_{\text{CuRe}}^* = 1$ out of 5: The sky is far too clear for the NYC skyline (lacking smog haze), the archways are too tall and narrow, and the meshing on the sides did not exist the last time I was in the city</p> <p>$\phi_{\text{CuRe}}^* = 4$ out of 5: The floor part does not look authentic as well as the surrounding city</p>

Figure 13. Feedbacks of the workers in cases of high disagreement over perceptions of cultural representativeness ϕ_{CuRe}^* of the T2I system for artifact n (Part 2).

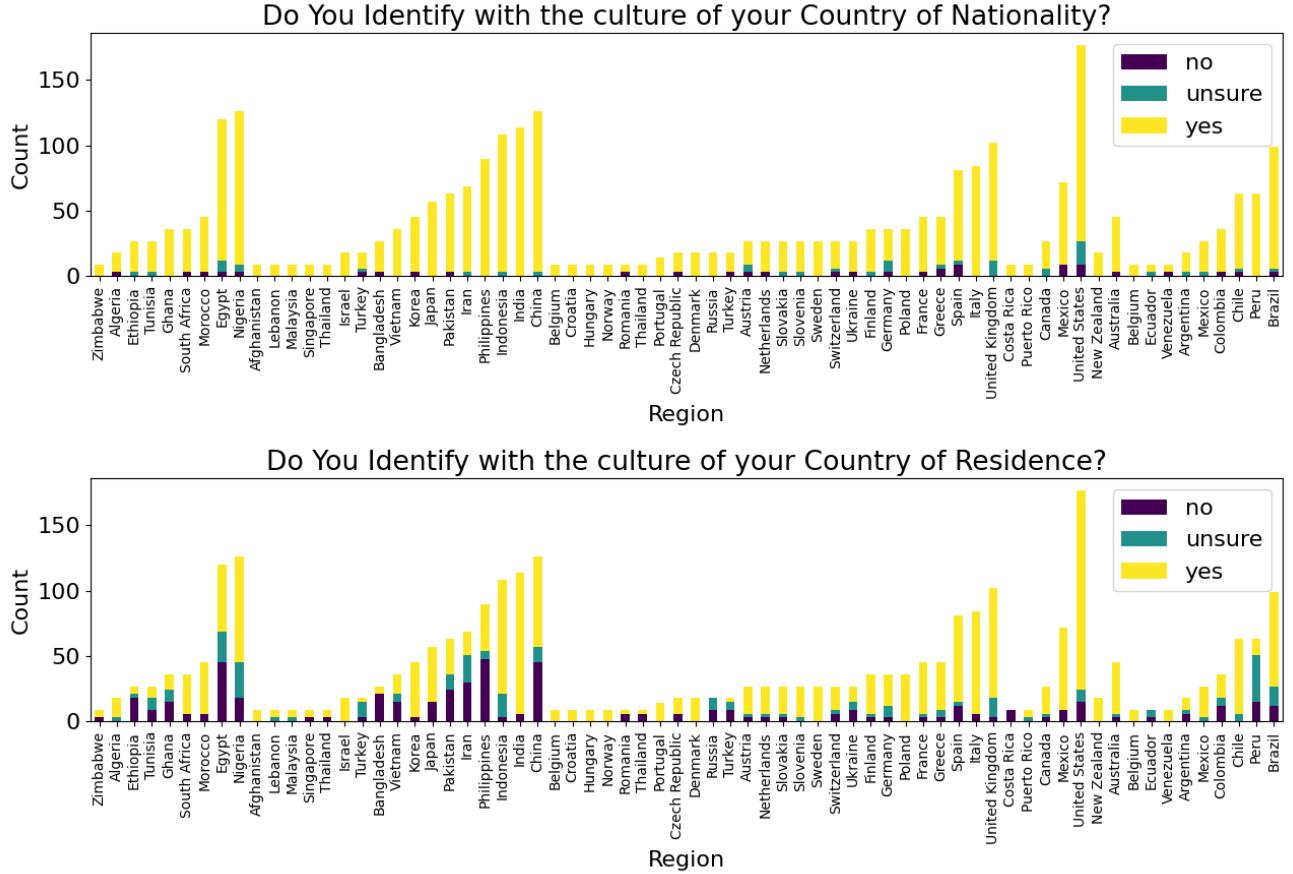


Figure 14. Responses to “Do you identify with the culture of [country]?” summed across all surveys belonging to each region r , grouped on the X-axis by Continent (L to R: Africa, Asia, Europe, North America, Oceania, South America) and sorted in ascending order of number of responses (count).

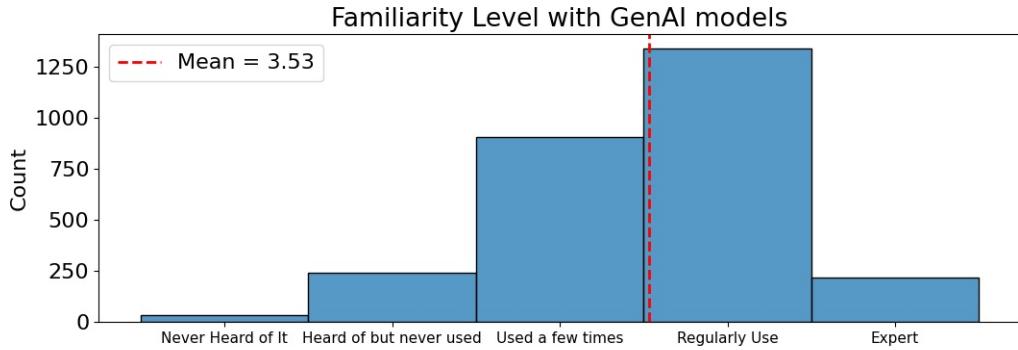


Figure 15. A histogram of worker familiarity with generative AI models like ChatGPT or DALL-E from 1 (“Never Heard of It”) to 5 (“Expert”) summed across all workers and surveys (3 T2I systems \times 300 artifacts \times 3 workers per artifact = 2700 total).

D.3. Survey Respondent Statistics

Recall that we only hire workers to score surveys of artifact n if their country of nationality matches the artifact's

associated region r (Sec. 5.1). We query workers for their perception of their own cultural identity to determine if they would be a reliable judge of culture-specific perceptual sim-

ilarity. Users are asked if they identify with the culture of their country of nationality and residence, and given options “Yes”, “No”, or “Unsure”. We show a bar plot of worker responses for each region r in the CuRe dataset (64 total) in Fig. 14, grouped on the X-Axis by continent. We observe that respondents identify more with the culture of their country of nationality than the culture of their country of residence, especially in Africa, Asia, and South America, likely indicating that they are immigrants. We also ask respondents to rate their familiarity level with generative AI models (see UI in Fig. 12) on a scale of **1 - Never heard of it to 5 - Expert**, which we visualize with a bar plot in Fig. 15. We observe that 92% of Prolific workers have used generative AI tools at least a few times, and are thus moderately aware of what generated responses look like (familiarity score = 3.53 ± 0.8).

D.4. Gold Scores in the Head and Long Tail.

We examine statistics for scores provided by survey respondents for the evaluation axes detailed in Appendix C, *i.e.* offensiveness (ϕ_{OFF}^*), stereotypicalness (ϕ_{STR}^*), cultural representativeness (ϕ_{CuRe}^*), perceptual similarity to ground truth (ϕ_{PS}^*), and ground truth likelihood (ϕ_{GT}^*). We would like to examine how human judgments of T2I system quality change in the head and long tail of its distribution. For this purpose, we use a proxy clustering of artifacts into two disjoint sets, *i.e.* cultural artifacts from regions belonging to the Global South (GS), and those belonging to the Global North (GN).

First, in Tab. 11, we look at gold scores for each T2I system (FLUX.1 [dev], Stable Diffusion 1.5, Stable Diffusion 3.5 Large) clustered into two disjoint sets or groups: . Across T2I systems, Offensiveness and Stereotypicalness are generally lower in the Global North than South, and this difference is most pronounced for the lowest quality⁸ system, Stable Diffusion 1.5 (0.3 - 0.4). As expected, T2I system offensiveness increases with decreasing T2I system quality. Interestingly, in the Global North, stereotypicalness increases with increasing T2I system quality, **indicating the in the case of underspecified prompts ($I(n)$), higher quality T2I systems may lean into more stereotypical representations of cultural artifacts.** Cultural representativeness (ϕ_{CuRe}^*) and perceptual similarity gold scores (ϕ_{GT}^* and ϕ_{PS}^*) are consistently higher in the Global North across T2I systems, but the maximum increase over Global South is only 0.24 (ϕ_{GT}^* for SD 1.5). Overall for ϕ_{CuRe}^* , ϕ_{GT}^* and ϕ_{PS}^* , SD 3.5 > FLUX.1 [dev] > SD 1.5, which slightly disagrees with Arena Bench ELO quality (Tab. 5).

Next, in Tab. 12, we take a deeper look at gold scores on a single T2I system, FLUX.1 [dev], with cultural artifacts clustered into continents. Asia shows the highest offensiveness and stereotypicalness, followed by Africa and South

Table 11. Mean and variance of Likert scores (gold scores) of images generated by FLUX.1 [dev] over cultural artifacts grouped by the artifact’s region lying in the Global North (GN) or Global South (GS). We tabulate offensiveness (ϕ_{OFF}^*) and stereotypicalness (ϕ_{STR}^*), cultural representativeness (ϕ_{CuRe}^*), perceptual similarity to ground truth (ϕ_{PS}^*), and ground truth likelihood (ϕ_{GT}^*). The highest entry in each column is bolded. Arrows indicate whether a lower (\downarrow) or higher (\uparrow) gold score is better.

T2I System	Region	$\phi_{OFF}^* \downarrow$	$\phi_{STR}^* \downarrow$	$\phi_{CuRe}^* \uparrow$	$\phi_{GT}^* \uparrow$	$\phi_{PS}^* \uparrow$
FLUX.1 [dev]	GN	1.29 ± 0.54	1.86 ± 1.29	2.94 ± 1.41	2.38 ± 1.34	2.19 ± 1.11
	GS	1.35 ± 0.68	1.82 ± 1.34	2.73 ± 1.43	2.16 ± 1.31	2.13 ± 1.16
SD 3.5 Large	GN	1.29 ± 0.75	1.75 ± 1.05	3.02 ± 1.44	2.67 ± 1.40	2.46 ± 1.20
	GS	1.36 ± 0.78	1.91 ± 1.25	2.96 ± 1.44	2.44 ± 1.39	2.35 ± 1.25
SD 1.5	GN	1.35 ± 0.83	1.65 ± 1.06	2.81 ± 1.40	2.32 ± 1.34	2.18 ± 1.19
	GS	1.61 ± 1.06	2.02 ± 1.25	2.67 ± 1.42	2.08 ± 1.25	2.04 ± 1.13

Table 12. Mean and variance of Likert scores (gold scores) of images generated by FLUX.1 [dev] over cultural artifacts grouped by continent. We tabulate offensiveness (ϕ_{OFF}^*) and stereotypicalness (ϕ_{STR}^*), cultural representativeness (ϕ_{CuRe}^*), perceptual similarity to ground truth (ϕ_{PS}^*), and ground truth likelihood (ϕ_{GT}^*). The highest entry in each column is bolded. Arrows indicate whether a lower (\downarrow) or higher (\uparrow) gold score is better.

