

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

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

¹University of Wisconsin-Madison, ²Google Deepmind

Abstract

Popular text-to-image (T2I) models 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 text-to-image systems as a proxy for human judgments. Our CuRe dataset has a novel categorical hierarchy built from the crowdsourced Wikimedia knowledge graph that enables benchmarking T2I systems in this manner. Our benchmark has 300 cultural artifacts across 32 cultural subcategories grouped into six broad cultural axes (food, art, fashion, architecture, celebrations, and people). Unlike existing benchmarks which suffer from “generative entanglement” due to overlap of T2I system and evaluation scorer training data, CuRe enables 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 Dalle-3.

1. Introduction

Text-to-Image (T2I) systems [44, 47, 49, 52] are trained on web-scale data [9, 20, 54, 55], which is long-tail in nature [43]. This translates to hallucinations when generating 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 [4, 10], including harmful and offensive content [6, 7]. We show a simple example of T2I system bias using a state-of-the-art T2I system [39] to gen-



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.

erate 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 measurement, which is non-trivial. A typical method to measure bias is through large-scale user studies on crowdsourcing platforms [25, 28], 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

*Corresponding author: aniketr@cs.wisc.edu

†Equal Advising

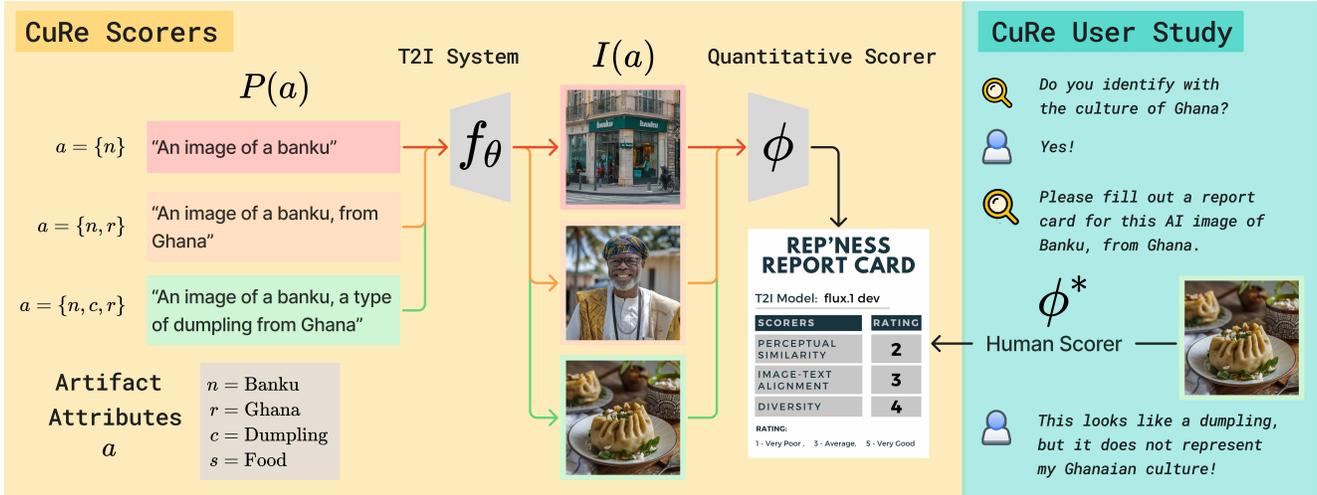


Figure 2. An overview of the scorers and user study of CuRe. Through the lens of marginal utility of attributes, we analyze images $I(a)$ generated from prompts templates $P(a)$ over attributes of the 300 cultural artifacts in our benchmark dataset. For example, we generate an image from name and region attributes as $I(a) = f_{\theta}(P(a|a = \{n, r\}))$. This is the middle image in the figure above. These images are then scored by quantitative scorers ϕ (see Sec. 2) and user judgments from a large-scale user study ϕ^* (see Appendix D.1). Crucially, cultural artifacts from region r are rated only by workers who identify with the culture of region r .

human judgments, such as computing similarity of generated images to real images with large neural encoders [29], realism metrics [5, 23], alignment of images to some desired attributes specified through text [29, 63] and cultural diversity [28]. 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. 3).

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. 2), which correlates better to human judgments than proxy scorers across T2I models (Sec. 3). We compare our dataset and metric design to prior work in Tab. 4 and summarize our contributions below:

- A **new dataset** constructed in a scalable fashion directly from the large crowdsourced Wikimedia knowledge graph [66] with a **novel coarse-to-fine categorical hierarchy** of 300 cultural artifacts across six cultural

axes, 32 cultural categories and 64 countries (see Fig. 3).

- 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. 3).
- For the first time, we evaluate the cultural capabilities of a natively multimodal large language model (Appendix J).

2. Finding a CuRe through Information

We introduce our novel lens of marginal utility of information into three prominent existing classes of quantitative scorers: measuring human **perceptual similarity** between the generated image and the ground truth (Sec. 2.1), evaluating **image-text alignment** with a desired attribute specified through text as a proxy for visual question-answering (Sec. 2.2), and measuring the **diversity** of the T2I system across global cultures (Sec. 2.3). We set up some useful notation in Appendix E.

Towards designing a more accurate, reliable scorer for

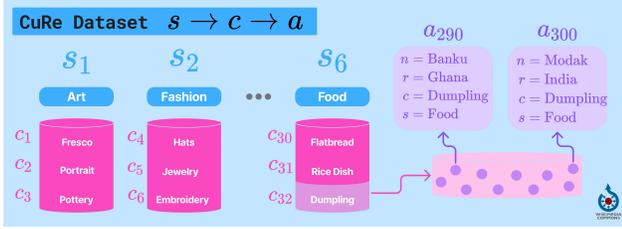


Figure 3. The CuRe dataset is constructed in a hierarchical fashion ($s \rightarrow c \rightarrow a$) from Wikimedia, with 6 cultural axes s , 32 cultural categories c , and 300 cultural artifacts a . For example, $s = \text{“Food”}$, $c = \text{“Dumpling”}$, $a = \text{“Banku”}$.

CuRe across global cultures, we ask: *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. 2. 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. 9). 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 much information must be *explicitly provided* to a T2I system reveals valuable insights into its cultural representativeness capabilities.

Unlike existing methods, this approach measures how well a model internalizes cultural knowledge by analyzing the **marginal utility of each additional attribute** specified during generation. For example, with the “Banku” artifact in Fig. 2, we need to marginally increase attributes specified to the T2I system via the prompt $n \rightarrow n, r \rightarrow n, c, r$ 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-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. 9. We demonstrate this behavior quantitatively over the entire

CuRe dataset in Sec. 3. We formally define the variant of MIA scorer for each class of proxy scorer in the relevant section below and explain further in Appendix E.

2.1. Perceptual Similarity Scorers

The goal of perceptual similarity (PS) scorers is to compute the similarity between a generated image $I(a)$ and another set of images representing the same artifact a , which are typically manually-curated ground-truth (GT) images $G(a)$. 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 $\phi_{PS}^*(a) = \text{Likert}(I(a), G(a))$, where a Likert score [33] 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 [16, 22], which we call $\phi_{GT} = \text{sim}(I(a), G(a))$. The proxy scorers often ignore when two images are semantically similar in favor of spatial and textural consistency [21], which causes a discrepancy between real human perceptual similarity ratings (see Fig. 9a).

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 (?? for $n = \text{“pierogi”}$). If these images are very different (?? 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 G , $\phi_{PS}(n)$ instead computes similarity of T2I generations of banku $I(n)$ to T2I generations of dumplings $I(c)$:

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

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

Following the status quo for semantic similarity [26], sim is a cosine distance between embeddings from state-of-the-art large vision encoders (SigLIP 2 [61], AIMV2 [18], and DINOv2 [42]).

2.2. 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 [46] popularized using textual descriptions of ImageNet classes as zero-shot labels for image

classification via unsupervised contrastive learning. We denote this class of image-text scorers by $sim(I(n), P(a))$. For example, to evaluate country-specific representativeness, Khanuja et al. [29] check CLIP similarity of $I(n)$ with $P(r) = \text{“This image is culturally relevant to } \{r\}\text{”}$. We note that prior works typically use CLIP trained on LAION-2B [54] as their VLM of choice for image-text alignment, which has been shown to have an Amero and Euro-centric bias [2, 37] and is also part of the pretraining datasets of many popular T2I systems [27, 44, 49, 52, 70]. This overlap causes misleading over-estimations of quality, which we call **generative entanglement** (see Sec. 3.3 and Tab. 15 in Appendix H).

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 [50]. We show in Tab. 2 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:

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

2.3. Diversity Scorers

The goal of diversity (**DIV**) scorers is to capture the heterogeneity or intra-class variance of images generated by T2I systems [28]. 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?

LPIPS [73] computes an average over pair-wise dissimilarity of images across all cultural artifacts a associated with category c using deep features extracted from convolutional neural networks. Similar to Perceptual Similarity scorers (Sec. 2.1), LPIPS (which uses encoders trained on ImageNet [51]) can ignore image semantics in favor of spatial, color, and textural consistency [21]. Another

recent approach to scoring diversity is via Vendi Scores (VS) [19, 72], 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. Kanen et al. [28] propose quality-weighting the Vendi score (QVS) with a human preference reward model [68] to address this limitation. We demonstrate qualitatively (Fig. 9c) and quantitatively (Tab. 3) that LPIPS, VS, and QVS do not correlate well with human judgments of CuRe.

We propose a modification to LPIPS that captures marginal information attribution. 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:

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

3. Experiments

We discuss our empirical setup and observations for each class of scorer below, *i.e.* Perceptual Similarity scorers ϕ_{PS} in Sec. 3.2, Image-Text Alignment scorers ϕ_{ITA} in Sec. 3.3, and Diversity scorers ϕ_{DIV} in Sec. 3.4. We provide details on T2I inference and seeding in Appendix B, discuss user study design in Appendix D.1, and discuss evaluating the capability of our scorers to approximate human judgments in Sec. 3.1. Lastly, we benchmark state-of-the-art T2I systems on our dataset and scorer suite in Appendix E.1.

3.1. 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. 1, 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. 3, 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. 9 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.

3.2. Perceptual Similarity

To compare our proposed scorers to human judgments of cultural representativeness $\phi_{\text{CuRe}}^*(a)$, perceptual similarity $\phi_{PS}^*(a)$ and ground-truth label likelihood $\phi_{GT}^*(a)$, we tabulate a Spearman rank correlation in Tab. 1. In our experiments, we set $a = \{n\}$ for $\phi_{PS}(a)$ in Eq. (2), to compare how perceptually similar images of cultural artifacts are to generated images of their cultural category, *e.g.* $I(n)$ (images of “banku”) and $I(c)$ (images of “dumpling”). To evaluate how much the *change* in attribute specification ($n \rightarrow a$) affects perceptual similarity, we also compare the divergence Δ of $\phi_{PS}(a)$ from $\phi_{PS}(n)$ across subsets of attributes a :

$$\begin{aligned}\phi_{GT}(n) &= \text{sim}(I(n), G(n)) \\ \phi_{PS}(n) &= \text{sim}(I(n), I(c)) \\ \Delta\phi_{PS}(a) &= \text{sim}(I(a), I(c)) - \text{sim}(I(n), I(c))\end{aligned}$$

Intuitively, we find that across all image encoders, all quantitative PS scorers 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. We find that our $\phi_{PS}(n)$ scorer, which uses no ground-truth information, is comparable to strong baseline $\phi_{GT}(n)$ in Spearman’s ρ with the gold scores ϕ_{GT}^* for Stable Diffusion models across all image encoders (Tab. 1). On FLUX.1 [dev], 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. We also evaluate Gemini 2.0 Flash [13], a strong natively multimodal large language model (see Appendix J). We query Gemini to score perceptual similarity between $I(n)$ and ground truth images $G(n)$ on a 1 to 5 scale, similar to ϕ_{PS}^* from the user study. We observe that Gemini correlates almost identically

Table 1. Spearman rank correlation 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.