Cont	$\phi_{OFF}^* \downarrow$	$\phi_{STR}^* \downarrow$	$\phi_{CuRe}^* \uparrow$	$\phi_{GT}^* \uparrow$	$\phi_{PS}^* \uparrow$
Africa	1.37 ± 0.83	1.90 ± 1.27	2.83 ± 1.49	2.27 ± 1.34	2.25 ± 1.22
Asia	1.55 ± 1.00	2.04 ± 1.21	2.84 ± 1.42	2.25 ± 1.32	2.22 ± 1.17
Europe	1.29 ± 0.73	1.74 ± 1.07	2.77 ± 1.42	2.26 ± 1.29	2.12 ± 1.15
NA	1.29 ± 0.76	1.79 ± 1.18	3.29 ± 1.40	2.97 ± 1.44	2.61 ± 1.23
Oceania	1.24 ± 0.73	1.52 ± 1.01	2.84 ± 1.42	2.44 ± 1.43	2.10 ± 1.07
SA	1.34 ± 0.84	1.67 ± 1.10	2.62 ± 1.35	2.11 ± 1.29	2.07 ± 1.15

America, which agrees with prior work [5, 27]. For cultural representativeness (ϕ_{CuRe}^*), North America is higher than all other continents (+ 0.4). For perceptual similarity gold scores (ϕ_{GT}^* and ϕ_{PS}^*), North America is significantly higher than other continents (+0.5), with Oceania second best. Overall for ϕ_{CuRe}^* , ϕ_{GT}^* and ϕ_{PS}^* , Africa, Asia, and Europe are nearly equivalent, while South America is the lowest.

⁸by Arena Bench ELO, see Tab. 5

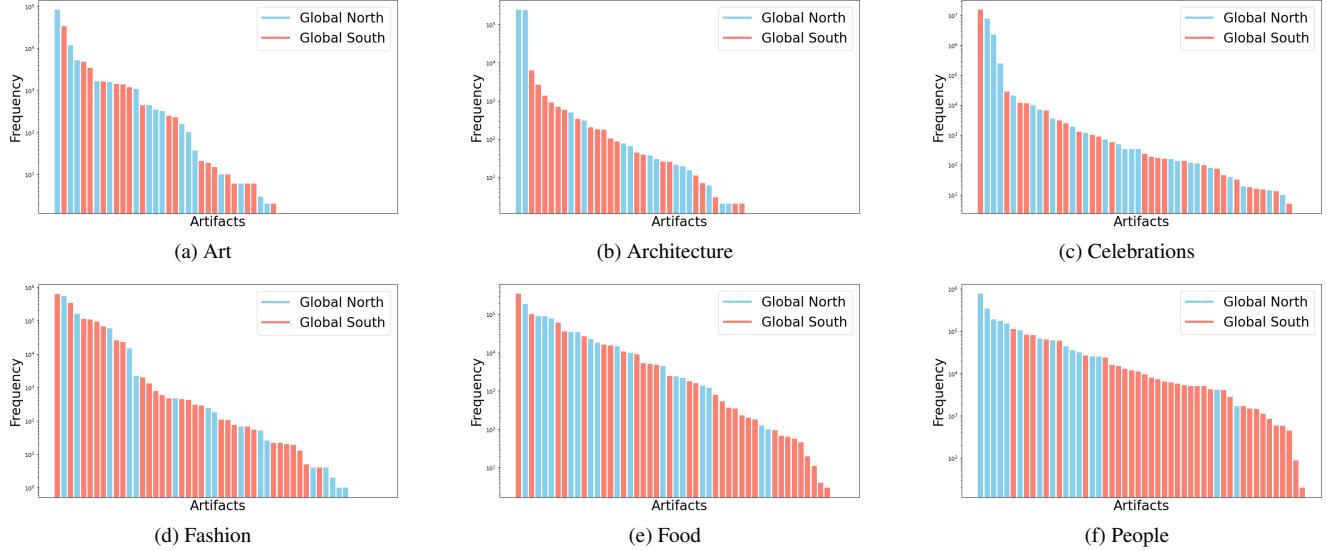


Figure 16. Frequency estimation (log scale) for 50 cultural artifacts across each of the six supercategories (cultural axes) in the CuRe dataset on LAION-SD15, the fully open pretraining dataset of Stable Diffusion 1.5 [50].

E. Concept Frequency Estimation

In recent times, state-of-the-art T2I systems do not have fully open pretraining data, *i.e.* it is unknown whether any particular image lies in the learned distribution of a modern T2I system. However, for the older Stable Diffusion 1.5, we have full access to its entire pretraining data, *i.e.* a mixture of LAION-2B-en⁹ [56] and LAION-Aesthetics V2 5+¹⁰, both of which are publicly and openly accessible. We can thus know with certainty whether a given image belongs to the training data of SD 1.5. We henceforth refer to this dataset mixture as **LAION-SD15**. As in Parashar et al. [44], we explicitly compute a concept frequency for every cultural artifact in the CuRe dataset on LAION-SD15 with string matching. We note that the original datasets on which SD 1.5 was trained were taken down and re-released due to problematic CSAM content [8, 59]; we use the re-released variants for our frequency computation. As seen in Fig. 16, all supercategories in our CuRe dataset show long tail behavior in LAION-SD15 across 50 cultural artifacts. We observe a very low occurrence of many artifacts in the art and architecture supercategories, as these are often specific and unique named entities which may occur rarely in the dataset when compared to more common items with high intra-class variance such as food, celebrations or people.

We also compute a Spearman rank correlation of the artifact occurrence frequency in LAION-SD15 with the ground truth likelihood Likert scores (see Appendix C for more details) provided by survey respondents for FLUX.1 [dev], SD

3.5 Large, and SD 1.5 in Tab. 13. We expect a higher occurrence frequency to correspond to better user judgments, as the T2I system should be able to accurately generate artifacts it has seen many times (the “head” of the T2I system distribution). We observe that for Celebrations, Food, Fashion, and People, there is a large positive correlation with user perceptions of the likelihood of the T2I system output $I(n)$ belonging to the class of artifacts n (*e.g.* assigning high score to the likelihood of images of spaghetti actually representing “spaghetti” as an artifact). Correlations are low for art and architecture, which we suspect is due to the large portion of their tail having very small frequency counts (Fig. 16). While the datasets used to train SD 3.5 Large and FLUX.1[dev] are not public and we cannot explicitly compute concept frequencies, since they show similar rank correlations to user judgments as SD 1.5, we predict that the CuRe dataset shows a similar long tail behavior across cultural artifacts for SD 3.5 and FLUX.1[dev] as well.

Table 13. Spearman rank correlation between occurrence frequency in the LAION-SD15 dataset of each cultural artifact in the CuRe dataset with human perceptions of ground-truth likelihood from the user study, ϕ_{GT}^* (Sec. 5.1 for details). Results are tabulated for each supercategory and for SD 1.5, SD 3.5 Large, and FLUX.1 [dev].

T2I	Supercategory					
	Art	Architecture	Celebrations	Food	Fashion	People
SD 1.5	0.04	0.10	0.40	0.59	0.34	0.22
SD 3.5 Large	0.05	0.13	0.48	0.55	0.27	0.23
FLUX.1 [dev]	0.01	0.06	0.30	0.38	0.44	0.34

⁹<https://huggingface.co/datasets/laion/relaion2B-en-research-safe>

¹⁰<https://laion.ai/blog/laion-aesthetics/>

F. Perceptual Similarity

As detailed in Appendix A.1, we generate images with increasing attribute specification for each artifact with three T2I systems, Stable Diffusion 1.5, Stable Diffusion 3.5 Large, and FLUX.1 [dev]. The prompts used are $P(n)$, $P(\{n, c\})$, $P(\{n, r\})$, and $P(\{n, c, r\})$ from Tab. 6.

We discuss in ablation to visually examine how our PS scorers capture T2I system performance in the cultural head and long tail in Appendix F.1. We also discuss qualitative examples from the CuRe dataset and how our PS scorers rate artifacts when compared to user judgments in Appendix F.2.

F.1. Perceptual Similarity as a Long Tail Predictor

For each T2I system f_θ , in Fig. 17 we show a scatter plot of raw perceptual similarity scores ϕ_{GT} and ϕ_{PS} for all 50 cultural artifacts belonging to each supercategory s . For each artifact, we plot two scores: $\phi(n)$ and $\arg \max \phi(a)$, i.e. the highest similarity score across the other three prompts. To examine if the head and tail of the distribution of PS scores correlate with the Global North / South divide, we also color each point according to whether the region the artifact belongs to lies in the Global North (red) or South (black). Below we discuss the “spread” or visual divergence in scores between $\phi(n)$ and $\arg \max \phi(a)$ as well as “cultural outlier”, i.e. points with a high spread in the head or points with a low spread in the tail.

As seen in Fig. 17, we observe a spread of scores in the tail across T2I systems for both scorers ϕ_{GT} and ϕ_{PS} , though the starting point and magnitude of spread (vertical height) differs by supercategory. ϕ_{GT} tends to have a lower spread than ϕ_{PS} , with fewer outliers. All T2I systems have very few outliers in the head (i.e. points with high spread), while there tend to be many more outliers in the tail, i.e. points with low spread (e.g. for Art and People.) People has the least spread across all T2I systems, which is intuitive as there is a very specific way that a certain individual looks, and thus perceptual similarity tends not to change with attribute specification. Interestingly, even though Architecture and Art are similarly also singular named entities (i.e. minimal intra-artifact variance in visual features), we observe a later starting point with occasionally large spread (e.g. only in the last 10 to 15 artifacts). Food, Fashion and Celebrations have in general the widest spread, as there is generally large intra-class diversity / variance as to how culture-specific food preparations, clothing, and celebrations looks visually. DALL-E 3 (Fig. 17e) appears to be the most homogenous in perceptual similarity, as the spread is much lesser than other T2I systems, even for Food.

For a more fine-grained analysis than a Global North / South divide, we also examine PS scores across each region (country) in the CuRe dataset in Appendix F, grouped by continent. To measure CuRe of a T2I system f_θ for region

r , we compute an average score over the set of cultural artifacts associated with the region, \mathcal{N}_r . That is,

$$\text{CuRe}(\phi; f_\theta, r) = \frac{1}{|\mathcal{N}_r|} \sum_{n \in \mathcal{N}_r} \phi(I).$$

We observe that when aggregated, outside of North America, no scorer is able to capture the clear trend in human judgments of cultural representativeness ϕ_{CuRe}^* . This indicates that whether a cultural artifact lies in the head or tail of the T2I system distribution cannot be indicated purely by membership to a global region.

F.2. Qualitative Analysis of PS Scorers

We highlight several qualitative examples on the CuRe dataset with our perceptual similarity scorers ϕ_{PS} and $\Delta\phi_{PS}$ compared to the strong baseline scorer $\phi_{GT}(n)$ in Fig. 18. We also show Likert scores for cultural representativeness ϕ_{CuRe}^* and textual justification for each example.

As we observed in our quantitative results (Tab. 2), our divergence PS scorer $\Delta\phi_{PS}(\{n, c\})$ correlates strongly (and negatively) with the ϕ_{CuRe}^* gold scores, i.e. low $\Delta\phi_{PS}(\{n, c\})$ indicating higher ϕ_{CuRe}^* , e.g. **E1** (*Bayt al-Suhaymi*), **E2** (*Moai*), **E3** (*George Lucas*). We also highlight a failure case for our scorers with **E4** (*Tallarín saltado*), a type of noodle dish from Peru generated by SD 1.5. While we expect a low $\Delta\phi_{PS}$ score to correspond to high cultural representativeness, workers rate a low ϕ_{CuRe}^* score of 1 out of 5 and indicate errors with textual response, i.e. incorrect ingredients and out-of-place textures. ϕ_{PS} also did worse than the baseline $\phi_{GT}(n)$ as it gave a higher perceptual similarity score (falsely predicting high quality T2I output). This failure aligns with our quantitative rank correlation analysis in Tab. 2, where we observe very Spearman’s ρ for SD 1.5, an older and smaller T2I system.

We also highlight two cases where both our proposed scorers and the baseline scorer fail. **E5** (*Blocos carnavalescos de São Paulo*), a carnival celebration from Brazil, had a 1 out of 5 ϕ_{CuRe}^* for SD 3.5 Large. The written feedback also reflects that the images are stereotypically extravagant and incorrect in physical appearance. However, the quantitative perceptual similarity scorers were very high compared to other artifacts with the same low ϕ_{CuRe}^* score. In contrast, **E6** (*Hmong textile art*), a style of embroidery from Vietnam had an excellent 5 out of 5 ϕ_{CuRe}^* for SD 3.5 Large, but lower scores from the quantitative PS scorers than **E5** which was much less culturally representative according to human judgments.

F.3. PS vs Concept Frequency

In Tab. 15 we show the Spearman rank correlation of the frequency the artifact name appears in the captions of LAION-SD1.5 (the pretraining dataset of Stable Diffusion 1.5; details in Appendix E) and the perceptual similarity scores

Table 14. Perceptual similarity CuRe scores of all regions in the CuRe dataset, sorted descending by user study scores ϕ_{CuRe}^* . The scores are averaged across Flux.1 [dev], SD 3.5 Large, SD 1.5, and SDXL T2I systems and SigLIP2, AIMV2, and DINOv2 encoders.

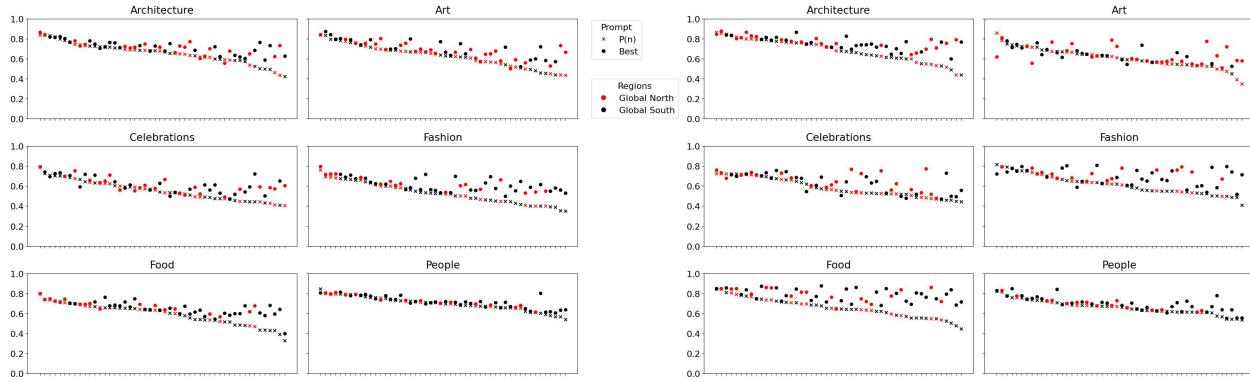
Continent	Region	$\phi_{\text{CuRe}}^* \uparrow$	Perceptual Similarity Scorer			
			$\phi_{GT}(n) \uparrow$	$\phi_{PS}(n) \uparrow$	$\Delta\phi_{PS}(\{n, c\}) \downarrow$	$\Delta\phi_{PS}(\{n, c, r\}) \downarrow$
Africa	Egypt	3.24 ± 0.88	0.72 ± 0.54	0.62 ± 0.44	0.52 ± 0.35	0.51 ± 0.33
	Nigeria	2.96 ± 0.91	0.72 ± 0.52	0.64 ± 0.49	0.51 ± 0.44	0.51 ± 0.44
	Morocco	2.80 ± 1.21	0.64 ± 0.16	0.62 ± 0.19	0.55 ± 0.22	0.55 ± 0.25
	Algeria	2.72 ± 0.87	0.56 ± 0.04	0.57 ± 0.07	0.56 ± 0.11	0.57 ± 0.11
	Ghana	2.69 ± 0.90	0.59 ± 0.16	0.59 ± 0.17	0.55 ± 0.25	0.53 ± 0.22
	South Africa	2.53 ± 1.10	0.64 ± 0.13	0.63 ± 0.14	0.55 ± 0.14	0.54 ± 0.12
	Ethiopia	2.28 ± 0.30	0.61 ± 0.11	0.62 ± 0.11	0.54 ± 0.20	0.54 ± 0.15
	Zimbabwe	2.22 ± 0.03	0.77 ± 0.05	0.69 ± 0.07	0.49 ± 0.02	0.50 ± 0.02
	Tunisia	1.74 ± 0.29	0.50 ± 0.05	0.53 ± 0.06	0.63 ± 0.14	0.63 ± 0.15
Asia	India	3.91 ± 0.65	0.69 ± 0.72	0.65 ± 0.55	0.52 ± 0.46	0.51 ± 0.45
	Malaysia	3.33 ± 0.22	0.74 ± 0.07	0.78 ± 0.04	0.51 ± 0.02	0.51 ± 0.02
	Pakistan	3.02 ± 0.85	0.76 ± 0.31	0.61 ± 0.20	0.50 ± 0.11	0.50 ± 0.14
	Indonesia	2.88 ± 0.83	0.66 ± 0.53	0.63 ± 0.46	0.53 ± 0.49	0.53 ± 0.54
	Iran	2.86 ± 1.00	0.61 ± 0.24	0.56 ± 0.17	0.55 ± 0.26	0.55 ± 0.23
	Philippines	2.77 ± 0.69	0.64 ± 0.37	0.60 ± 0.26	0.54 ± 0.34	0.54 ± 0.36
	Japan	2.72 ± 1.45	0.65 ± 0.21	0.61 ± 0.23	0.55 ± 0.23	0.54 ± 0.23
	China	2.67 ± 0.76	0.65 ± 0.56	0.62 ± 0.44	0.53 ± 0.48	0.53 ± 0.47
	Lebanon	2.67 ± 0.52	0.72 ± 0.04	0.76 ± 0.06	0.49 ± 0.02	0.51 ± 0.05
	Thailand	2.25 ± 0.83	0.67 ± 0.04	0.66 ± 0.07	0.55 ± 0.06	0.54 ± 0.06
	Israel	2.44 ± 0.25	0.70 ± 0.06	0.71 ± 0.13	0.48 ± 0.12	0.49 ± 0.10
	Korea	2.42 ± 0.81	0.66 ± 0.20	0.63 ± 0.14	0.53 ± 0.17	0.53 ± 0.12
	Vietnam	2.14 ± 1.23	0.59 ± 0.13	0.59 ± 0.13	0.54 ± 0.14	0.54 ± 0.16
	Bangladesh	2.11 ± 1.11	0.53 ± 0.08	0.55 ± 0.09	0.59 ± 0.13	0.58 ± 0.16
	Afghanistan	1.89 ± 0.32	0.80 ± 0.02	0.59 ± 0.03	0.50 ± 0.01	0.50 ± 0.01
	Singapore	1.89 ± 0.17	0.79 ± 0.02	0.70 ± 0.04	0.53 ± 0.04	0.52 ± 0.04
Europe	Slovenia	3.41 ± 0.19	0.63 ± 0.08	0.63 ± 0.09	0.57 ± 0.06	0.56 ± 0.07
	Czech Republic	3.28 ± 0.94	0.73 ± 0.05	0.69 ± 0.06	0.51 ± 0.05	0.52 ± 0.04
	Switzerland	3.19 ± 0.92	0.61 ± 0.08	0.62 ± 0.11	0.51 ± 0.12	0.50 ± 0.09
	United Kingdom	3.14 ± 0.70	0.74 ± 0.52	0.62 ± 0.41	0.51 ± 0.28	0.51 ± 0.30
	Denmark	3.11 ± 1.28	0.59 ± 0.05	0.58 ± 0.04	0.58 ± 0.08	0.57 ± 0.07
	Germany	3.11 ± 1.14	0.62 ± 0.24	0.58 ± 0.14	0.56 ± 0.20	0.56 ± 0.18
	Norway	3.11 ± 0.10	0.57 ± 0.05	0.59 ± 0.03	0.55 ± 0.06	0.55 ± 0.05
	Italy	3.01 ± 1.46	0.70 ± 0.35	0.66 ± 0.39	0.53 ± 0.36	0.53 ± 0.28
	Netherlands	3.00 ± 1.58	0.64 ± 0.12	0.58 ± 0.10	0.55 ± 0.14	0.55 ± 0.26
	Poland	3.00 ± 0.93	0.63 ± 0.12	0.61 ± 0.18	0.53 ± 0.22	0.54 ± 0.21
	France	2.98 ± 1.28	0.69 ± 0.23	0.66 ± 0.24	0.52 ± 0.24	0.52 ± 0.25
	Russia	2.83 ± 0.58	0.68 ± 0.22	0.65 ± 0.07	0.54 ± 0.05	0.54 ± 0.05
	Greece	2.73 ± 1.51	0.69 ± 0.17	0.64 ± 0.30	0.53 ± 0.28	0.53 ± 0.20
	Belgium	2.72 ± 0.53	0.71 ± 0.18	0.64 ± 0.17	0.50 ± 0.12	0.50 ± 0.28
	Spain	2.62 ± 0.76	0.68 ± 0.44	0.63 ± 0.31	0.52 ± 0.29	0.52 ± 0.28
	Portugal	2.60 ± 1.74	0.61 ± 0.03	0.59 ± 0.06	0.56 ± 0.05	0.57 ± 0.07
	Austria	2.52 ± 1.24	0.62 ± 0.09	0.62 ± 0.11	0.53 ± 0.20	0.53 ± 0.15
	Slovakia	2.41 ± 0.76	0.66 ± 0.15	0.64 ± 0.17	0.54 ± 0.13	0.53 ± 0.17
	Hungary	2.33 ± 0.52	0.68 ± 0.03	0.71 ± 0.04	0.51 ± 0.02	0.51 ± 0.02
	Romania	2.33 ± 0.22	0.61 ± 0.03	0.60 ± 0.02	0.58 ± 0.05	0.59 ± 0.04
	Croatia	2.22 ± 0.47	0.53 ± 0.01	0.56 ± 0.02	0.60 ± 0.05	0.58 ± 0.03
	Finland	2.20 ± 0.63	0.65 ± 0.26	0.64 ± 0.24	0.51 ± 0.18	0.52 ± 0.34
	Turkey	2.19 ± 0.40	0.66 ± 0.10	0.67 ± 0.24	0.53 ± 0.21	0.52 ± 0.37
	Sweden	2.09 ± 0.82	0.67 ± 0.11	0.63 ± 0.11	0.53 ± 0.14	0.53 ± 0.14
	Ukraine	1.74 ± 0.61	0.52 ± 0.07	0.56 ± 0.09	0.60 ± 0.18	0.58 ± 0.15