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	-	-	0.40	-	-	0.39	-	-	0.40
SigLIP 2 [61]	$\phi_{GT}(n)$	0.25	0.36	0.44	0.26	0.29	0.46	0.23	0.33	0.48
	$\phi_{PS}(n)$	0.18	0.25	0.32	0.17	0.32	0.42	0.21	0.27	0.30
	$\Delta\phi_{PS}(n, c)$	-0.16	-0.27	-0.31	-0.19	-0.30	-0.41	-0.05	-0.04	0.08
	$\Delta\phi_{PS}(n, c, r)$	-0.17	-0.25	-0.30	-0.19	-0.31	-0.42	-0.06	-0.02	0.05
AIMV2 [18]	$\phi_{GT}(n)$	0.20	0.28	0.39	0.27	0.31	0.46	0.19	0.29	0.42
	$\phi_{PS}(n)$	0.08	0.17	0.25	0.17	0.25	0.34	0.19	0.18	0.16
	$\Delta\phi_{PS}(n, c)$	-0.17	-0.30	-0.35	-0.22	-0.29	-0.34	0.00	-0.05	0.05
	$\Delta\phi_{PS}(n, c, r)$	-0.18	-0.28	-0.32	-0.20	-0.28	-0.35	0.03	0.02	0.07
DINOv2 [42]	$\phi_{GT}(n)$	0.16	0.29	0.40	0.28	0.34	0.46	0.26	0.34	0.51
	$\phi_{PS}(n)$	0.13	0.22	0.30	0.24	0.37	0.44	0.24	0.32	0.36
	$\Delta\phi_{PS}(n, c)$	-0.19	-0.27	-0.32	-0.17	-0.28	-0.38	-0.14	-0.08	0.03
	$\Delta\phi_{PS}(n, c, r)$	-0.19	-0.25	-0.31	-0.16	-0.27	-0.37	-0.06	0.03	0.00

well with ϕ_{PS}^* across T2I systems, and slightly lags behind $\phi_{GT}(n)$ in Spearman’s ρ . We highlight that while $\phi_{GT}(a)$ 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 [18, 42, 61]. We also highlight that our divergence scorers $\Delta\phi_{PS}$ degrade in rank correlation for SD 1.5, an older lower-resolution T2I system trained on much less data while the baseline scorer $\phi_{GT}(n)$ does not. This suggests that our scorer may be more sensitive to the coverage of the pretraining data or the quality of the base T2I system.

3.3. 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. 2. To evaluate the **generative entanglement** (impact of training data overlap) between T2I system and scorer, we use a suite of vision-language models (VLMs) with different pretraining datasets as our similarity measure: four versions of OpenCLIP with different pretraining datasets, SigLIP 2 (details in Appendix H). We compare our scorer to several baselines $\text{sim}(I(n), P(a))$: two prior works evaluating cultural relevance [29] and cultural identity [63], the best performing prompt $P(r)$ suggested by o3-mini [41], Gemini 2.0 Flash, and several state-of-the-art human preference scorers [30, 68, 69]. We also evaluate image alignment to specific artifact attribute specified through text varying $P(a) = P(n), P(c), P(r)$, and $P(c, r)$.

Table 2. Spearman rank correlation 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 in Tab. below) compute a cosine distance with SigLIP 2 embeddings.

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 [30]	0.20	0.29	0.34	0.23	0.37	0.40	0.23	0.39	0.45
Imagereward [69]	0.19	0.21	0.26	0.16	0.24	0.30	0.23	0.30	0.35
HPS v2 [68]	0.23	0.29	0.33	0.18	0.35	0.37	0.24	0.40	0.43
Gemini 2.0 Flash	0.23	0.41	-	0.27	0.37	-	0.18	0.39	-
Khanuja et al. [29]	0.13	0.08	0.11	0.05	0.04	0.00	0.06	-0.02	-0.04
Ventura et al. [63]	0.19	0.15	0.14	0.10	0.07	0.05	0.11	0.02	0.00
o3-mini [41]	0.17	0.13	0.14	0.07	0.03	0.00	0.06	-0.01	-0.04
$\text{sim}(I(n), P(n))$	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	-0.05
$\text{sim}(I(n), P(c, r))$	0.25	0.35	0.38	0.22	0.31	0.35	0.25	0.34	0.37
$\phi_{ITA}(c)$	0.24	0.37	0.40	0.20	0.34	0.40	0.22	0.37	0.44
$\phi_{ITA}(r)$	0.28	0.33	0.35	0.20	0.27	0.29	0.23	0.30	0.39
$\phi_{ITA}(c, r)$	0.27	0.38	0.42	0.22	0.34	0.39	0.24	0.36	0.44

Our evaluation reveals that comparing images to $P(n)$, which uses only the artifact’s name (e.g., “sushi”) in the prompt, consistently has a stronger positive rank correlation with human judgments than all methods that compare to $P(r)$ across all T2I systems (Tab. 2). This supports our claim in Sec. 2.2 that explicitly querying VLMs for image similarity to prompts describing the image ($\text{sim}(I(n), P(n))$) is more effective at assessing region-specific cultural relevance than directly querying for cultural relevance with region-specific prompts $P(r)$. Replacing name-specific prompts $P(n)$ with category-specific prompts $P(c)$ (e.g. “food” for an image of sushi), 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 outperform name-specific prompts.

Our scorers 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) accurately generate textual descriptions of the artifact ($\text{sim}(I(n), P(n))$), b) capture 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, and is slightly outperformed by $\phi_{ITA}(r)$ on Stable Diffusion models for ϕ_{PS}^* and ϕ_{GT}^* . $\phi_{ITA}(r)$ is worse than the best-performing baseline across T2I systems, indicating that adding categorical information ($n \rightarrow c$) is important to capture region-specific human judgments. Our scorers also outperform or match strong T2I human preference reward models trained on hundreds of thousands of human-labeled preferences across all gold scores ($\phi_{ITA} \geq \text{PickScore} > \text{HPS v2} > \text{Imagereward}$ in rank correlation).

Lastly, we discuss the impact of our choice of VLM as a similarity measure in the context of **generative entanglement**. In Tab. 15 (Appendix H), 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 pretraining 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.

3.4. Diversity

We show a Spearman’s ρ 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. 3. 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 high-quality models (FLUX.1 [dev] and SD 3.5 Large).

Table 3. Spearman rank correlation between diversity (DIV) scorers and user judgments ϕ_{CuRe}^* , ϕ_{PS}^* , and ϕ_{GT}^* across T2I systems on the CuRe dataset.

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	-0.03	-0.02
qVS(c)	-0.02	-0.01	-0.01	-0.03	0.02	0.05	-0.01	-0.03	-0.02
$\phi_{DIV}(a)$	-0.20	-0.23	-0.30	-0.22	-0.23	-0.29	-0.08	-0.11	-0.11

As the Vendi score is computed as an aggregate at a coarse category level on static assignments over cultural artifacts and has no sense of image quality (see Appendix I 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]), (Tab. 3). Even the more recent quality-weighted Vendi Score [28] fails to resolve these limitations, as it simply scales VS(c) by a scalar quality measure q averaged over all artifacts n associated with category c (HPS v2 score [68]), leaving the rank correlation with human judgments unchanged.

In contrast, our MIA-based scorer $\phi_{DIV}(a)$ 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]). This result indicates that with our scorer, if incrementally increasing attribute specification decreases diversity (*i.e.* the cultural artifact is homogenous across attribute specification and likely lies in the head of the T2I distribution), cultural representativeness and perceptual similarity as perceived by humans tends to increase. This aligns with our hypotheses in Sec. 2.3 and indicates the potential of marginal utility of information specification as a measure of diverse cultural representativeness.

4. 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.

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. 3) outperform or match all baselines, including a strong multimodal large language model (Gemini 2.0 Flash, Appendix J). 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. 2).

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. 11
- [2] Abhipsa Basu, R Venkatesh Babu, and Danish Pruthi. Inspecting the geographical representativeness of images from text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5136–5147, 2023. 4, 10, 11, 17
- [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. 11, 17
- [4] 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, 10
- [5] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018. 2, 10
- [6] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021. 1, 10
- [7] 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
- [8] 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. 11, 30
- [9] 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
- [10] 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
- [11] Carole Counihan, Penny Van Esterik, et al. *Food and culture*. Routledge New York, NY, 2013. 12
- [12] 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. 10
- [13] Google DeepMind. gemini, 2025. Accessed: March 5, 2025. 5, 11, 30
- [14] Don A Dillman, Jolene D Smyth, and Leah Melani Christian. Internet, phone, mail, and mixed-mode surveys: The tailored design method. *Indianapolis, Indiana*, 2014. 17
- [15] 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. 10
- [16] 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. 3
- [17] 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. 26
- [18] Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor Guilherme Turrissi da Costa, Louis Béthune, Zhe Gan, et al. Multimodal autoregressive pre-training of large vision encoders. *arXiv preprint arXiv:2411.14402*, 2024. 3, 5
- [19] Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning, 2023. 4, 28
- [20] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruva Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023. 1, 26
- [21] 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. 3, 4
- [22] 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. 3
- [23] 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, 10
- [24] Geert Hofstede. *Culture’s consequences: Comparing values, behaviors, institutions and organizations across nations*. Sage publications, 2001. 12
- [25] 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. 1, 11

- [26] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 3
- [27] 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. 4, 10
- [28] Nithish Kannan, 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 *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. 1, 2, 4, 6, 10, 11, 12, 17, 28
- [29] 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, 4, 5, 10, 11, 12, 17, 26, 27
- [30] 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. 5
- [31] 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. 10
- [32] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. *Research methods in human-computer interaction*. Morgan Kaufmann, 2017. 17
- [33] Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932. 3, 17
- [34] 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. 10
- [35] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 11
- [36] Zhixuan Liu, Youeun 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. 10, 11, 12, 17
- [37] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408*, 2023. 4, 10
- [38] Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens. *arXiv preprint arXiv:2304.06034*, 2023. 10
- [39] OpenAI. Dall-e 3, 2023. Accessed: March 5, 2025. 1
- [40] OpenAI. Openai o1, 2024. Accessed: March 5, 2025. 30
- [41] OpenAI. o3-mini, 2025. Accessed: March 5, 2025. 5, 26
- [42] Maxime Oquab, Timothée Darcet, 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. 3, 5, 17
- [43] 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, 10, 28, 35
- [44] 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, 4
- [45] Prolific. Prolific, 2014. Accessed: March 5, 2025. 10
- [46] 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, 11, 26
- [47] 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, 10
- [48] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*, 2022. 10
- [49] 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, 4, 10
- [50] 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. 4, 10, 11, 12
- [51] 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. 4, 10
- [52] 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, 4, 10
- [53] 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. 10
- [54] 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, 4
- [55] 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, 10, 11, 26, 35
- [56] 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. 10
- [57] 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. 10
- [58] 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. 10
- [59] 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. 11
- [60] UN Trade and Development. All groups compositions. https://unctadstat.unctad.org/EN/Classifications/DimCountries_All_Hierarchy.pdf, 2025. Accessed: 2025-03-21. 13
- [61] 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. 3, 5, 17
- [62] Victoria Turk. How ai reduces the world to stereotypes. <https://restofworld.org/2023/ai-image-stereotypes/>, 2023. Accessed: 2024-09-07. 10
- [63] 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, 5, 10, 11, 17, 26, 27
- [64] 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. 10
- [65] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiyang Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 11
- [66] Wikimedia. Wikimedia Commons, 2004. Accessed: 2025-03-06. 2, 11
- [67] 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. 10
- [68] 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. 4, 5, 6
- [69] 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. 5
- [70] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gungjan 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. 4, 10
- [71] 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. 11
- [72] Lili Zhang, Xi Liao, Zaijia 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. 4, 10, 11, 12, 17
- [73] 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. 4, 28

Contents

1. Introduction	1
2. Finding a CuRe through Information	2
2.1. Perceptual Similarity Scorers	3
2.2. Image-Text Alignment Scorers	3
2.3. Diversity Scorers	4
3. Experiments	4
3.1. Scorer Correlation to Human Judgments	4
3.2. Perceptual Similarity	5
3.3. Image-Text Alignment	5
3.4. Diversity	6
4. Conclusion	6
A Related Work	10
B T2I Inference Details	11
B.1. Seeding	11
B.2. Safety Filter Refusal	11
C Dataset Design	12
D User Study Design	13
D.1. User Study Setup	13
D.2. Disclosure	13
D.3. Perceptual Similarity	14
D.4. Artifact Familiarity Questionnaire	14
D.5. CuRe: Cultural Representativeness	14
D.6. Offensiveness and Stereotypes	15
D.7. User Metadata	15
E Scorers	16
E.1. Benchmark Results	16
F. User Study Analytics	17
F.1. Comparison to prior User Studies	17
F.2. Inter-Annotator and Encoder Agreement	17
F.3. Survey Respondent Statistics	18
G Perceptual Similarity Ablations	22
G.1. Head-Tail Split using PS Scorers	22
G.2. Qualitative Analysis of PS Scorers	24
H Image-Text Alignment	26
H.1. Choice of ITA Scorer’s VLM Backbone	26
H.2. Qualitative Analysis of ITA Scorers	26
I. Diversity	28
I.1. Diversity as a Long Tail Predictor	28
I.2. Qualitative Analysis of DIV scorers	28

J. MLLM as a Judge	30
J.1. Gemini 2.0 Flash as a Scorer	30
J.2. o1 as a Judge	31
J.3. Analysis of Gemini 2.0 Flash Responses.	32

K Concept Frequency Estimation	35
---------------------------------------	-----------

A. Related Work

Dataset Biases Datasets used to train generative models are known to have biases, including the *geographical distribution* biasworks [12, 57] of crowd-labeled datasets such as ImageNet [51], Open Images [31] and MS COCO [34], which have been shown to have a strong Americentric and Eurocentric bias. These works also highlight the *data collection bias*; e.g., on YFCC100m [58], the data from underrepresented countries is often taken by tourists (~47% [64]), and does not capture the true local distribution of objects, people, and language. These dataset biases propagate towards the T2I systems.