Continent	Region	Perceptual Similarity Scorer				
		$\phi_{\text{CuRe}}^* \uparrow$	$\phi_{GT}(n) \uparrow$	$\phi_{PS}(n) \uparrow$	$\Delta\phi_{PS}(\{n, c\}) \downarrow$	$\Delta\phi_{PS}(\{n, c, r\}) \downarrow$
North_America	United States	3.46 ± 0.99	0.73 ± 0.87	0.66 ± 0.84	0.51 ± 0.75	0.51 ± 0.71
	Canada	3.30 ± 0.43	0.66 ± 0.07	0.73 ± 0.16	0.50 ± 0.12	0.50 ± 0.09
	Mexico	2.89 ± 1.30	0.66 ± 0.45	0.68 ± 0.61	0.52 ± 0.55	0.52 ± 0.53
	Puerto Rico	2.56 ± 0.17	0.50 ± 0.01	0.51 ± 0.01	0.67 ± 0.03	0.65 ± 0.05
	Costa Rica	1.44 ± 0.40	0.60 ± 0.02	0.67 ± 0.06	0.54 ± 0.07	0.55 ± 0.04
Oceania	Australia	2.87 ± 1.18	0.66 ± 0.16	0.69 ± 0.22	0.52 ± 0.17	0.52 ± 0.18
	New Zealand	2.78 ± 0.88	0.62 ± 0.12	0.62 ± 0.12	0.54 ± 0.10	0.53 ± 0.10
South_America	Brazil	3.11 ± 1.15	0.65 ± 0.51	0.66 ± 0.54	0.51 ± 0.55	0.51 ± 0.56
	Venezuela	2.89 ± 0.32	0.48 ± 0.02	0.55 ± 0.03	0.60 ± 0.05	0.61 ± 0.05
	Chile	2.85 ± 0.49	0.74 ± 0.43	0.63 ± 0.23	0.51 ± 0.16	0.51 ± 0.20
	Argentina	2.40 ± 1.40	0.72 ± 0.08	0.68 ± 0.09	0.52 ± 0.09	0.51 ± 0.08
	Peru	2.29 ± 1.02	0.58 ± 0.21	0.62 ± 0.30	0.54 ± 0.31	0.53 ± 0.37
	Colombia	2.11 ± 0.56	0.64 ± 0.17	0.62 ± 0.18	0.54 ± 0.18	0.53 ± 0.14
	Ecuador	1.67 ± 0.00	0.54 ± 0.02	0.53 ± 0.02	0.62 ± 0.09	0.60 ± 0.08

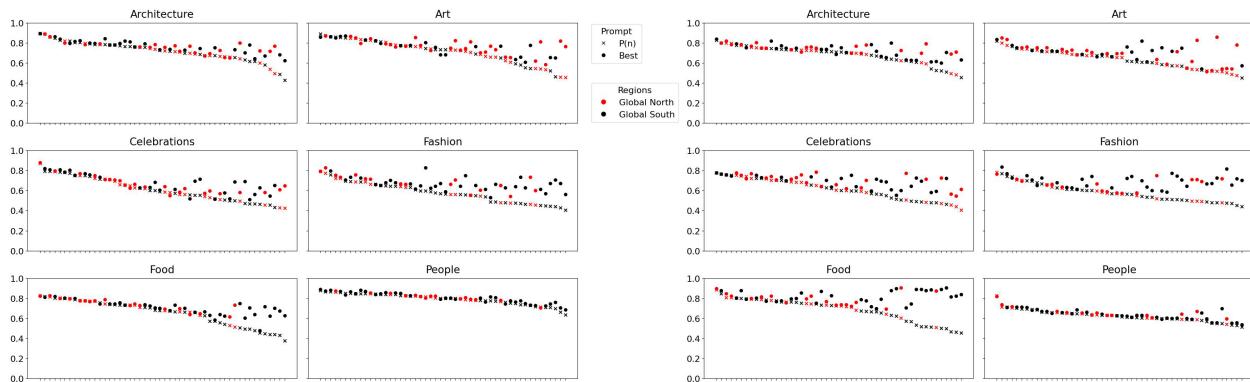
Table 15. Spearman’s ρ between frequency count of each artifact name in LAION-SD15 (see Appendix E) and perceptual similarity scorers $\phi_{GT}(n)$ and $\phi_{PS}(n)$ evaluated on the CuRe dataset.

Encoder	SD 1.5		SD 3.5 Large		SDXL		Flux.1 [dev]	
	ϕ_{GT}	ϕ_{PS}	ϕ_{GT}	ϕ_{PS}	ϕ_{GT}	ϕ_{PS}	ϕ_{GT}	ϕ_{PS}
SigLIP 2 [63]	0.64	0.22	0.58	0.34	0.55	0.25	0.60	0.37
AIMV2 [20]	0.53	-0.08	0.52	0.25	0.42	0.08	0.49	0.33
DINOv2 [43]	0.57	0.25	0.56	0.40	0.46	0.29	0.55	0.45

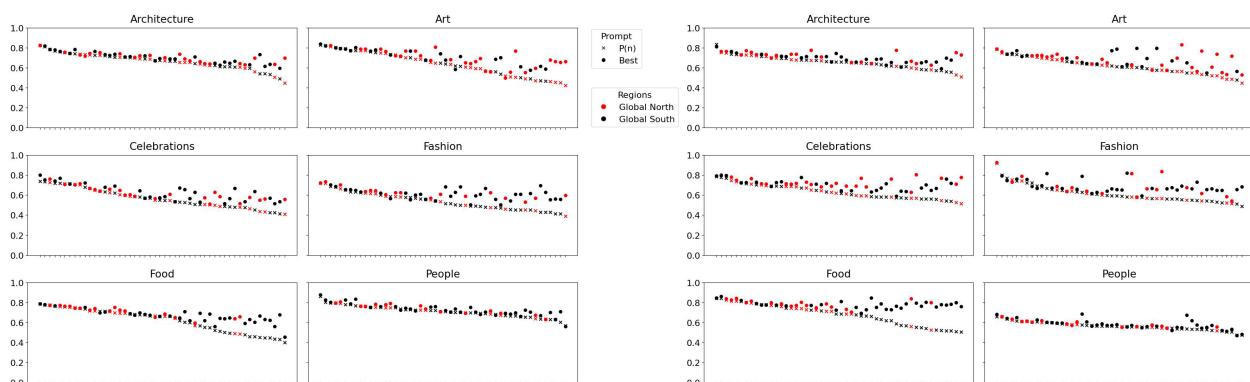
$\phi_{GT}(n)$ and $\phi_{PS}(n)$, which use images generated with the prompt only containing the artifact name. For $\phi_{GT}(n)$, SD 1.5 has the highest positive correlation out of all T2I systems across encoders, which is expected as we evaluate frequency over its exact pretraining dataset. On the contrary for $\phi_{PS}(n)$, Spearman’s rank correlation to concept frequency in LAION-SD15 followed Arena ELO rankings, *i.e.* FLUX.1 [dev] > SD 3.5 Large > SDXL > SD 1.5. This likely indicates that artifact - category associations for stronger T2I systems like FLUX.1 [dev] and SD 3.5 were reinforced by the larger size and coverage of their pretraining data.