Text-to-Image System Biases T2I systems are predominantly built on diffusion models [49, 52], autoregressive Transformer models [15, 47, 70] or GANs [27, 53]. The pretraining datasets of these models are web-scale [55] and long-tail [43], and there have been several recent works examining the biases present within T2I systems. Several recent works analyze the biases of text-to-image models from a geographical [2], gender occupation [56], cultural [4, 62] and social [37, 38] perspective. [37] provides an excellent overview of TTI system biases, including from the data collection [6] and filtering [48] process, as well as model training [67].

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 [29], which contain similar dataset biases as T2I models themselves [56]. Another class of scorers use realism metrics [5, 23], which ignore culture-specific nuance. Lastly, prior works evaluate image-text alignment to carefully chosen prompts [29, 63] and cultural diversity as a proxy for representativeness [28], which often empirically do not correlate well human judgments across global cultures (see Sec. 3).

Cultural Benchmarks To create cultural benchmarks to measure T2I biases, previous works typically crowdsource data, either directly from workers on online platforms [29, 45] or with massive inter-organization efforts [50]. Some works also rely on cultural experts to create the data [36, 72]. 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

Table 4. 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, $|\mathcal{C}|$ the number of cultural categories, and $|\mathcal{A}|$ 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{A} $	New Metric	Img-Txt Sim	Img-Img Sim	Diversity	In User Study	Only Quant
Liu et al. [36]	✓			8	9	-	1095					1	1
Basu et al. [2]				27	-	10	-		✓	✓		2	2
Ventura et al. [63]				10	8	200	-	✓	✓	✓		3	6
Jha et al. [25]				135	-	-	-	✓	✓			1	1
Khanuja et al. [29]	✓			7	-	17	580		✓	✓		2	2
Kannen et al. [28]	✓	✓		8	3	-	1000	✓			✓	2	2
Zhang et al. [72]				10	9	-	595	✓		✓		3	3
Bayramli et al. [3]	✓	✓		10	3	-	150		✓			3	3
CuRe (Ours)	✓	✓	✓	64	6	32	300	✓	✓	✓	✓	3	6

time. To measure cultural biases, prior works rely on querying users across global cultures for their judgment on images generated by T2I systems [2, 28, 29]. We overcome these limitations by constructing our benchmark CuRe directly from the live Wikimedia graph [66], 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 [13, 35, 59, 65, 71]. 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 [46, 55] and perform very well on complex visual question answering benchmarks [1, 50]. They can thus be directly queried for culture-specific knowledge, similar to existing works that use MLLMs as a judge [8]. Even at this scale of pretraining data, we show that MLLMs still fall short at evaluating cultural representativeness (Appendix J).

B. T2I Inference Details

B.1. Seeding

We use four of the prompt styles outlined in Tab. 5 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 seeds¹. For

¹DALLE-3 has no dedicated random seed parameter. We follow prior work on passing random seeds to DALLE-3 via prompting: <https://community.openai.com/t/consistent-variability-using-seeding-with-dall-e-3/457823>

Table 5. Text prompts P describing cultural artifact a given to generative model f_θ with differing levels of informativeness indicated by artifact attributes $X = \{n, c, s, r\}$.

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

all models 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. 2.1 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,

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

B.2. Safety Filter Refusal

Due to the inbuilt safety filters of Dalle-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. 6 shows the percentage of each supercategory that was successfully generated, calculated as:

[//community.openai.com/t/consistent-variability-using-seeding-with-dall-e-3/457823](https://community.openai.com/t/consistent-variability-using-seeding-with-dall-e-3/457823)

Table 6. Percent of each supercategory that was generated by Dalle-3 and SD 1.5.

Supercategory	Dalle-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)

$$\begin{aligned}
 &= \frac{N_{img_gen}(s)}{N_{img_total}(s)} * 100 \\
 &= \frac{N_{img_gen}(s)}{N_{artifact}(s) * N_{promptstyle} * N_{seeds}}
 \end{aligned}$$

where $N_{artifact}(s) = 50$ for all supercategories, $N_{promptstyle} = 4$, and N_{seeds} depends on the T2I system (see Appendix B.1 for details).

C. Dataset Design

Creating a high quality cultural benchmark is non-trivial for two primary reasons: a) it requires significant crowdsourcing efforts [28, 29, 50] or hiring domain experts [36, 72] b) it requires good “cultural coverage”, *i.e.* collecting cultural artifacts across a large number of cultural regions r and cultural categories c . 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 Wikimedia. We address the second difficulty by collecting cultural artifacts across 64 countries, which is higher than all existing cultural benchmarks (Tab. 4). We will fully open-source our code to easily add new data to our benchmark.

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) [24]. 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 [11], 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

Table 7. Supercategories and corresponding categories for our CuRe benchmark dataset.

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

measured by Vendi scores in Sec. 3.4). 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, 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 is made up of a set of 300 cultural artifacts organized in a categorical hierarchy, denoted by \mathcal{A} . These artifacts are partitioned into six cultural axes (or supercategories) \mathcal{S} , *i.e.* architecture, art, celebrations, fashion, food,

people. We describe each cultural artifact $a \in \mathcal{A}$ by a set 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). To construct the CuRe dataset, within each supercategory, we search for Wikimedia categories structured as “[category name] by country”. 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. 2.1) and our user study (Appendix D.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 [60], which we decided based on UNCTAD categorization as developed economies (Global North) or developing economies (Global South). We show an illustration of the dataset structure in Fig. 3 and tabulate the entire categorical hierarchy of the CuRe dataset in Tab. 7.

D. User Study Design

D.1. User Study Setup

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 specific to their region. We ask each worker from region r to rate on a 1-5 Likert scale a generated image of artifact a for

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 $\{a\}$ ”

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 a . We discuss the UI, survey design choices, and survey questions asked to workers in Appendix D. We also compare our study

²<https://www.prolific.com/>

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 Slightly similar Reasonably Similar Very Similar Extremely Similar

Answer:

Figure 4. **Q1a:** Querying users for perceptual similarity of T2I system generated image to ground-truth images.

to prior works in Tab. 9 (Appendix F). We pay workers the platform set minimum of \$8 per hour.

For each artifact a 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 D.6) 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 D.2)
2. Perceptual Similarity (Appendix D.1)
3. Artifact Familiarity Questionnaire (Appendix D.4)
4. CuRe: Cultural Representativeness (Appendix D.5)
5. Offensiveness and Stereotypes (Appendix D.6)
6. User Metadata (Appendix D.7)

D.2. 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 5. **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’.

D.3. 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 B.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. 4).

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. 5. We analyze worker disagreement over perceptual similarity in Appendix F.2.

D.4. 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”

D.5. 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. 6.

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),



	1. Highly Unlikely	2. Slightly Likely	3. Somewhat Likely	4. Quite Likely	5. Extremely Likely
How likely can the item in this image be found in your country?	<input type="radio"/>				
How likely is this an image of spaghetti with meatballs?	<input type="radio"/>				

Figure 6. 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 *a*, 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 7. Description and examples given about stereotypes.

D.6. 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. 7 for details), as we observed a vast gulf in understanding of what “stereotypical” means in the context of T2I systems during our pilot study.

D.7. 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. 8). 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)?

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

Figure 8. Demographic information questions.

Table 8. We evaluate several state-of-the-art T2I systems on our CuRe benchmark across our three scorer classes: perceptual similarity (Sec. 2.1), image-text alignment (Sec. 2.2), and diversity (Sec. 2.3). The best entry of each column is bolded, and the next best if the T2I system is Dall-E 3 or SD 1.5*

T2I System	$\phi_{PS} \downarrow$			$\phi_{ITA} \uparrow$			$\phi_{DIV} \uparrow$
	SigLIP 2	DINOv2	AIMV2	SigLIP 2	LAION-2B	WIT	AlexNet
FLUX.1 [dev]	0.061	0.075	0.024	0.094	0.218	0.209	0.708
SD 3.5 Large	0.067	0.104	0.019	0.115	0.251	0.225	0.670
SD 1.5*	0.059	0.076	0.019*	0.107	0.240	0.229	0.755*
SDXL	0.057	0.079	0.020	0.113	0.255	0.230	0.753
Ideogram 2.0	-	-	-	0.096	0.214	0.195	0.693
Dall-E 3*	0.057*	0.073*	0.022	0.105	0.219	0.222	0.789*

* Dall-E 3 and Stable Diffusion 1.5 have moderate refusal rates due to safety filters (see Appendix B.2)

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.”

E. Scorers

E.1. Benchmark Results

We evaluate several popular state-of-the-art T2I systems on the CuRe benchmark dataset with our three scorer classes in Tab. 8. We note that due to inbuilt safety filters, Dall-E and Stable Diffusion 1.5 refuse to generate 17% and 1.5% respectively (detailed breakdown in Appendix B.2). For perceptual similarity, we use our $\Delta\phi_{PS}(n, c)$ scorer, which had the highest negative Spearman’s ρ with gold scores on average (Tab. 1). As this scorer is a divergence, a value closer to zero is better. SDXL and Dall-E 3 perform best for SIGLIP 2 cosine distance, while DALL-E 3 and FLUX.1 [dev] slightly edge out SD 1.5 and SDXL with DINOv2.

AIMV2 has lower separability between T2I systems, with SD 1.5 and 3.5 Large 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. 3.2), PS scores on SD 1.5 are likely to be overestimates. For ϕ_{ITA} , we use $\phi_{ITA}(c, r)$, which is a cosine distance using SigLIP 2 dense vector embeddings. We note that ϕ_{PS} for Ideogram is omitted due to API compute constraints, as our metric requires generating multiple seeds of $I(c)$. Across VLMs, the Stable Diffusion class of models perform quite strongly ($XL \geq 3.5 > 1.5$) compared to FLUX.1 [dev], Ideogram 2.0 and Dall-E 3, and this is especially pronounced for CLIP trained on LAION-2B. As Stable Diffusion models 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. 2.1 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.

Notation: To generate an image of artifact a , we first construct a text prompt $P(a)$ with templates over a subset of the artifact attributes. We illustrate an example with *banku*, a type of dumpling from Ghana in Fig. 2, and a more exhaustive list of example prompt templates over subsets of a in Tab. 5 (Appendix C). For example, $P(a|a = \{n\}) = \text{“An image of } \{n\}\text{”}$, $P(a|a = \{n, r\}) = \text{“An image of } \{n\}\text{, from } \{r\}\text{”}$ (the red and orange prompts in Fig. 2 respectively). Let f_θ denote a T2I system f parametrized by weights θ , which takes in a text prompt $P(a)$ for an artifact a , and generates an image

$$I(a) := f_\theta(P(a)).$$

For example, to generate “An image of $\{n\}$, a type of $\{c\}$ from $\{r\}$ ”, $I(n, c, r) = f_\theta(P(a|a = \{n, c, r\}))$; e.g., the image generated by the green prompt in Fig. 2.

To measure cultural representativeness capability of f_θ for region r , we need a method of scoring the quality of generated images. Let $\phi : I \rightarrow \mathbb{R}$ be a quality scorer for CuRe (e.g. similarity of generated images to ground-truth images of artifact a). To measure CuRe of f_θ for region r , we compute an average score over the set of cultural artifacts $a \in \mathcal{A}_r$:

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

	Images		Metric		ϕ
a_1			[42]	0.62	0.62
			[61]	0.71	
a_2			ϕ_{CuRe}^*	0.31	0.31
			ϕ_{PS}^*	0.44	
			$\phi_{PS} \downarrow$	0.15	
	AI : $I(n)$	Real : $G(n)$			

	Images		Metric		ϕ
a_1			[29]	0.13	0.13
			[63]	0.11	
a_2			ϕ_{CuRe}^*	1.00	1.00
			ϕ_{PS}^*	0.75	
			$\phi_{ITA} \uparrow$	0.14	
	AI : $I(n)$	Real : $G(n)$			

	Images		Metric		ϕ
a_1			LPIPS(n)	0.72	0.72
			VS(c)	0.24	
a_2			ϕ_{CuRe}^*	0.93	0.93
			ϕ_{PS}^*	0.66	
			$\phi_{DIV} \uparrow$	0.57	
	AI : $I(n)$	Real : $G(n)$			

(a) Perceptual Similarity Scorer (ϕ_{PS}). **Top:** “Omurice”, **Bottom:** “Chicken Biryani”. Images were generated with Stable Diffusion 3.5 Large.