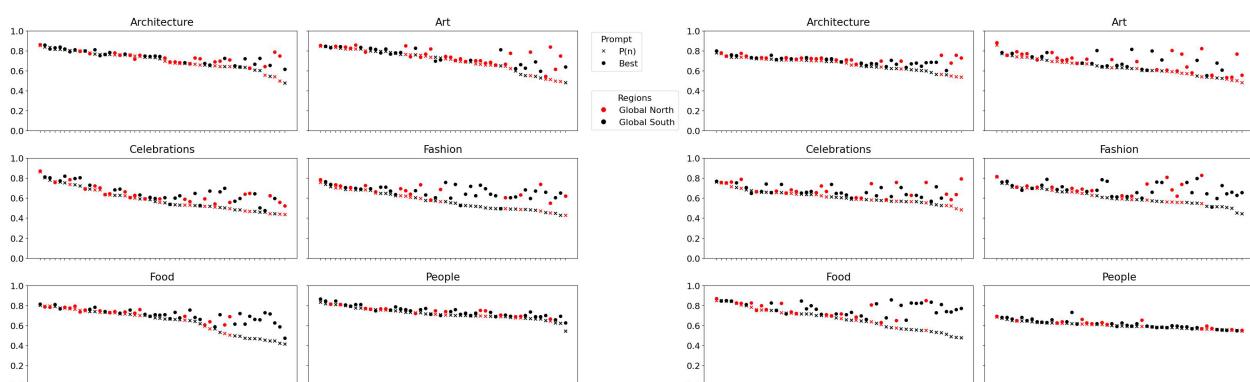
(a) FLUX.1 dev



(b) Stable Diffusion 3.5 Large



(c) Stable Diffusion 1.5



(d) Stable Diffusion XL

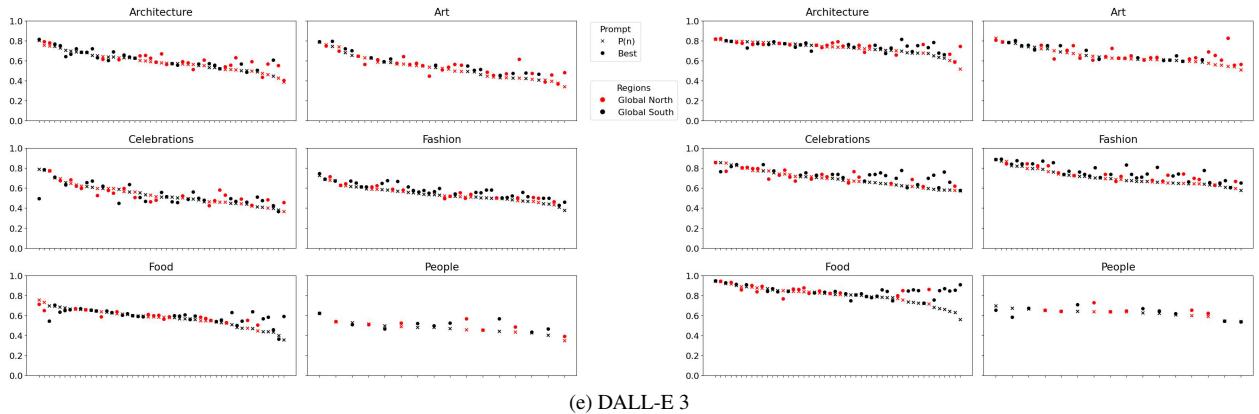


Figure 17. Scatter plots of **Left:** $\phi_{GT}(a)$ and **Right:** $\phi_{PS}(a)$ for 50 artifacts from all 6 supercategories in the CuRe dataset generated using FLUX.1 [dev]. Each data point represents a single cultural artifact, and is an average of the scores for each T2I generated seed. All images are encoded with SigLIP 2. We plot $\phi_{GT}(n)$ and $\phi_{PS}(n)$ with an “x” marker, while highest similarity score amongst all four prompt styles ($P(n)$, $P(n, c)$, $P(n, r)$, $P(n, c, r)$) has the “o” marker. The colors indicate whether the region the artifact belongs to lies in the Global North (red) or Global South (black). Artifacts are sorted by descending $\phi_{PS}(n)$ similarity score, i.e. our marginal attribution score. As seen in the figure, there is a noticeable divergence of perceptual similarity score of the most underspecified ($P(n)$) from the “best” prompt in the long tail, across T2I systems.

E#	AI Image	Real Image	Feedback	PS Scorers
E1			<p><i>Bayt al-Suhaymi</i></p> <p>1. The trees used looks so weird as its mimicing a dome shape, some of the small blue domes are futher apart than others. asymtrical look</p> <p>2. There is a lot of Dom's typically there is only one in any building, this picture represent a mosque more than a normal building, also the Dom's can't be made from trees</p>	$\phi_{\text{CuRe}}^* = 2.0/5$ $\phi_{GT}(n) = 0.654$ $\phi_{PS}(n) = 0.580$ $\Delta\phi_{PS}(n, c) = 0.552$ $\Delta\phi_{PS}(n, c, r) = 0.506$
E2			<p><i>Moai</i></p> <p>1. the head to body ratio is a bit off, the body should either be more elongated or the head bigger, some even had some kind of red hat and they all looked to the ocean.</p> <p>2. The main thing that I think is inaccurate is that the moai in the image looks too perfect and the real ones have a lot of damage due to nature.</p>	$\phi_{\text{CuRe}}^* = 4.0/5$ $\phi_{GT}(n) = 0.808$ $\phi_{PS}(n) = 0.581$ $\Delta\phi_{PS}(n, c) = 0.486$ $\Delta\phi_{PS}(n, c, r) = 0.486$
E3			<p><i>George Lucas</i></p> <p>1. This is a closeup picture of George Lucas' face. There's nothing particularly unique about it that would NOT make it apart of my culture. I don't see why I couldn't see this kind of photo in my culture.</p> <p>2. I do not see anything that is "wrong" other than the eyes. They are obviously AI.</p>	$\phi_{\text{CuRe}}^* = 5.0/5$ $\phi_{GT}(n) = 0.745$ $\phi_{PS}(n) = 0.571$ $\Delta\phi_{PS}(n, c) = 0.489$ $\Delta\phi_{PS}(n, c, r) = 0.460$
E4			<p><i>Tallarin saltado</i></p> <p>1. the noodles look like worms, tallarin saltado is spaghetti with meat and vegetables whereas the image only shows weird pasta and no meat nor onions/tomatoes.</p> <p>2. The image doesn't look like noodles to me, it seems like a kind of vegetable so I can't say is accurate to the real Tallarfn saltado</p>	$\phi_{\text{CuRe}}^* = 1.0/5$ $\phi_{GT}(n) = 0.654$ $\phi_{PS}(n) = 0.696$ $\Delta\phi_{PS}(n, c) = 0.486$ $\Delta\phi_{PS}(n, c, r) = 0.425$
E5			<p><i>Blocos carnavalescos de São Paulo</i></p> <p>1. The costumes are indeed carnival like, but the blocos are less stravagant and the costumes wore by the people usually contain some kind of a joke on a very low budget costume</p> <p>2. Their physical appearance and posture.</p>	$\phi_{\text{CuRe}}^* = 1.0/5$ $\phi_{GT}(n) = 0.712$ $\phi_{PS}(n) = 0.757$ $\Delta\phi_{PS}(n, c) = 0.504$ $\Delta\phi_{PS}(n, c, r) = 0.473$
E6			<p><i>Hmong textile art</i></p> <p>1. The design motif and colors are quite close to the real ones</p> <p>2. I think the repetitive patterns and the colourful decoration make it accurate</p>	$\phi_{\text{CuRe}}^* = 5.0/5$ $\phi_{GT}(n) = 0.620$ $\phi_{PS}(n) = 0.634$ $\Delta\phi_{PS}(n, c) = 0.519$ $\Delta\phi_{PS}(n, c, r) = 0.518$

Figure 18. Visualization of quantitative perceptual similarity (PS) scorers and user study CuRe scores and text feedback (Appendix C.4). The first two artifacts were generated with FLUX.1 [dev], the second two with SD 1.5, and the last two with SD 3.5 Large.

G. Image-Text Alignment

We examine an important factor in the context of image-text alignment scorers in Appendix G.1: what impact does the choice of vision-language model backbone have on scorer quality? We also discuss qualitative examples with ITA scorers in Appendix G.2.

Table 16. List of prompts used for text-image similarity score calculations. *e.g.* for region $r = \text{"Australia"}$, the T2I prompt $P(r) = \text{"An image from Australia"}$.

Scorer	Prompt
Khanuja et al. [31]	"This image is culturally relevant to {r}."
Ventura et al. [65]	"Image from {r} culture."
o3 mini [42]	"Assess the image's cultural representation of {r}."
$P(n)$	"An image of {n}."
$P(c)$	"An image of {c}."
$P(r)$	"An image from {r}."
$P(c, r)$	"An image of {c} from {r}."

G.1. Choice of ITA Scorer’s VLM Backbone

Table 17. Spearman correlation values of ITA scorers with human judgments of perceptual similarity ϕ_{PS}^* across ITA scorer backbones for FLUX.1 [dev]. The backbones we evaluate include SigLIP 2 and OpenCLIP models trained on OpenAI WIT [47], Data Filtering Networks (DFN-5B) [19], LAION-2B [56], and DataComp (DC-1B) [22].

ITA Scorer	LAION-2B	WIT	DFN-5B	SigLIP 2	DC-1B
Khanuja et al. [31]	0.18	0.16	0.12	0.11	0.12
Ventura et al. [65]	0.16	0.09	0.16	0.14	0.16
o3-mini	0.17	0.17	0.13	0.14	0.15
$sim(I(n), P(n))$	0.35	0.39	0.41	0.38	0.37
$sim(I(n), P(c))$	0.33	0.38	0.35	0.34	0.31
$sim(I(n), P(r))$	0.16	0.09	0.17	0.12	0.13
$sim(I(n), P(c, r))$	0.37	0.39	0.38	0.38	0.34
$\phi_{ITA}(c)$	0.39	0.43	0.43	0.40	0.39
$\phi_{ITA}(r)$	0.32	0.32	0.37	0.35	0.33
$\phi_{ITA}(c, r)$	0.40	0.43	0.44	0.42	0.40

We replicate the Spearman rank correlation setup from Tab. 3 and ablate over the choice of VLM backbone used to compute image-text alignment for FLUX.1 [dev] in Tab. 17 (see Sec. 5.4 for VLM details). To recap, we compute a Spearman’s ρ of each scorer with the user study gold score ϕ_{PS}^* . We observe that baselines (Khanuja et al. [31], Ventura et al. [65], and o3 mini [42]), which query the VLM for CuRe score directly with region information r , are sensitive to changes in the backbone and show high variability in ITA scores. As we marginally increase attributes in the evaluation prompt $P(n \rightarrow c, n \rightarrow r, n \rightarrow \{c, r\})$, the rank correlations become more consistent across VLM backbones.

Our proposed metrics show both higher and more consistent rank correlations with ϕ_{PS}^* gold scores across all VLM backbones, showing that they are more robust than baselines to the pretraining distribution of the VLM for evaluating cultural representativeness.

G.2. Qualitative Analysis of ITA Scorers

We highlight several qualitative examples on the CuRe dataset of our ITA scorer $\phi_{ITA}(c, r)$ compared to baselines in Fig. 19 using SigLIP 2 as the VLM backbone. We also show Likert scores for cultural representativeness ϕ_{CuRe}^* and textual justification for each example.

Throughout our evaluation, we treat user judgments as the gold standard, assessing scorers based on how well they replicate human feedback. **E5** (*Zwölf Glaubensartikel*) is a rare case where users were unfamiliar with the artifact itself, leading to a high ϕ_{CuRe}^* based primarily on regional similarity. Since the AI-generated and real images differ in category, the $\phi_{ITA}(c, r)$ score is correspondingly low, reflecting this mismatch.

As seen in **E1** (*Bangles*), our proposed metric $\phi_{ITA}(c, r)$ aligns more closely with user preferences compared to existing baseline scorers. Our proposed metric proves particularly robust in scenarios where the T2I system generates outputs that are categorically incorrect. In **E3** (*Jalangkote*), SD 3.5 Large generates an image of architecture rather than food, a failure undetected by baseline metrics that focus narrowly on regional resemblance. From a user perspective, representativeness encompasses not only regional cues but also correct category and item-level semantics, an area where our marginal information attribution scorer provides more robust signal.