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

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

Figure 9. A qualitative comparisons of our proposed MIA scorers compared to baselines in three scorer classes: a) Perceptual Similarity (Sec. 2.1) b) Image-Text Alignment (Sec. 2.2) c) Diversity (Sec. 2.3). While human judgments ϕ^* from the user study are on a 1 - 5 Likert Scale, we normalize them to a 0 - 1 scale for a more direct comparison to quantitative scorers. When human judgments of representativeness ϕ_{CuRe}^* and perceptual similarity ϕ_{PS}^* are able to distinguish between a generated image and ground truth image pair (rows a_1 and a_2), our scorers are able to *differentiate* between them (ϕ_{PS} changes significantly), but baselines are not.

Choosing a Scorer: The gold standard for quality scorer ϕ is to survey a large number of people from region r to rate the model’s performance along pre-defined rubrics such as realism and image-text alignment, typically with a 1-5 scale Likert score [33], which we denote ϕ^* (see Fig. 2). 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 [32]. 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 model after deployment [14]. 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. 3).

F. User Study Analytics

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).

F.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 a from country r , we ask them if

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. [36]		✓		✓	✓		
Basu et al. [2]		✓	✓				✓
Ventura et al. [63]		✓					
Khanuja et al. [29]		✓	✓		✓		✓
Kannen et al. [28]		✓	✓				✓
Zhang et al. [72]				✓		✓	
Bayramli et al. [3]			✓	✓			
CuRe	✓	✓		✓	✓	✓	✓

they identify with the culture of $\{r\}$) via worker metadata (see Fig. 10 for details).

F.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. 5). We compute a Kendall’s Tau distance between the rankings from the user study and the ranking given by our scorer for each encoder. 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 (i, j) :

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

For example, for agreement over only three survey respondents for each survey (“Worker Only”), we have $3c2$ pairs of rankings to compute agreement over, which we average. When we add the ranking given by the image encoder (SigLIP 2, AIMV2, DINOv2), we have $4c2$ pairs of rankings to average over. 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, *i.e.* ϕ_{CuRe}^* , alongside their textual justification for their score in Fig. 12. A major cause of disagreement is a miscalibration between worker thoughts and the Likert score selected. For example, for the 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.” Similarly for the 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. Some workers base their rating more on semantic content, while others critique details more harshly (*e.g.* Hogmanay, where one worker highlights incorrect details, while another focuses on the fireworks and crowd presence being accurate). In another case, workers appear to agree with their justification, but mark vastly different Likert scores (cowboy hat, pabellón criollo). In the case of cowboy hat, this appears to be an incorrect interpretation of which Likert score indicates high CuRe, while for pabellón criollo one worker assigns a high score even as though the image is not of the dish, the worker believes it can still be found in their country and justifies the mistake made by the T2I system was due to underspecification in the prompt (*i.e.* the T2I system should be told in the prompt that pabellón criollo is a type of food).

Table 11. Mean and Variance of Likert scores over offensiveness (ϕ_{OFF}) and stereotypicalness (ϕ_{STR}) of images generated by FLUX.1 [dev].

Continent	ϕ_{OFF}	ϕ_{STR}
Africa	1.22 ± 0.64	1.74 ± 1.19
Asia	1.46 ± 0.93	1.90 ± 1.09
Europe	1.31 ± 0.77	1.89 ± 1.17
North America	1.22 ± 0.60	1.88 ± 1.22
Oceania	1.05 ± 0.21	1.62 ± 1.01
South America	1.28 ± 0.81	1.68 ± 1.10

Table 12. Mean and Variance of Likert scores over offensiveness (ϕ_{OFF}) and stereotypicalness (ϕ_{STR}) of the T2I systems split by Global North and Global South.

Model	Global North/South	ϕ_{OFF}	ϕ_{STR}
FLUX.1 [dev]	Global North	1.29 ± 0.54	1.86 ± 1.29
	Global South	1.35 ± 0.68	1.82 ± 1.34
SD 3.5 Large	Global North	1.29 ± 0.75	1.75 ± 1.05
	Global South	1.36 ± 0.78	1.91 ± 1.25
SD 1.5	Global North	1.35 ± 0.83	1.65 ± 1.06
	Global South	1.61 ± 1.06	2.02 ± 1.25

F.3. Survey Respondent Statistics

Recall that we only hire workers to score surveys of artifact a if their country of nationality matches the artifact’s associated region r (Appendix D.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 similarity. 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. 10, 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. 8) on a scale of **1 - Never heard of it** to **5 - Expert**, which we visualize with a bar plot in Fig. 11. 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).

We also report statistics for how offensive and stereotypical of local culture FLUX.1 [dev] generations are, as rated by survey respondents on a 1 to 5 Likert scale, in Tab. 11. We observe that across continents, both offensiveness and stereotypicalness scores are between 1 (“Not at all”) and 2

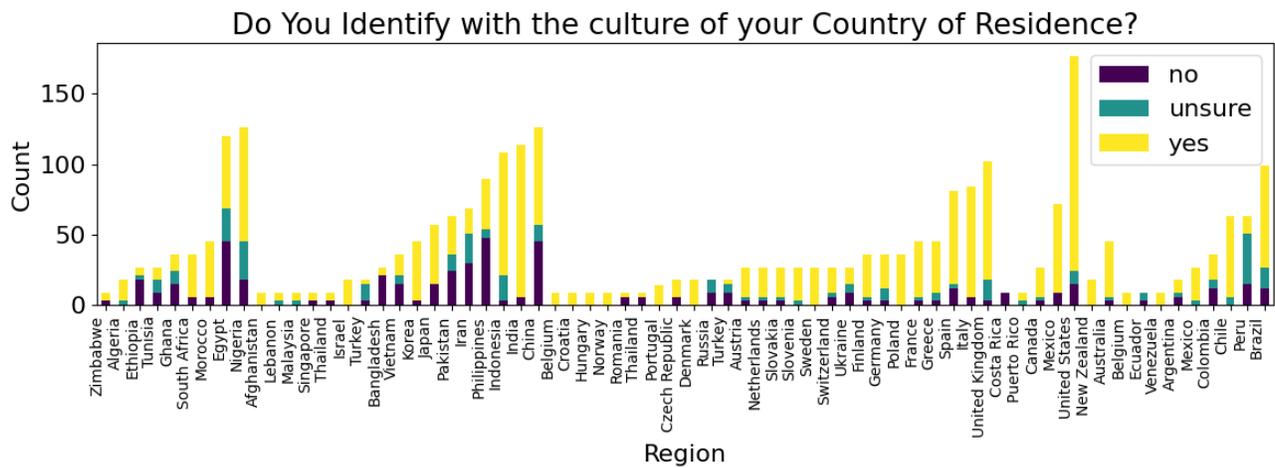
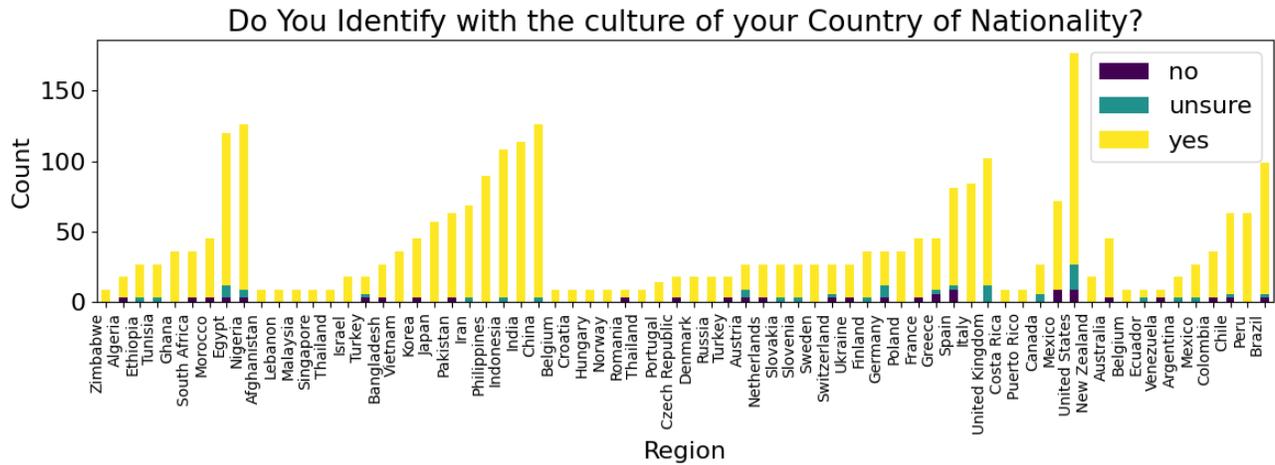


Figure 10. Responses to “Do you identify with the culture of [country]?” summed across all surveys grouped by Continent and sorted by total count of responses.

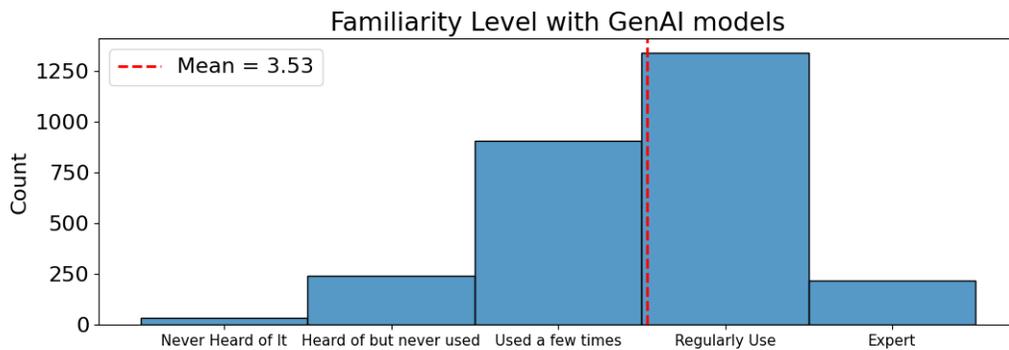


Figure 11. Responses to “Familiarity with GenAI models” summed across all surveys.

(“Slightly”) out of 5, indicating that FLUX.1 [dev] is perceived to be minimally offensive and slightly stereotypical of local cultures. Asia ranks highest for offensiveness in

terms of mean and variance, followed by Europe and South America. Asia also ranks highest for stereotypicalness, followed by Europe and North America.

AI Image	Real Image	Feedback
		<p>$\phi_{CuRe}^* = 1$ 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>$\phi_{CuRe}^* = 5$ 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>
Yangpu Bridge		
		<p>$\phi_{CuRe}^* = 1$ 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>$\phi_{CuRe}^* = 5$ 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>
Rostás Pál Monument		
		<p>$\phi_{CuRe}^* = 1$ out of 5: fireworks are too low, faces blurry and distorted</p> <p>$\phi_{CuRe}^* = 5$ out of 5: The image shows displays of fireworks which are often associated with celebrating Hogmanay, and also large crowds of people who gather t[o] celebrate and watch the fireworks</p>
Hogmanay		
		<p>$\phi_{CuRe}^* = 1$ 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>$\phi_{CuRe}^* = 5$ out of 5: Cowboy hats from my country are made from straws or leather which is shown in the picture here.</p>
Cowboy Hat		
		<p>$\phi_{CuRe}^* = 1$ 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>$\phi_{CuRe}^* = 5$ out of 5: Ok, the image created by the AI is a creole Pavillion 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>
Pabellòn criollo		
		<p>$\phi_{CuRe}^* = 2$ out of 5: The image is more bright and shows different features of him</p> <p>$\phi_{CuRe}^* = 5$ 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>
Fawad Khan		

Figure 12. Comparison between user study written feedback when the same artifact for the same T2I system was scored by users on opposite ends for ϕ_{CuRe}^* (Part 1).

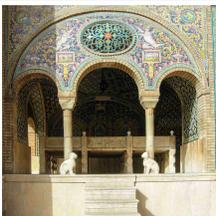
AI Image	Real Image	Feedback
		<p>$\phi_{CuRe}^* = 5$ out of 5: This isn't a picture of a bunad at all, but rather a picture associated with Norwegian cultural heritage, a beautiful landscape with mountains and fjords.</p> <p>$\phi_{CuRe}^* = 1$ out of 5: This image doesn't show anything connected to Hardangerbunad which is an item of clothing, it shows a landscape with mountains and fjords.</p>
Hardangerbunad		
		<p>$\phi_{CuRe}^* = 1$ out of 5: The image used the wrong race of the person</p> <p>$\phi_{CuRe}^* = 5$ out of 5: It appears much accurate and more clear</p>
Kaapse Klopse		
		<p>$\phi_{CuRe}^* = 1$ 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>$\phi_{CuRe}^* = 5$ 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>
Kue Nastar		
		<p>$\phi_{CuRe}^* = 2$ 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>$\phi_{CuRe}^* = 5$ 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>
Golestan Palace		
		<p>$\phi_{CuRe}^* = 1$ 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>$\phi_{CuRe}^* = 4$ out of 5: The image shows two lovers touching and almost kissing, while in the real painting they aren't.</p>
Stanza dell' Amore Coniugale		
		<p>$\phi_{CuRe}^* = 1$ 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>$\phi_{CuRe}^* = 4$ out of 5: The floor part does not look authentic as well as the surrounding city</p>
Brooklyn Bridge		

Figure 12. Comparison between user study written feedback when the same artifact for the same T2I system was scored by users on opposite ends for ϕ_{CuRe}^* (Part 2).

G. Perceptual Similarity Ablations

As detailed in Appendix B.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. 5.

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

G.1. Head-Tail Split using PS Scorers

For each T2I system f_θ , in Fig. 13 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 four 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.

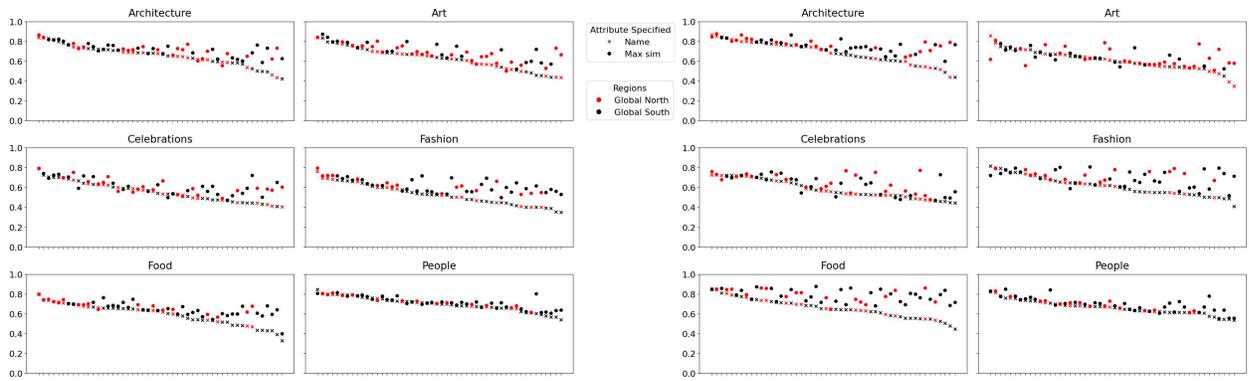
As seen in Fig. 13, we observe a spread of scores in the tail across T2I systems for both scorers ϕ_{GT} and ϕ_{PS} , though the starting point and magnitude of spread (vertical height) differs by supercategory. ϕ_{GT} tends to have a lower spread than ϕ_{PS} , with fewer outliers. All T2I systems have 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. 13e) appears to be the most homogenous in perceptual similarity, as the spread is much lesser than other models, even for Food.

For a more fine-grained analysis than aggregating at the Global North / South divide, we also examine PS scores across each region (country) in the CuRe dataset in Tab. 13. In some continents, such as North America, we observe a

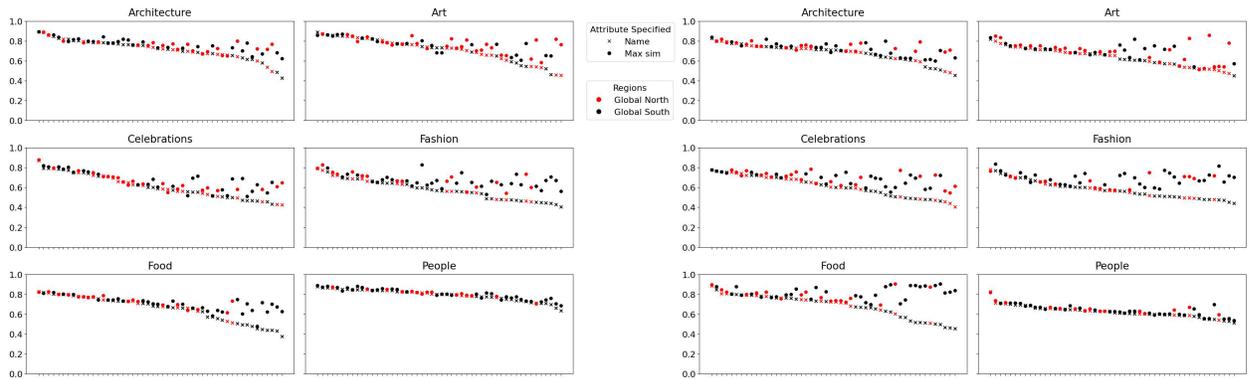
Table 13. Perceptual similarity CuRe scores of all regions in the CuRe dataset, sorted descending by user study scores ϕ_{CuRe}^* .

Continent	Region	Perceptual Similarity Scorer				
		ϕ_{CuRe}^*	$\phi_{GT}(n)$	$\phi_{PS}(n)$	$\Delta\phi_{PS}(n, c)$	$\Delta\phi_{PS}(n, c, r)$
Africa	Egypt	3.265	0.719	0.617	0.019	0.010
	Nigeria	2.959	0.716	0.638	0.010	0.006
	Morocco	2.800	0.638	0.618	0.048	0.045
	Algeria	2.722	0.556	0.569	0.058	0.065
	Ghana	2.694	0.589	0.588	0.049	0.034
	South Africa	2.528	0.641	0.631	0.052	0.043
	Ethiopia	2.259	0.606	0.623	0.039	0.039
	Zimbabwe	2.222	0.770	0.686	-0.011	-0.003
	Tunisia	1.741	0.498	0.529	0.132	0.127
Asia	India	3.910	0.690	0.654	0.016	0.014
	Malaysia	3.333	0.735	0.779	0.010	0.007
	Pakistan	3.024	0.760	0.612	0.002	-0.000
	Indonesia	2.875	0.656	0.629	0.032	0.028
	Iran	2.861	0.614	0.560	0.051	0.049
	Philippines	2.772	0.644	0.598	0.043	0.039
	Japan	2.722	0.650	0.610	0.046	0.040
	China	2.667	0.652	0.624	0.033	0.030
	Lebanon	2.667	0.720	0.763	-0.006	0.006
	Israel	2.444	0.701	0.709	-0.019	-0.013
	Korea	2.422	0.662	0.627	0.033	0.025
	Thailand	2.250	0.674	0.662	0.053	0.038
	Vietnam	2.139	0.587	0.588	0.042	0.041
	Bangladesh	2.111	0.529	0.547	0.091	0.082
	Singapore	1.889	0.787	0.696	0.028	0.022
Afghanistan	1.889	0.798	0.586	0.003	-0.002	
Europe	Slovenia	3.407	0.634	0.629	0.065	0.060
	Czech Republic	3.278	0.732	0.694	0.012	0.024
	Switzerland	3.185	0.605	0.622	0.013	0.004
	United Kingdom	3.139	0.737	0.623	0.012	0.013
	Germany	3.111	0.624	0.582	0.058	0.056
	Denmark	3.111	0.591	0.575	0.080	0.068
	Norway	3.111	0.572	0.586	0.052	0.053
	Italy	3.014	0.695	0.662	0.028	0.028
	Netherlands	3.000	0.638	0.582	0.048	0.046
	Poland	3.000	0.626	0.610	0.034	0.041
	France	2.978	0.687	0.655	0.023	0.015
	Russia	2.833	0.675	0.651	0.035	0.035
	Greece	2.733	0.690	0.642	0.030	0.028
	Belgium	2.722	0.706	0.638	-0.001	0.003
	Spain	2.617	0.684	0.631	0.021	0.018
	Portugal	2.600	0.614	0.593	0.060	0.065
	Austria	2.519	0.618	0.619	0.026	0.026
	Slovakia	2.407	0.660	0.643	0.042	0.034
	Romania	2.333	0.608	0.597	0.077	0.092
Hungary	2.333	0.684	0.710	0.005	0.013	
Croatia	2.222	0.529	0.561	0.097	0.078	
Finland	2.201	0.654	0.637	0.014	0.015	
Turkey	2.194	0.656	0.668	0.032	0.016	
Sweden	2.093	0.666	0.630	0.029	0.028	
Ukraine	1.741	0.524	0.556	0.098	0.084	
North America	United States	3.463	0.727	0.664	0.009	0.010
	Canada	3.296	0.656	0.728	-0.002	-0.001
	Mexico	2.889	0.662	0.675	0.020	0.015
	Puerto Rico	2.556	0.499	0.508	0.170	0.147
	Costa Rica	1.444	0.598	0.666	0.040	0.048
Oceania	Australia	2.867	0.659	0.688	0.024	0.021
	New Zealand	2.778	0.617	0.618	0.035	0.026
South America	Brazil	3.108	0.653	0.660	0.014	0.014
	Venezuela	2.889	0.483	0.545	0.102	0.106
	Chile	2.849	0.735	0.629	0.009	0.009
	Argentina	2.403	0.716	0.677	0.020	0.014
	Peru	2.286	0.575	0.623	0.036	0.026
	Colombia	2.111	0.635	0.619	0.043	0.034
Ecuador	1.667	0.537	0.528	0.116	0.097	

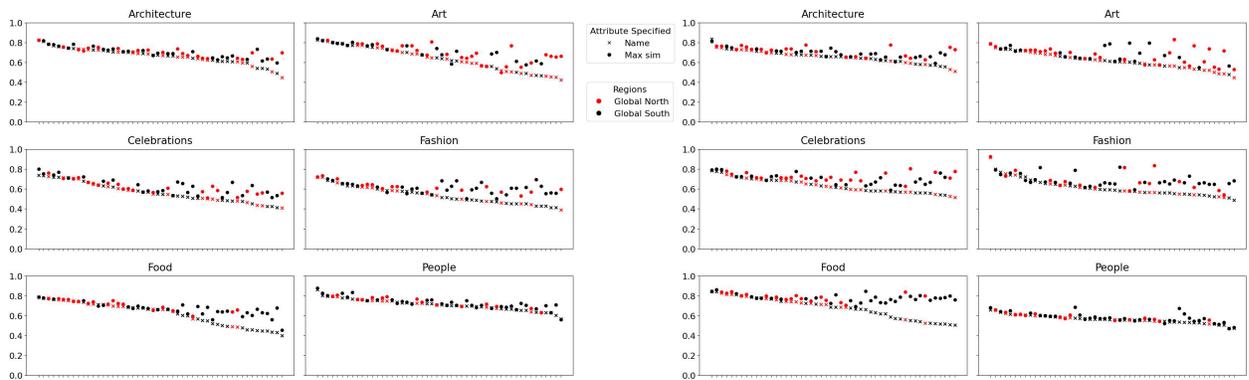
clear correspondence between the expected availability of training data and the performance of our perceptual similarity scorers. Regions like the United states and Canada have high ϕ_{CuRe}^* which aligns with the scores from our ϕ_{PS} scorers, while Puerto Rico and Costa Rico have much lower scores for both. However, this trend is not uniformly observed across all continents. Other continents, like Africa and South America have little relationship be-