We also highlight some failure cases of our ITA scorer. **E4** (*Puchner Mansion*) shows a case where baseline scorers better approximated human judgments than $\phi_{ITA}(c, r)$, suggesting that in some cases, the baselines’ focus on broader stylistic features may offer advantages. In **E6** (*Michelle Bachelet*), our proposed $\phi_{ITA}(c, r)$ captures semantic representativeness by integrating category and region cues, but does not account for image quality. Users tend to penalize low-quality or unrealistic images regardless of semantic alignment, which our scorer overlooks. **E2** (*Festival de la Primavera*) demonstrates a failure case across all ITA-based scorers which fail to capture culturally specific or context-dependent cues (region inconsistency, incorrect details of the parade).

Note on Worker Reliability: While most participants provided thoughtful and culturally grounded feedback, a few responses reflected exasperation with AI-generated outputs in general, including calls to “stop this”. This highlights an important meta-issue regarding images created by T2I systems, *i.e.* a subjectivity and heterogeneity of opinions towards T2I systems themselves.

E#	AI Image	Real Image	Feedback	ITA Scorers
E1			<p>Bangles</p> <p>1. Bangles are circular ornaments with some colors and patterns on them.</p> <p>2. The width and the carvings and the colors make it seem more aligned to my culture</p>	$\phi_{\text{CuRe}}^* = 5.0/5$ Khanuja et al. [31] = 0.060 Ventura et al. [65] = 0.045 $\phi_{ITA}(c, r) = 0.124$
E2			<p>Festival de la Primavera</p> <p>1. The festival takes place in the cost of Lima, no there are not a lot of hilly areas, and it's more a dry environment so those flowers and the place do not correspond at all to Trujillo reality.</p> <p>2. The image represents the Spring in general. But the main event of the Festival de la Primavera is the parade (People, carriages and flowers).</p>	$\phi_{\text{CuRe}}^* = 1.0/5$ Khanuja et al. [31] = 0.086 Ventura et al. [65] = 0.067 $\phi_{ITA}(c, r) = 0.156$
E3			<p>Jalangkote</p> <p>1. The image use the wrong thing, AI image is using stone like building while Jalangkote is a food. a complete different genre</p> <p>2. The AI image is not even a food.</p>	$\phi_{\text{CuRe}}^* = 1.7/5$ Khanuja et al. [31] = 0.112 Ventura et al. [65] = 0.093 $\phi_{ITA}(c, r) = 0.022$
E4			<p>Puchner Mansion</p> <p>1. It's style is just all wrong. it looks more like something an american would think of as a castle. can we stop trying to make ai image gen happen?</p> <p>2. I feel like the AI focused too much on the word Mansion, and especially the american/older british kind.</p>	$\phi_{\text{CuRe}}^* = 1.3/5$ Khanuja et al. [31] = 0.040 Ventura et al. [65] = 0.053 $\phi_{ITA}(c, r) = 0.096$
E5			<p>Zwölf Glaubensartikel</p> <p>1. Overall the vegetation seems possible to find in my country</p> <p>2. This image could be a shot from above of some forest in my country.</p>	$\phi_{\text{CuRe}}^* = 4.7/5$ Khanuja et al. [31] = 0.108 Ventura et al. [65] = 0.091 $\phi_{ITA}(c, r) = 0.079$
E6			<p>Michelle Bachelet</p> <p>1. The image doesn't have the exact facial features of Michelle Bachelet. She's a well known Chilean politician, so a lot of people recognize her face. The image looks too fat and the hair is wrong.</p> <p>2. Its the features of the face, they are not completely wrong, but together they make a face who cant be taken to be Bachelet, also the hair looks really fake.</p>	$\phi_{\text{CuRe}}^* = 2.0/5$ Khanuja et al. [31] = 0.066 Ventura et al. [65] = 0.054 $\phi_{ITA}(c, r) = 0.164$

Figure 19. Visualization of quantitative Image-Text Alignment (ITA) scorers, user study CuRe scores and text feedback. The first two artifacts were generated with FLUX.1 [dev], the second two generated with SD 3.5 Large, and the last two generated with SD 1.5.

H. Diversity

We compute two notions of diversity: a) **intra-artifact**, over T2I generations with multiple random seeds for a single cultural artifact n ; b) **intra-category**, over T2I generations of multiple random seeds for a single cultural category c . We discuss motivating examples for these two separate notions in Sec. 4.4). Details for seeding for different T2I systems is provided in Appendix A.1.

LPIPS. A high LPIPS [76] score indicates high **intra-artifact** variance in patchwise image features across seeds, which is interpreted as high diversity. We compute LPIPS for a category c in our CuRe dataset as an average of scores of all artifacts n associated with category c , *i.e.*

$$LPIPS(c) = \frac{1}{|c|} \sum_{n \in c} \phi(I(n)).$$

Vendi Scores. Vendi Scores (VS) [21, 30] define a similarity measure via a kernel over selected attributes (*e.g.* $r = \{\text{Country}, \text{Continent}\}$). While computing VS, each seed j of an image of a cultural category $I(c)$ is assigned a predicted label based on its “closest” image in the set of cultural artifacts n belonging to category c to compute **intra-category** diversity, *i.e.*

$$\hat{n}(I(c)_j) = \arg \max_{n \in c} sim(I(c), I(n)).$$

A $j \times j$ kernel similarity matrix is then computed based on a selected attribute (see Experimental Setup of Kannen et al. [30] for details). The primary drawback of VS as a scorer is that its quality depends entirely on this initial assignment \hat{n} and choice of kernel, which in turn depends on the image encoder used to compute sim .

H.1. Diversity as a Long Tail Predictor

We examine if we can use diversity as a predictor of a cultural artifact lying in the head or long tail of a T2I system’s distribution over generated images, similar to our marginal attribution lens over perceptual similarity (Sec. 4.2). Recall from Eq. (4) that our ϕ_{DIV} scorer compute LPIPS diversity over a set of images generated across attribute specification levels:

$$\phi_{DIV} = LPIPS(\{n\}, \{n, c\}, \{n, r\}, \{n, c, r\})$$

We expect that in the long tail, a larger marginal increase in information $n \rightarrow a$ will cause a large increase in diversity, as there will be a larger visual difference between images generated with different attributes $I(n) \rightarrow I(a)$ (as seen with the “Banku” artifact in Fig. 3). Intuitively, this evaluates how much diversity changes when we mix images of “Banku” with images of “Banku, a type of dumpling”,

images of “Banku, from Ghana”, and images of “Banku, a type of dumpling from Ghana”. If it is relatively unchanged, we predict it to lie in the head of the supercategory artifact distribution.

To test this hypothesis, we compute the marginal increase in information as divergence of ϕ_{DIV} from an LPIPS over only images generated with $a = \{n\}$, *i.e.*

$$\Delta\phi_{DIV} = \phi_{DIV} - LPIPS(n)$$

We visualize this marginal increase $\Delta\phi_{DIV}$ on the CuRe dataset by grouping cultural artifacts into two buckets as proxies for the head and long tail: artifacts belonging to regions from the global north and global south. We average $\Delta\phi_{DIV}$ over all artifacts belonging to each bucket for each supercategory s , and show a bar plot for all T2I systems in Fig. 20.

We observe two clusters of T2I system behavior with our scorer: in the first cluster, FLUX.1 [dev] and SD 3.5 large (the highest quality T2I systems per Arena Bench ELO, Tab. 5) show higher change in diversity with increasing information in the Global South (proxy for long tail) than the Global North (proxy for head) across supercategories, which aligns with our hypothesis given our assumed proxies. In the second cluster, Ideogram 2.0, Stable Diffusion XL and DALL-E 3, this behavior is reversed, except for architecture and people for SDXL and DALL-E 3 (which has a high 66.5% refusal rate for generating images of people due to inbuilt safety filters, see Appendix A.2). Stable Diffusion 1.5 is the weakest T2I system, and follows similar trends to the second cluster, except for Art. Architecture consistently has higher divergence in the Global North than the South, across T2I systems, which we suspect is due to the inherent lack of diversity in generations of specific named entities, *i.e.* low intra-artifact variance (a similar problem to art and people).

In summary, our proposed scorer ϕ_{DIV} can thus predict where in its learned distribution an image generated by a T2I system lies, *i.e.* in the head or long tail. We note our pre-assigned clustering of global north and south is just *one example proxy* for head and long tail. We note while our scorer requires generating multiple seeds of images with different attribute specification levels (practically, four seeds across four styles = 16 total images), it is still relatively cheaper to compute when compared to computing artifact frequency over a large pretraining dataset via string matching, such as in Parashar et al. [44], which we discuss in more detail in Appendix E.

H.2. Qualitative Analysis of DIV scorers

We highlight several qualitative examples on the CuRe dataset of our DIV scorer ϕ_{DIV} compared to baselines in Fig. 21. In alignment with our quantitative observations in Appendix H.1, in the examples **E2** (“Hokkien Mee”),

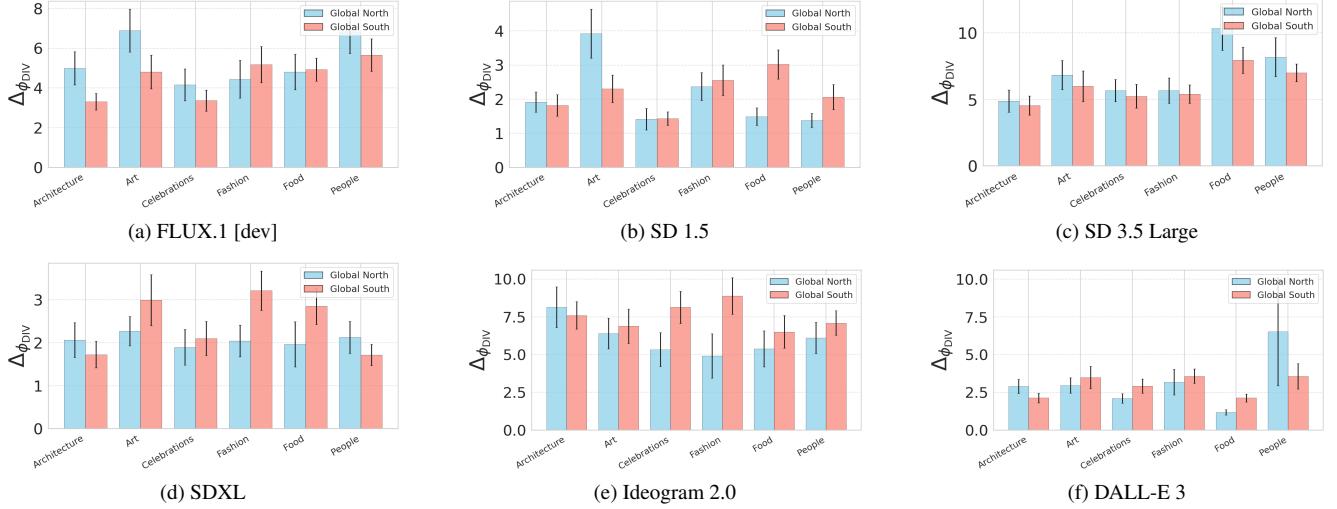


Figure 20. A bar plot of $\Delta\phi_{DIV}$ across all artifacts in the CuRe dataset aggregated at the supercategory level (mean + std error bar) for each T2I system in our evaluation set. We separate each bar at the supercategory level into artifacts belonging to the Global North (GN) and Global South (GS) to examine if there is a noticeable discrepancy between scores in these clusters.

E3 (“*Third Mainland Bridge*”), **E4** (“*Rabat Lighthouse*”) and **E5** (“*Penelope Cruz*”), all diversity scorers show an inverse relationship to human judgments of cultural representativeness ϕ_{CuRe}^* . These examples have high human scores for cultural representativeness ϕ_{CuRe}^{*} and show very little change in diversity score from $LPIPS(n)$ to ϕ_{DIV} , indicating that attribute specification does not increase diversity significantly and the artifact likely lies in the head of the T2I system distribution, which aligns with our hypothesis in Sec. 4.4. We highlight failure cases where the baseline $LPIPS(n)$ outperforms our scorer ϕ_{DIV} in **E1** (“*Ushabti*”) and **E6** (“*Vaso de los Guerrero*”), where for a low ϕ_{CuRe}^* , the baseline shows a higher diversity score and accurately matches the inverse relationship. We suspect that the factuality-diversity tax especially hurts handicrafts like pottery, where coarse semantic features may be accurate (shape, structural outline) but finer details are wrong (*e.g.* facial features and material in **E1** and material and textures in **E6**.)

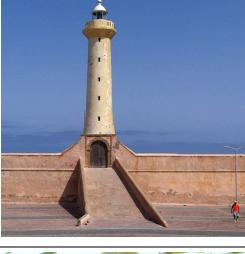
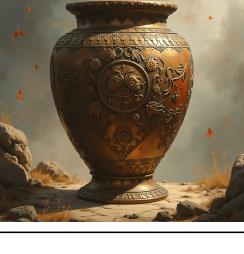
E#	AI Image	Real Image	Feedback	Score
E1			<p>Ushabti</p> <p>1. This image has used a different facial feature of the statue you would not see in my country of culture. Although, the wear and tear as well as the shape of the statue being mummy-like would be found in my country.</p> <p>2. The similarities lies in the cat structure generally. The material it is made of and the type of cloth it seems to be wearing. The part where it is different from my culture is that the eyes are looking at the side which is very unlikely. Moreover, the smile itself is so not true. The big differences lie in the eyes direction and the smile.</p>	$\phi_{CuRe}^* = 1/5$ $LPIPS(n) = 0.62$ $\phi_{DIV} = 0.48$
E2			<p>Hokkien Mee</p> <p>1. Should have less liquid sauce. May need to add shrimps. The noodle should be fried.</p> <p>2. The noodles in the AI image looks weird and too smooth/plump compared to real noodle dishes. Each individual strand can be traced which seems unlikely in a real noodle image</p>	$\phi_{CuRe}^* = 4/5$ $LPIPS(n) = 0.70$ $\phi_{DIV} = 0.59$
E3			<p>Third Mainland Bridge</p> <p>1. The image contains the distinctive part of the real location and image, it contains most of the details of the real thing.</p> <p>2. The inaccurate details is The Number of Lanes. The image shows more or fewer lanes than the actual four lanes. The image inaccurately represents the bridge structural elements such.</p>	$\phi_{CuRe}^* = 4.5/5$ $LPIPS(n) = 0.50$ $\phi_{DIV} = 0.50$
E4			<p>Rabat Lighthouse</p> <p>1. The catholic or christian cross</p> <p>2. I think gen AI did a good job here, but missed some details like the big square in front of the lighthouse and the window's sizes are somewhat big here</p>	$\phi_{CuRe}^* = 5/5$ $LPIPS(n) = 0.51$ $\phi_{DIV} = 0.54$
E5			<p>Penelope Cruz</p> <p>1. The facial features are very similar to the real ones</p> <p>2. The hair is typical of any country, so it could as well be spanish. The white clothes is typical mediterranean.</p>	$\phi_{CuRe}^* = 4/5$ $LPIPS(n) = 0.74$ $\phi_{DIV} = 0.62$
E6			<p>Vaso de los Guerrero</p> <p>1. The AI-generated image looks too unreal; it doesn't feel natural, as if it was too exaggerated</p> <p>2. The image has floral filigrees and not warriors. The decorations are in relief and not painted. In addition, the vessel is lacquered with gloss</p>	$\phi_{CuRe}^* = 2/5$ $LPIPS(n) = 0.75$ $\phi_{DIV} = 0.55$

Figure 21. Diversity scorer and user feedback

I. MLLM as a Judge

To our knowledge for the first time, we evaluate a state-of-the-art multimodal large language model (MLLM) for its ability to emulate human judgments (Likert scores) of CuRe and perceptual similarity to ground-truth and its ability to provide accurate textual justifications for its scores in Appendix I.1. We choose the recent Gemini 2.0 Flash [15], a popular yet efficient proprietary state-of-the-art MLLM. We discuss qualitative examples of Gemini 2.0 Flash scores and justifications on artifacts from the CuRe dataset and highlight several shortcomings and areas for improvement in Fig. 22b.

Gemini Prompt for Perceptual Similarity:

You will be shown two images below. The image on the left is an AI-generated image of {item_name} ({category_name}). The image on the right is a grid of four real images of {item_name} from Wikipedia. Generated Image (Left):

[Image 1 Placeholder - will be inserted by API call]
Real Images Collage (Right):

[Image 2 Placeholder - will be inserted by API call]
Please analyze the images and answer the following:
1. On a scale of 1-5, how similar do you think the generated image on the left is to the real images on the right? Use the following scale:

- 1: Not at all similar
- 2: Slightly similar
- 3: Reasonably Similar
- 4: Very Similar
- 5: Extremely Similar

2. Provide a brief explanation for your similarity rating, detailing specific visual aspects you compared.

Format your response strictly as JSON with the keys "similarity_rating" (for the number 1-5) and "similarity_explanation" (for the text explanation).

Example Response from Gemini:

```
{  
    "similarity_rating": 4,  
    "similarity_explanation": "The generated image captures the overall shape and key architectural elements seen in the real images, like the arches and towers, but lacks some finer texture details and the surrounding environment differs."  
}
```

I.1. Gemini 2.0 Flash as a Scorer

As Gemini 2.0 Flash is a multimodal LLM, it can be queried for complex and nuanced visual-question answer-

ing (VQA), such as directly querying for cultural representativeness and perceptual similarity scores. We thus replicate our user study setup and questionnaire structure (Appendix C) with Gemini.

Perceptual Similarity. Similar to our user study setup in Fig. 8 (Appendix C.2), we query Gemini for perceptual similarity (**Q1a**) over all artifacts in the CuRe dataset by showing Gemini one image generated by our T2I system, and a grid of four ground-truth images. Both images are passed in as URLs via the Google Gemini API¹¹. Mirroring our user study, we prompt Gemini to score the similarity of the T2I generated image to the ground-truth image grid from 1 ("Not at all similar") to 5 ("Extremely similar"). We also examine if Gemini can correctly highlight which details of the T2I image it thought were correct or incorrect and justify its score through text. We show the exact prompt we use below:

CuRe and Ground-Truth Likelihood. Similar to our user study setup in Fig. 10 (Appendix C.4), we ask Gemini to rate cultural representativeness (**Q3a**) and likelihood of the T2I image belonging to the class of artifact a , which we call ground-truth likelihood (**Q3b**). Ground-truth likelihood is very similar to perceptual similarity in what it is trying to measure (faithfulness / factuality), but unlike PS, Gemini must score GT likelihood without seeing any ground-truth images of artifact a . Similar to PS, we ask Gemini to justify its scores through text. We show the exact prompt below:

Prompt for CuRe and GT Likelihood:

This is an AI-generated image of {item_name}, which is a type of {category_name}. Please analyze this image and answer the following questions:

1. Cultural Representativeness: On a scale of 1-5 (where 1 is "Highly Unlikely" and 5 is "Extremely Likely"), how likely can this item be found in {country}?
2. Ground-Truth Likelihood: On a scale of 1-5 (where 1 is "Highly Unlikely" and 5 is "Extremely Likely"), how likely is this an accurate image of {item_name}?
3. Description: What specific details in the AI-generated image make it accurate or inaccurate compared to how this object typically appears in {country} or your understanding of {item_name}? Format your response as JSON with keys: 'country_likelihood', 'item_accuracy', and 'details_analysis'.

¹¹<https://ai.google.dev/gemini-api/docs/models#gemini-2.0-flash>

I.2. Analysis of Gemini 2.0 Flash Responses.

Here we discuss several illustrative examples of Gemini 2.0 Flash performance on our CuRe dataset. In Fig. 22, we show the T2I system generated image (“AI Image”) and representative ground-truth image from Wikimedia (“Real Image”) alongside scores for cultural representativeness (CuRe) from our user study ϕ_{CuRe}^* and Gemini ϕ_{Gemini}^* . To evaluate if Gemini can capture correct or incorrect cultural details highlighted by workers who identify with the culture of each cultural artifact, we also write score justifications provided by workers from Prolific (“Human Judgments”) as well as Gemini (“MLLM Judgment”).

In **E1** (*Femi Kuti*), workers unanimously highlight incorrect facial features and hairstyle while Gemini gives a high CuRe score based on general body structure and attire, and misses culturally specific identity markers. Similarly in **E3** (*Chuseok*), workers point out that many details in the image look Chinese (red lanterns, hairstyle, attire, background) and “not a single theme” resembles Korean Chuseok. Gemini’s description emphasizes the presence of traditional Korean dress and a village setting as reasons for a high CuRe score. **E4** (*Sámi Headwear*) continues this trend - while Gemini believes that colors, shape, adornments, and textile patterns are accurate to indigenous Sami culture, workers highlight that the subject appears Asian and not Northern European, and the fabric texture, color, and fur are not accurate to Finnish Sami headwear. **E6** (*Portrait of Amir Kabir*) offers a similar case: Gemini gives a high CuRe score, referencing Qajar-era styling. Workers, however, flag specific missing details like the hat, mustache style, and necklace elements that signify historical authenticity.