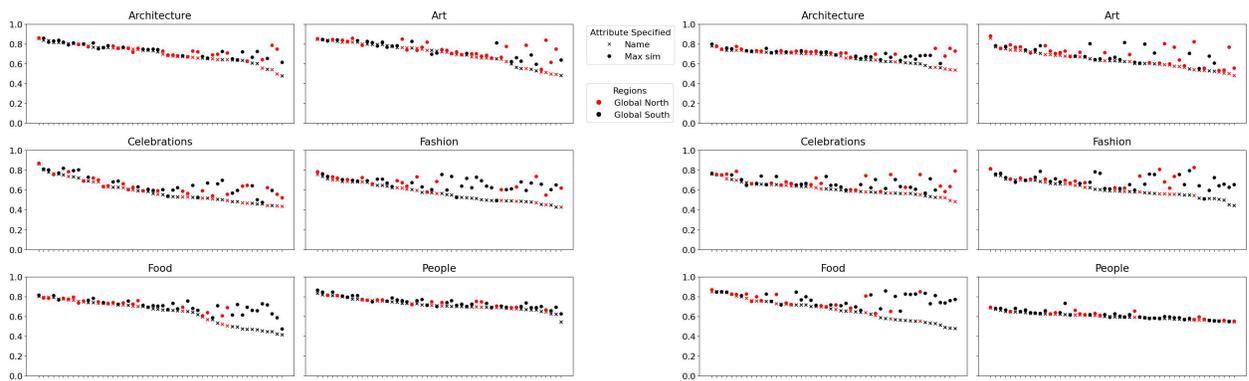
(a) FLUX.1 dev



(b) Stable Diffusion 3.5 Large



(c) Stable Diffusion 1.5



(d) Stable Diffusion XL

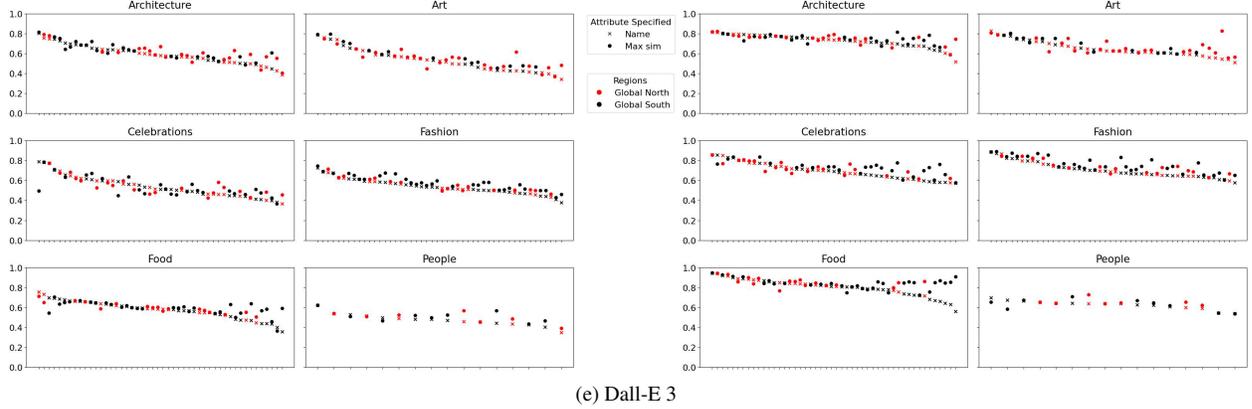


Figure 13. **Left:** ϕ_{GT} for artifacts from the supercategories generated using FLUX.1 [dev] with similarity averaged to get one value for each artifact. **Right:** ϕ_{PS} for artifacts from the supercategories generated using FLUX.1 [dev] with similarity averaged to get one value for each artifact. $\phi_{GT}(n)$ and $\phi_{PS}(n)$ (Vanilla) has the “x” marker while highest similarity score (Best) amongst the other 3 prompts has the “o” marker. The colors indicate whether the artifact falls in the Global North (red) or Global South (black). Artifacts are sorted by descending Vanilla similarity score.

tween the quantitative perceptual similarity scorers and the user study gold scores. While certain regions fall in line with the ϕ_{CuRe}^* scores, for example, Tunisia has the worst scores across all quantitative perceptual similarity scorers and also the lowest ϕ_{CuRe}^* in Africa, the ranking of the scorers rarely match with the ranking of ϕ_{CuRe}^* .

G.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. 14. We also show Likert scores for cultural representativeness ϕ_{CuRe}^* and textual justification for each example.

As we observed in our quantitative results (Tab. 1), 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.* Bayt al-Suhaymi, Moai, and George Lucas). We also highlight a failure case for our scorers with *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. 1, 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. The artifact *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, the artifact *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 *Blocos carnavalescos de São Paulo* which was much less culturally representative according to human judgments.

AI Image	Real Image	Feedback	PS Scorers
 <p>$\phi_{\text{CuRe}}^* = 2 \text{ out of } 5$</p>	 <p>Bayt al-Suhaymi</p>	<ol style="list-style-type: none"> 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 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 	$\phi_{GT}(n) = 0.654$ $\phi_{PS}(n) = 0.580$ $\Delta\phi_{PS}(n, c) = 0.052$ $\Delta\phi_{PS}(n, c, r) = 0.006$
 <p>$\phi_{\text{CuRe}}^* = 4 \text{ out of } 5$</p>	 <p>Moai</p>	<ol style="list-style-type: none"> 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. 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. 	$\phi_{GT}(n) = 0.808$ $\phi_{PS}(n) = 0.581$ $\Delta\phi_{PS}(n, c) = -0.014$ $\Delta\phi_{PS}(n, c, r) = -0.014$
 <p>$\phi_{\text{CuRe}}^* = 1 \text{ out of } 5$</p>	 <p>Blocos carnavalescos de São Paulo</p>	<ol style="list-style-type: none"> The costumes are indeed carnival like, but the blocos are less stravgant and the costumes wore by the people usually contain some kind of a joke on a very low budget costume Their physical appearance and posture. 	$\phi_{GT}(n) = 0.712$ $\phi_{PS}(n) = 0.757$ $\Delta\phi_{PS}(n, c) = 0.004$ $\Delta\phi_{PS}(n, c, r) = -0.027$
 <p>$\phi_{\text{CuRe}}^* = 5 \text{ out of } 5$</p>	 <p>Hmong textile art</p>	<ol style="list-style-type: none"> The design motif and colors are quite close to the real ones I think the repetitive patterns and the colourful decoration make it accurate 	$\phi_{GT}(n) = 0.620$ $\phi_{PS}(n) = 0.634$ $\Delta\phi_{PS}(n, c) = 0.019$ $\Delta\phi_{PS}(n, c, r) = 0.018$
 <p>$\phi_{\text{CuRe}}^* = 1 \text{ out of } 5$</p>	 <p>Tallarín saltado</p>	<ol style="list-style-type: none"> the noodles look like worms, tallarín saltado is spaghetti with meat and vegetables whereas the image only shows weird pasta and no meat nor onions/tomatoes. 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 Tallarín saltado 	$\phi_{GT}(n) = 0.654$ $\phi_{PS}(n) = 0.696$ $\Delta\phi_{PS}(n, c) = -0.014$ $\Delta\phi_{PS}(n, c, r) = -0.075$
 <p>$\phi_{\text{CuRe}}^* = 5 \text{ out of } 5$</p>	 <p>George Lucas</p>	<ol style="list-style-type: none"> 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. I do not see anything that is "wrong" other than the eyes. They are obviously AI. 	$\phi_{GT}(n) = 0.745$ $\phi_{PS}(n) = 0.571$ $\Delta\phi_{PS}(n, c) = -0.011$ $\Delta\phi_{PS}(n, c, r) = -0.040$

Figure 14. Visualization of quantitative perceptual similarity (PS) scorers and user study CuRe scores and text feedback (Appendix D.5). The first two artifacts were generated with FLUX.1 [dev], the second two with SD 3.5 Large, and the last two with SD 1.5.

H. Image-Text Alignment

We examine an important factor in the context of image-text alignment scorers in Appendix H.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 H.2.

Table 14. 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. [29]	“This image is culturally relevant to {r}.”
Ventura et al. [63]	“Image from {r} culture.”
o3 mini [41]	“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}.”

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

Table 15. Spearman correlation values of ITA scorers with human judgments of perceptual similarity ϕ_{PS}^* across ITA scorer models for FLUX.1 [dev]. ITA scorer models we evaluate are OpenCLIP models trained on OpenAI WIT [46], Data Filtering Networks (DFN-5B) [17], LAION-2B [55], and DataComp (DC-1B) [20].

ITA Scorer	LAION-2B	WIT	DFN-5B	SigLIP 2	DC-1B
Khanuja et al. [29]	0.18	0.16	0.12	0.11	0.12
Ventura et al. [63]	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. 2 and ablate over the choice of VLM backbone used to compute image-text alignment for FLUX.1 [dev] in Tab. 15 (see Sec. 3.3 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. [29], Ventura et al. [63], and o3 mini) which query the VLM for CuRe score directly with region information r are sensitive to changes in the backbone, showing high variability in ITA scores. As we marginally increase attributes specified to the T2I system ($n \rightarrow c \rightarrow r \rightarrow c, r$), the rank correlations become more consistent across VLM backbones. Our proposed metrics show both higher and more consis-

tent rank correlations with ϕ_{PS}^* gold scores across all VLM backbones, showing that they are less sensitive to the pre-training distribution of the VLM for evaluating cultural representativeness.

H.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. 14 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. **Example 5** (*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 **Example 1** (*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 **Example 3** (*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. **Example 4** (*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 **Example 6** (*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. **Example 2** (*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 outright pleas to “stop this”. This highlights an important issue regarding images created by generative AI systems: a sense of subjectivity and heterogeneity of opinions involved in evaluating cultural artifacts and towards T2I systems themselves.

AI Image	Real Image	Feedback	ITA Scorers
 <p>$\phi_{\text{CuRe}}^* = 5$ out of 5</p>	 <p>Bangles</p>	<ol style="list-style-type: none"> 1. Bangles are circular ornaments with some colors and patterns on them. 2. The width and the carvings and the colors make it seem more aligned to my culture 	<p>Khanuja et al. [29] = 0.060 Ventura et al. [63] = 0.045 $\phi_{ITA}(c, r) = 0.124$</p>
 <p>$\phi_{\text{CuRe}}^* = 1$ out of 5</p>	 <p>Festival de la Primavera</p>	<ol style="list-style-type: none"> 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. 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>Khanuja et al. [29] = 0.086 Ventura et al. [63] = 0.067 $\phi_{ITA}(c, r) = 0.156$</p>
 <p>$\phi_{\text{CuRe}}^* = 1.67$ out of 5</p>	 <p>Jalangkote</p>	<ol style="list-style-type: none"> 1. The image use the wrong thing, AI image is using stone like building while Jalangkote is a food. a complete different genre 2. The AI image is not even a food. 	<p>Khanuja et al. [29] = 0.112 Ventura et al. [63] = 0.093 $\phi_{ITA}(c, r) = 0.022$</p>
 <p>$\phi_{\text{CuRe}}^* = 1.34$ out of 5</p>	 <p>Puchner Mansion</p>	<ol style="list-style-type: none"> 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? 2. I feel like the AI focused too much on the word Mansion, and especially the american/older british kind. 	<p>Khanuja et al. [29] = 0.040 Ventura et al. [63] = 0.053 $\phi_{ITA}(c, r) = 0.096$</p>
 <p>$\phi_{\text{CuRe}}^* = 4.67$ out of 5</p>	 <p>Zwölf Glaubensartikel</p>	<ol style="list-style-type: none"> 1. Overall the vegetation seems possible to find in my country 2. This image could be a shot from above of some forest in my country. 	<p>Khanuja et al. [29] = 0.108 Ventura et al. [63] = 0.091 $\phi_{ITA}(c, r) = 0.079$</p>
 <p>$\phi_{\text{CuRe}}^* = 2$ out of 5</p>	 <p>Michelle Bachelet</p>	<ol style="list-style-type: none"> 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. 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>Khanuja et al. [29] = 0.066 Ventura et al. [63] = 0.054 $\phi_{ITA}(c, r) = 0.164$</p>

Figure 15. Visualization of ITA scorers and user 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.

I. Diversity

LPIPS. A high LPIPS [73] 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:

$$LPIPS(c) = \frac{1}{|\mathcal{A}_c|} \sum_{a \in \mathcal{A}_c} \phi(I(a)).$$

Vendi Scores. Vendi Scores (VS) [19, 28] define a similarity measure via a kernel over selected attributes (e.g. $r = \{\text{Country, Continent}\}$). 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 artifacts belonging to c , i.e. $\hat{a}(I(c)_j) = \arg \max_{a \in c} \text{sim}(I(c), I(a))$.

A $j \times j$ kernel similarity matrix is then computed based on a selected attribute (see Experimental Setup of Kannen et al. [28] for details). The primary drawback VS as a scorer is that its quality depends entirely on this initial assignment \hat{a} and choice of kernel, which in turn depends on the image encoder used to compute sim .