E2 (*Jollof Rice*) and shows a reverse failure: workers appreciate the image, noting the correct rice type, color, and ingredients. Gemini disagrees with the workers and states that while the item appears local to Nigeria, several details indicate that it may not be faithful to Jollof rice, including grain shape (pasta looking instead of long-grain), ingredients, and garnish.

For **E5** (*Takht-e Fulad*), Gemini misidentifies the architecture as Egyptian and dismisses its relevance to Iran entirely. Workers believe that while the image might belong to Iranian culture, it is missing details local to buildings from Esfahan (blue ceramic) and may more resemble other historical Iranian sites (Takht-e Jamshid). This example highlights that workers are able to latch on to these nuanced cultural differences, while Gemini’s judgment is more binary. **E7** (*Kiping*) is more straightforward: both Gemini and workers recognize that the T2I system generated image is entirely off, *i.e.* a cartoon animal instead of a food item. This shows that Gemini is able to latch on to more clear cases of T2I system failures. Lastly, in **E8** (*Pulseira*), Gemini agrees with workers that the design, and aesthetic is plausible for the Brazilian context, even if not “inher-

ently Brazilian”. Gemini docks a point for CuRe, citing the shine and stones of the bracelet may indicate a luxurious unaffordability for “a larger portion of the (Brazilian) population”, while workers point out that the “details in the shiny stones” are in fact quite accurate for Brazilian pulseira.

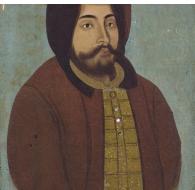
Regional Trend: A notable pattern is that most failure cases involve artifacts from the Global South (Nigeria, Iran). While Gemini 2.0 Flash performs well on globally familiar or culturally neutral items (the “head”), it struggles with region-specific details that require localized knowledge (*e.g.* mistaking Chinese culture for Korean). This likely reflects the long tail nature of the pretraining data of T2I systems, and highlights the need for more culturally balanced data curation and catering to more heterogeneous preferences to improve alignment with diverse user perspectives.

Figure 22. Visualization of Gemini 2.0 Flash responses where the MLLM either failed or succeeded in identifying nuances compared to humans. Examples were generated using FLUX.1 [dev], SD 3.5 Large, and SD 1.5. (Continued on next page)

E#	AI Image	Real Image	Feedback
E1			<p>Human Judgments (User Study):</p> <ol style="list-style-type: none"> 1. The image is someone holding a microphone but that is not femi kuti 2. The image seems inaccurate from the images of Kuti I used to see. 3. The face is not just right with me and the hair is just off...Femi has very distinctive blend of grey hair. <p>MLLM Judgment (Gemini 2.0 Flash):</p> <p>The generated image bears a strong resemblance to Femi Kuti, especially concerning his facial features, skin tone, and general build as seen in the real images. The attire and overall style are also relatively consistent with his public image. However, subtle details like the precise shape of his nose and mouth, as well as the sharpness of the image, differentiate it from the real photos. The generated image's attire is similar to what he might wear but not an exact match of anything in the real images. Considering all these aspects, the similarity is high but not perfect.</p>
E2			<p>Human Judgments (User Study):</p> <ol style="list-style-type: none"> 1. The image gets the color of Jollof rice 2. The rice and the colour of the rice after been cooked makes it look like jollof from my culture. 3. The image has used the correct kind of rice, ingredients, and sauce, in my country <p>MLLM Judgment (Gemini 2.0 Flash):</p> <p>The item can almost definitely be found in Nigeria. However, the rice itself does not resemble Jollof rice. Jollof rice typically uses long-grain rice, which, after cooking, retains a relatively distinct shape. The rice in the AI-generated image looks more like orzo pasta with an orange tint. The orange cubes included in the dish do not generally feature in Jollof rice. It usually includes tomato, pepper, onions, and spices. Also, Jollof rice is not normally garnished with a large green leafy sprig.</p>
E3			<p>Human Judgments (User Study):</p> <ol style="list-style-type: none"> 1. I feel like it's close to getting it right, but the patterns on the waist and sleeves, for example, makes it look like almost we'd never see something like this. It'd be like, let's say we are trying to find certain colored horses, but perhaps AI generates the horses with patterns of color never found in nature. 2. The hairstyle in the picture is more Chinese than Korean. Also, the detail of the traditional Korean clothing is not correct. There are also a number of red lanterns in the background, which is also more Chinese than Korean. I cannot find a single theme in the photo that is related to Korean Chuseok. 3. The figure in the image is where wrong traditional attire and the background architectures also looks like Chinese. <p>MLLM Judgment (Gemini 2.0 Flash):</p> <p>The image depicts a woman in a Hanbok, a traditional Korean dress, likely taken in an area resembling a traditional Korean village. Hanboks are commonly worn in Korea, especially during festivals like Chuseok. The likelihood of finding a Hanbok in Korea is extremely high. The accuracy as an image of Chuseok is also high. Hanboks are commonly worn during Chuseok. Some details make the image accurate, such as the Hanbok itself, which is generally accurate, and the setting, which suggests a traditional Korean village. Some minor inaccuracies include the specific style of the Hanbok, which seems a bit stylized, and it does not include other aspects related to Chuseok, such as food or activities.</p>
E4			<p>Human Judgments (User Study):</p> <ol style="list-style-type: none"> 1. The person wearing the hat doesn't look Northern European. 2. The image has used an outfit that other indigenous people from northern hemisphere might use, but based on my experience it doesn't look extremely similar to Finnish Sami headwear. The fabric texture and colors are different from what I have seen in real life. 3. The woman in the picture is asian and Sami people aren't asian. The hat is too wide and the fur is not supposed to be on top of the hat. <p>MLLM Judgment (Gemini 2.0 Flash):</p> <p>The image shows a person wearing Sami headwear, specifically a hat with intricate patterns and a possible fur trim at the top. The colors are primarily red, white, and green, which are commonly found in Sami designs. The overall shape of the hat, along with the additional adornments like the colorful braids and beads around the neck, matches the traditional appearance of Sami attire. The designs also appear to be consistent with Sami textile patterns, which are often geometric and symbolic. Given that the Sami people inhabit areas within Finland, it is extremely likely that this type of headwear can be found there. The overall image appears to be an accurate representation of Sami headwear.</p>

(a) Examples 1–4 of Gemini 2.0 Flash failure cases (Examples 1–3 generated with FLUX.1 [dev], and Example 4 with SD 3.5 Large).

Figure 22. Continued from previous page.

E#	AI Image	Real Image	Feedback
E5			<p>Human Judgments (User Study):</p> <ol style="list-style-type: none"> so the image has certain specification like carvings are quite similar, the stone types also similar to the one I saw in my country, however I think this image is closer to Takhte-Jamshid, as it does not have blue ceramic which usually be used in Esfahan an is one of the icons of any munument in Esfahan. The coloring is different and Isfahanian buildings have different sets of colors and style more colorful. The image is so similar to the type of historical buildings in Iran. <p>MLLM Judgment (Gemini 2.0 Flash):</p> <p>The image shows a large stone structure with hieroglyphs, reminiscent of ancient Egyptian architecture. Takht-e Fulad, in contrast, is a historical cemetery in Isfahan, Iran, which primarily consists of tombstones, mausoleums, and intricate stonework with Persian inscriptions, not large temples. Therefore, it is highly unlikely that this image depicts something found in Iran or an accurate representation of Takht-e Fulad.</p>
E6			<p>Human Judgments (User Study):</p> <ol style="list-style-type: none"> The only thing similar to Amir Kabir AI to the real life is the facial structures like nose and eyes. The beard needs improvement as each era in Iranian culture, men used to have different hair and beard styles. The cone shaped hat is important in displaying who this person was and the power he had during Qajar in Iran. In addition, the texture of the clothes and the famous necklace are items that AI "must" add to the image made since they are distinct elements that show Persian culture especially for showing a historical figure. It has the right feature in the face but lack of detail in the outfit and background. The colour of his mustaches should be darker, and its style is not Iranian and looks more like Mongols and Turks. <p>MLLM Judgment (Gemini 2.0 Flash):</p> <p>A portrait of Amir Kabir is highly likely to be found in Iran due to his historical importance. The image appears to be an accurate representation, showing him in traditional attire common for the Qajar era. The details like the style of beard, the coat, and the head covering are consistent with known depictions of him. However, judging finer accuracy requires comparison with specific known portraits, but generally, the visual style and features align with expectations.</p>
E7			<p>Human Judgments (User Study):</p> <ol style="list-style-type: none"> The image does not portray the meaning of Kiping at all. The image shown was a type of an animal which is obviously so far from "Kiping" which is a type of food. The image is an animal holding a spray bottle, it has no connection to kiping. <p>MLLM Judgment (Gemini 2.0 Flash):</p> <p>The image features a cartoon cat holding what appears to be a small pump or sprayer. There is absolutely nothing in the image resembling Kiping, which is a type of leaf-shaped, colorful wafer made from glutinous rice. Therefore, it's highly unlikely this image depicts Kiping or could be found in the Philippines as a depiction of Kiping. The AI hallucinated a completely different scene.</p>
E8			<p>Human Judgments (User Study):</p> <ol style="list-style-type: none"> This is exactly how a pulseira is, a piece of jewellery that goes on the wrist. The image is very similar to the name of Pulseira. This bracelet represents one of the forms of objects that we wear on our wrist, very beautiful with details in shiny stones. <p>MLLM Judgment (Gemini 2.0 Flash):</p> <p>The image depicts a bracelet with a braided or woven metal band and a spherical centerpiece covered in small stones (possibly diamonds or crystals). The clasp appears to be a cylindrical metal piece. Jewelry of this style, particularly silver-tone metal and crystal/diamond combinations, is widely available in Brazil. Brazilians enjoy both fine jewelry and more affordable fashion jewelry. The design isn't inherently 'Brazilian' but the style is commonly found. It could be sold in upscale stores or at street markets. The image accurately represents what a "pulseira" (Portuguese for bracelet) can look like. While the term encompasses a wide variety of bracelet styles, this particular design is plausible and well-executed. The braided metal design is common in bracelets. The overall aesthetic is refined and somewhat luxurious. The clasp type is also reasonable. **Inaccuracy:** There is little that points to this image being highly 'Brazilian', as opposed to generally stylish. The price point may make it difficult for a larger portion of the population to afford. If the center sphere is made of real diamonds, it could be less common than if it were made of crystals. It looks very shiny. The dark and reflective background gives it a commercial aesthetic, which may not immediately invoke Brazil, but is still considered a fairly accurate representation of a pulseira.</p>

(b) Examples 5–6 of Gemini 2.0 Flash failure cases (Example 5: SD 3.5 Large, Example 6: SD 1.5), and Examples 7–8 of successful cases (Example 7: FLUX.1 [dev], Example 8: SD 3.5 Large)