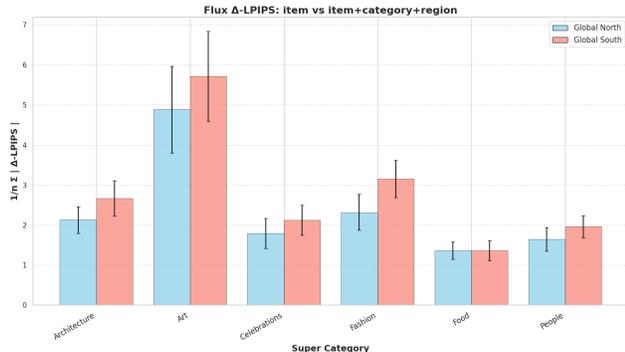


Figure 16. $\Delta\phi_{DIV}(a)$ for FLUX.1 [dev] images of artifacts in the CuRe dataset belonging to the Global North (GN) and Global South (GS).

I.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. 2.1). Recall from Eq. (4) that our ϕ_{DIV} scorer compute LPIPS diversity over a set of images generated across attribute specification levels:

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

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. 2). 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 long tail of the T2I system. To test this hypothesis, we compute the marginal increase in information as divergence of $\phi_{DIV}(a)$ from an LPIPS over only images generated with $a = \{n\}$, i.e.

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

We visualize this marginal increase $\Delta\phi_{DIV}(a)$ 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 $\Delta\phi_{DIV}(a)$ over all artifacts belonging to each bucket for each supercategory s , and show a bar plot for FLUX.1 [dev] in Fig. 16. We observe that the change in diversity score is higher for artifacts belonging to the global south than for artifacts belonging to the global north across all supercategories, which aligns with our hypothesis. Our proposed scorer $\phi_{DIV}(a)$ can thus serve as a good proxy to predict whether an image generated by a T2I system lies. 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 cheap to compute when compared to computing artifact frequency over a large pretraining dataset via string matching, such as in Parashar et al. [43], which we discuss in more detail in Appendix K.

I.2. Qualitative Analysis of DIV scorers

We highlight several qualitative examples on the CuRe dataset of our DIV scorer $\phi_{DIV}(a)$ compared to baselines in Fig. 17. In alignment with our quantitative observations in Appendix I.1, in the examples “Third Mainland Bridge”, “Rabat Lighthouse” and “Penelope Cruz”, all diversity scorers show an inverse relationship to human judgments of cultural representativeness ϕ_{CuRe}^* . In the examples of “Vaso de los Guerrero” and “Hokkien Mee”, $\phi_{DIV}(a)$ captures this negative rank correlation better than the baseline (lower score than baseline for high ϕ_{CuRe}^*). We highlight a failure case where the baseline outperforms in scorer with “Ushabti”, where for a low ϕ_{CuRe}^* , the baseline shows a higher diversity score.

AI Image	Real Image	Feedback	Score
 <p>$\phi_{CuRe}^* = 1$ out of 5</p>	 <p>Ushabti</p>	<ol style="list-style-type: none"> 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. 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>LPIPS(n) = 0.62 $\phi_{DIV}(n) = 0.48$</p>
 <p>$\phi_{CuRe}^* = 4$ out of 5</p>	 <p>Hokkien Mee</p>	<ol style="list-style-type: none"> 1. Should have less liquid sauce. May need to add shrimps. The noodle should be fried. 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>LPIPS(n) = 0.70 $\phi_{DIV}(n) = 0.59$</p>
 <p>$\phi_{CuRe}^* = 4.5$ out of 5</p>	 <p>Third Mainland Bridge</p>	<ol style="list-style-type: none"> 1. The image contains the distinctive part of the real location and image, it contains most of the details of the real thing. 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>LPIPS(n) = 0.50 $\phi_{DIV}(n) = 0.50$</p>
 <p>$\phi_{CuRe}^* = 5$ out of 5</p>	 <p>Rabat Lighthouse</p>	<ol style="list-style-type: none"> 1. The catholic or christian cross 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>LPIPS(n) = 0.51 $\phi_{DIV}(n) = 0.54$</p>
 <p>$\phi_{CuRe}^* = 4$ out of 5</p>	 <p>Penelope Cruz</p>	<ol style="list-style-type: none"> 1. The facial features are very similar to the real ones 2. The hair is typical of any country, so it could as well be spanish. The white clothes is typical mediterranean. 	<p>LPIPS(n) = 0.74 $\phi_{DIV}(n) = 0.62$</p>
 <p>$\phi_{CuRe}^* = 2$ out of 5</p>	 <p>Vaso de los Guerreros</p>	<ol style="list-style-type: none"> 1. The AI-generated image looks too unreal; it doesn't feel natural, as if it was too exaggerated 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>LPIPS(n) = 0.75 $\phi_{DIV}(n) = 0.55$</p>

Figure 17. Diversity scorer and user feedback

J. MLLM as a Judge

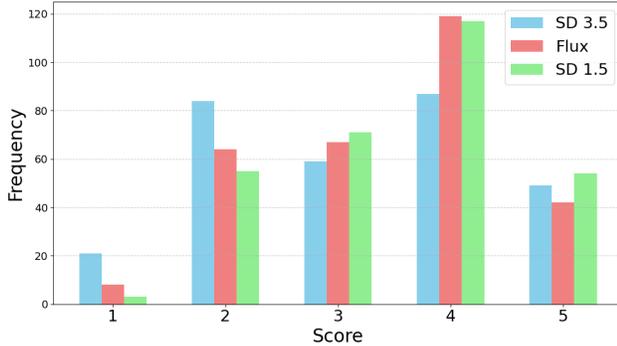


Figure 18. We evaluate how well Gemini 2.0 Flash CuRe score justifications match real human CuRe score justifications with OpenAI o1, which gives a score from 1 (poor) to 5 (excellent).

To our knowledge for the first time, we evaluate the capacity of a MLLM at approximating human judgments (gold scores) of cultural representativeness (ϕ_{CuRe}^*) and image perceptual similarity (ϕ_{PS}^* , ϕ_{GT}^*). Specifically, we query Gemini 2.0 Flash [13] with the same set of questions asked to real humans in our user study (details in Appendix J), and output a score from 1 (low) to 5 (high), similar to user study Likert scores. We also query Gemini for a textual justification of these scores, so we can inspect its reasoning.

We follow the same evaluation setup as PS, ITA and DIV scorers (Sec. 3.1) and compute a Spearman rank correlation between Gemini scores and gold scores from our user study. For PS scorers Tab. 1, 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. 2), 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 . This intuitively suggests that their pretraining dataset has wider cultural coverage of artifact - region associations compared to smaller VLMs like OpenCLIP and SigLIP 2. While these results are promising, our examination of Gemini’s textual justifications (Fig. 19) show that it 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).

In addition to the scores it provides for CuRe, perceptual similarity and ground-truth likelihood, we ask Gemini 2.0 Flash to provide a textual justification for its scores (details in Appendix J.1). Following previous work using an independent MLLM as a judge [8], we evaluate the quality of Gemini 2.0 Flash’s text justification with a state-of-the-art reasoning LLM, OpenAI o1 [40] (details in Appendix J.2). Specifically, for a given artifact, we ask o1 to output a score from 1 (low) to 5 (high) for how closely Gemini 2.0

Flash’s score justification matches the Likert score justifications provided by real humans in the user study. As seen in Fig. 18, according to o1, Gemini 2.0 Flash is able to most accurately justify scores for SD 1.5 (3.6 / 5), followed by Flux.1 [dev] (3.4 / 5), and lastly SD 3.5 Large (3.2 / 5). Even with these promising initial results at providing fine-grained textual justifications that resemble real human justifications, we show that Gemini 2.0 Flash still has many failure cases in Fig. 19b.

J.1. Gemini 2.0 Flash as a Scorer

We summarize the prompts administered to the Gemini 2.0 Flash model to elicit scores and textual justifications regarding cultural representativeness (ϕ_{CuRe}^*), perceptual similarity (ϕ_{PS}^*), and ground truth similarity (ϕ_{GT}^*). The core evaluation prompt administered to Gemini 2.0 Flash (as mentioned in item 1 above) was structured around five key questions. These questions were designed to mirror the assessments performed by human participants in our user study (Appendix D.5, Q3c), focusing on critical aspects of the generated images:

- PS Likert (Perceptual Similarity):** Assesses the visual resemblance between the AI-generated image and provided real reference images.
(Measured on a Likert scale from 1 "Not at all similar" to 5 "Extremely similar")
- Representativeness:** Evaluates the likelihood that the depicted item could be found within a specific cultural context ({country}).
(Measured on a Likert scale from 1 "Highly Unlikely" to 5 "Extremely Likely")
- GT Likert (Ground Truth Similarity):** Judges the accuracy of the AI’s depiction of the item ({item_name}) itself, irrespective of the cultural context.
(Measured on a Likert scale from 1 "Highly Unlikely" to 5 "Extremely Likely")
- Similarity Explanation:** Requests a textual justification identifying specific visual elements contributing to the similarity or difference compared to real images.
- Description (Accuracy/Inaccuracy Details):** Asks for textual details explaining the image’s accuracy or inaccuracy relative to its typical appearance in the specified {country} or general knowledge of the {item_name}.

Prompt for perceptual similarity Likert:

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:

```
{  
  "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."  
}
```

Prompt for representativeness and GT Likert:

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. 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. GT Likert: 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'.

J.2. o1 as a Judge

We discuss the precise prompt provided to the OpenAI o1 model, instructing it to evaluate the semantic similarity between the justifications generated by Gemini 2.0 Flash and the aggregated justifications from our human user study

Subsequent to the initial analysis by Gemini 2.0 Flash, a meta-evaluation step was employed using the OpenAI o1 model. The objective of this stage was to quantitatively assess how closely the textual justifications generated by Gemini 2.0 Flash align with the reasoning provided by human evaluators.

To prepare the necessary inputs for this comparative analysis, all relevant free-text descriptions collected from human participants for a given item for a specific model during the user study were first concatenated.

The prompt designed for the o1 model, detailed below, was then supplied with both the specific justification text produced by Gemini 2.0 Flash and this corresponding block of aggregated human descriptions. The core instruction within this prompt directed the o1 model to perform a semantic similarity evaluation between these two distinct sets of responses.

o1 Prompt for Evaluating description Similarity:

You are an AI evaluator tasked with assessing the semantic similarity between two sets of responses.

One set consists of human ratings (2 to 5 responses in a list), and the other is a single response from a Vision-Language Model (VLM).

Your task is to evaluate how well the VLM’s response captures the key cultural details mentioned in the human responses. Consider the following scoring criteria:

1 - Very Low Similarity: The VLM response fails to capture key details or is largely unrelated to the human descriptions.

2 - Low Similarity: Some relevant aspects are present, but major cultural details or themes are missing or incorrect.

3 - Moderate Similarity: The VLM response captures some important details but lacks nuance or specificity compared to the aggregated human responses.

4 - High Similarity: The VLM response aligns well with the human descriptions, with only minor omissions or differences in emphasis.

5 - Very High Similarity: The VLM response is highly aligned with the human descriptions, capturing all key cultural details accurately.

Provide your answer as a single number (1-5) along with a brief explanation (1-2 sentences) that outlines the main reasons behind your rating in this format:

```
{ "score": A, "explanation": "Explanation for score A" }
```

J.3. Analysis of Gemini 2.0 Flash Responses.

Overall Gemini 2.0 Flash demonstrates strong performance, often providing detailed and coherent explanations. However, while the model is generally reliable, it is not a complete solution for modeling nuanced cultural feedback. In this section, we walk through a set of illustrative cases (Fig. 19) that show where Gemini 2.0 Flash either fails or succeeds in matching the representational cues prioritized by human users. Most examples expose limitations especially when cultural, semantic, or regional subtleties are involved though a few clear successes demonstrate the model’s potential when the visual content is unambiguous or culturally neutral.

In **Example 1** (*Femi Kuti*), users were unanimous in rejecting the likeness citing incorrect facial features and hairstyle as key indicators that the image was not representative. Gemini 2.0 Flash, however, rates the similarity as high, based on general body structure and attire. This reveals a pattern where the model focuses on surface-level alignment but misses culturally specific identity markers.

Example 2 (*Jollof Rice*) shows a reverse failure: users appreciate the image, noting the correct rice type, color, and overall presentation. Gemini 2.0 Flash, however, discredits it due to less relevant details like garnish or grain shape. The model’s critique prioritizes visual specifics that don’t necessarily carry the same weight in cultural context.

In **Example 3** (*Chuseok*), Gemini 2.0 Flash’s description emphasizes the presence of traditional Korean dress and a village setting. But users notice features that don’t belong Chinese-style lanterns, architectural elements, and hairstyle. While the model sees broad alignment, it overlooks cross-cultural leakage that human evaluators catch immediately.

Example 4 (*Sámi Headwear*) continues this trend. The model focuses on color and form, affirming the image as accurate, while users reject it based on incorrect materials, proportions, and representation of the subject. These errors are small but significant, when dealing with indigenous artifacts where authenticity is tightly linked to detail.

For **Example 5** (*Takht-e Fulad*), Gemini 2.0 Flash misidentifies the architecture as Egyptian and dismisses its relevance to Iran entirely. Users instead interpret the image as resembling other historical Iranian sites, even if not the exact artifact. The model’s binary judgment misses this more flexible, human interpretation of regional similarity.

Example 6 (*Portrait of Amir Kabir*) offers a similar case: Gemini 2.0 Flash gives a broadly positive assessment, referencing Qajar-era styling. Users, however, flag specific missing details like the hat, mustache style, and necklace elements that signify historical authenticity. The model gets the context, but not the execution.

Example 7 (*Kiping*) is more straightforward. Both Gemini 2.0 Flash and users recognize that the image is entirely off the AI generated a cartoon animal instead of a food item. In this case, the model correctly identifies the error and does not attempt to justify the image.

Example 8 (*Pulseira*) Gemini 2.0 Flash agrees with the users that the bracelet is a valid representation, even if not explicitly regional. It correctly identifies visual plausibility and acknowledges the cultural neutrality of the design. These are the kinds of examples where Gemini 2.0 Flash’s descriptive abilities align well with human expectations.

Regional Trend: A notable pattern is that most failure cases involve artifacts from the Global South. While Gemini 2.0 Flash performs well on globally familiar or culturally neutral items, it struggles with region-specific representations that require localized knowledge. This may reflect imbalances in training data coverage, highlighting the need for broader cultural representation to improve alignment with diverse user perspectives.

Figure 19. 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)

AI Image	Real Image	Feedback
 <p data-bbox="186 554 337 596">Artifact: Femi Kuti Category: Musician</p>	 <p data-bbox="402 554 570 596">Region: Nigeria Super Category: People</p>	<p data-bbox="613 310 867 331">Human Judgments (User Study):</p> <ol data-bbox="613 338 1398 401" 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 with me and the hair is just off....Femi has very distinctive blend of grey hair. <p data-bbox="613 407 906 428">MLLM Judgment (Gemini 2.0 Flash):</p> <p data-bbox="613 434 1442 541">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> <p data-bbox="613 548 857 569">o1 critique of MLLM accuracy:</p> <p data-bbox="613 575 1425 617">The VLM's description contradicts the humans' consensus that the image is not Femi Kuti and misses key details about the distinctive hair and facial features.</p>
 <p data-bbox="186 875 337 917">Artifact: Jollof Rice Category: Rice dish</p>	 <p data-bbox="402 875 570 917">Region: Nigeria Super Category: Food</p>	<p data-bbox="613 625 867 646">Human Judgments (User Study):</p> <ol data-bbox="613 653 1305 716" 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 data-bbox="613 722 906 743">MLLM Judgment (Gemini 2.0 Flash):</p> <p data-bbox="613 749 1442 856">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> <p data-bbox="613 863 857 884">o1 critique of MLLM accuracy:</p> <p data-bbox="613 890 1442 953">The VLM's response directly contradicts the human descriptions that emphasize the color, rice type, and ingredients pointing to Jollof rice. It dismisses the dish as not resembling Jollof rice, missing the key cultural details the humans highlighted.</p>
 <p data-bbox="167 1268 357 1310">Artifact: Chuseok Category: Harvest festival</p>	 <p data-bbox="386 1268 586 1310">Region: South Korea Super Category: Celebrations</p>	<p data-bbox="613 961 867 982">Human Judgments (User Study):</p> <ol data-bbox="613 989 1476 1136" 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 data-bbox="613 1142 906 1163">MLLM Judgment (Gemini 2.0 Flash):</p> <p data-bbox="613 1169 1458 1318">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> <p data-bbox="613 1325 857 1346">o1 critique of MLLM accuracy:</p> <p data-bbox="613 1352 1442 1394">The VLM's response conflicts directly with the human descriptions, which emphasize Chinese elements (hairstyle, architecture, lanterns) rather than Korean ones, and thus it fails to capture the key cultural details mentioned.</p>
 <p data-bbox="167 1688 357 1730">Artifact: Sami headwear Category: Hat</p>	 <p data-bbox="402 1688 570 1730">Region: Finland Super Category: Fashion</p>	<p data-bbox="613 1402 867 1423">Human Judgments (User Study):</p> <ol data-bbox="613 1430 1442 1556" 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 data-bbox="613 1562 906 1583">MLLM Judgment (Gemini 2.0 Flash):</p> <p data-bbox="613 1589 1476 1717">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> <p data-bbox="613 1724 857 1745">o1 critique of MLLM accuracy:</p> <p data-bbox="613 1751 1458 1793">The VLM response contradicts the human descriptions, which emphasize that the hat is not an accurate representation of Sami attire. It overlooks key details about the hat's shape, colors, and origins mentioned by the human observers.</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 19. Continued from previous page.

AI Image	Real Image	Feedback
 <p>Artifact: Takht-e Fulad Category: Monument/Memorial</p>	 <p>Region: Iran Super Category: Architecture</p>	<p>Human Judgments (User Study):</p> <ol style="list-style-type: none"> 1. 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 monument in Esfahan. 2. The coloring is different and Isfahanian buildings have different sets of colors and style more colorful. 3. 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> <p>o1 critique of MLLM accuracy:</p> <p>The VLM response contradicts the human descriptions by focusing on Egyptian rather than Iranian architecture, ignoring key details about Takhte-Jamshid and the stylistic elements discussed.</p>
 <p>Artifact: Portrait-Amir Kabir Category: Oil Painting</p>	 <p>Region: Iran Super Category: Art</p>	<p>Human Judgments (User Study):</p> <ol style="list-style-type: none"> 1. 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. 2. It has the right feature in the face but lack of detail in the outfit and background. 3. 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> <p>o1 critique of MLLM accuracy:</p> <p>While it mentions the Qajar era and general attire, it overlooks key cultural details (e.g., the conical hat style, necklace, correct beard/mustache style) that the human descriptions emphasize.</p>
 <p>Artifact: Kiping Category: Flatbread</p>	 <p>Region: Philippines Super Category: Food</p>	<p>Human Judgments (User Study):</p> <ol style="list-style-type: none"> 1. The image does not portray the meaning of Kiping at all. 2. The image shown was a type of an animal which is obviously so far from "Kiping" which is a type of food. 3. 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> <p>o1 critique of MLLM accuracy:</p> <p>The VLM's response precisely matches the human descriptions by stating that the image is of an animal (a cat) holding a spray bottle and has no connection to kiping, while also accurately describing kiping as a leaf-shaped, glutinous rice wafer. Thus, it captures all key cultural details mentioned by the humans.</p>
 <p>Artifact: Pulseira Category: Jewellery</p>	 <p>Region: Brazil Super Category: Fashion</p>	<p>Human Judgments (User Study):</p> <ol style="list-style-type: none"> 1. This is exactly how a pulseira is, a piece of jewellery that goes on the wrist 2. The image is very similar to the name of Pulseira. 3. 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 details, such as the braiding of the metal and the studded sphere, are consistent with what one would expect. 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> <p>o1 critique of MLLM accuracy:</p> <p>The VLM response fully captures the main points from the human descriptions, including that it is a wrist-worn pulseira/bracelet with shiny stones, aligning closely with all key details.</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)

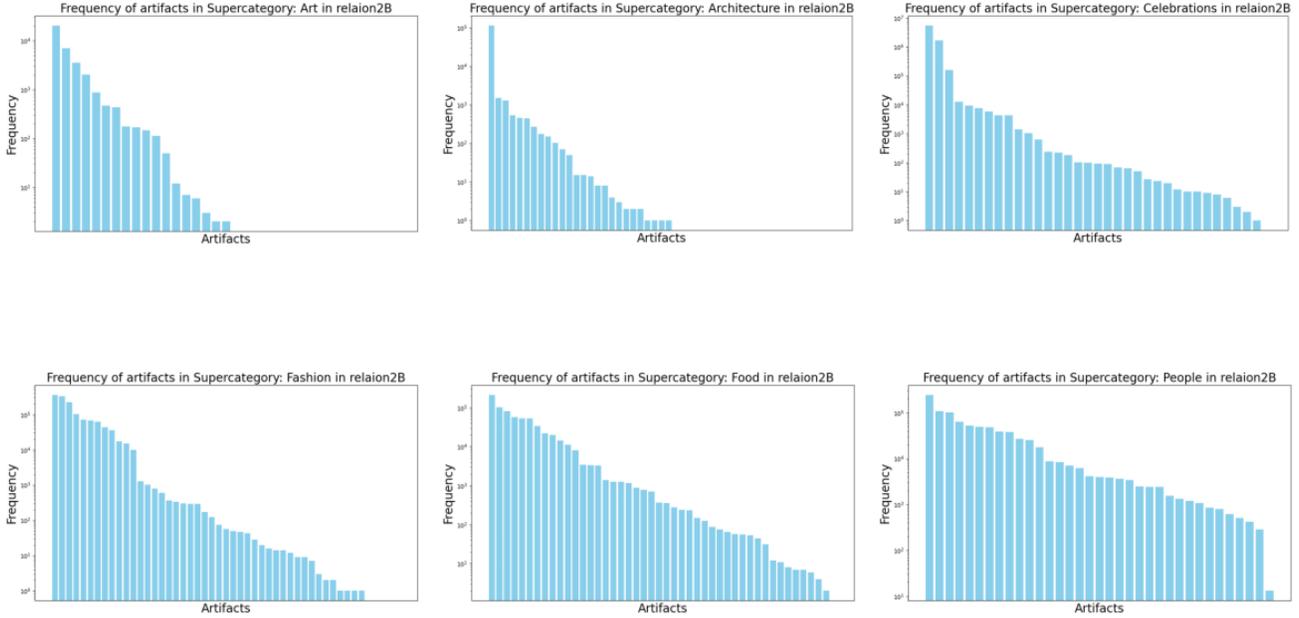


Figure 20. Concept Frequency Estimation for all artifacts across the six supercategories in the CuRe dataset on Re-LAION-2B.

K. Concept Frequency Estimation

In the modern era, 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-en [55] and LAION-Aesthetics V2 5+³, 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. Inspired by the string matching search system of Parashar et al. [43], we explicitly compute a concept frequency for every cultural artifact in the CuRe dataset on reLAION2B-en-research-safe⁴, the recently re-released version of the original dataset. As seen in Fig. 20, all supercategories in our CuRe dataset also show a long tail behavior in Re-LAION-2B across 50 cultural artifacts each. We observe especially low occurrence of many artifacts in art and architecture, 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 concept frequency with the ground truth likelihood Likert scores (see Appendix D for details) provided by survey respondents for FLUX.1 [dev], SD 3.5 Large, and SD 1.5 in Tab. 16. We observe that for Celebrations, Food, Fash-

ion, 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. 20). 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 16. Spearman rank correlation between occurrence frequency in Re-LAION-2B of each cultural artifact in the CuRe dataset with human perceptions of ground-truth likelihood from the user study. 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	0.37	0.59	0.31	0.21
SD 3.5 Large	0.05	0.12	0.44	0.55	0.23	0.24
FLUX.1 [dev]	0.01	0.06	0.33	0.37	0.40	0.36

Table 17. Spearman’s correlation of diversity scorers with frequency of artifact in reLAION2B dataset

³<https://laion.ai/blog/laion-aesthetics/>

⁴<https://huggingface.co/datasets/laion/relaion2B-en-research-safe>

T2I	Correlation w freq(n)	Supercategory					
		Art	Architecture	Celebrations	Food	Fashion	People
SD 1.5	LPIPS(n)	0.21	0.05	0.07	0.19	0.19	0.35
	$\phi_{DIV}(n)$	0.09	0.01	0.02	0.03	0.08	0.27
SD 3.5 Large	LPIPS(n)	0.09	0.28	0.23	0.22	0.13	0.11
	$\phi_{DIV}(n)$	0.11	0.32	0.09	0.43	0.16	0.23
FLUX.1 [dev]	LPIPS(n)	0.25	0.20	0.14	0.19	0.07	0.08
	$\phi_{DIV}(n)$	0.17	0.34	0.49	0.43	0.02	0.14