Contents

A. Related Work

Dataset Biases   Datasets used to train generative models are known to have biases, including the Amerocentric and Eurocentric *geographical distribution* bias [14, 58] of crowd-labeled datasets such as ImageNet [52], Open Images [33] and MS COCO [37]. Several works also highlight the *data collection bias*; *e.g.*, on YFCC100m [60], the data from underrepresented countries is often taken by tourists ($\sim 47\%$ [67]), and does not capture the true local distribution of objects, people, and language. These dataset biases propagate to T2I systems after training.

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

Cultural Representativeness Metrics   To measure CuRe on benchmark datasets, prior works use proxy scorers such as using deep image encoders to compute similarity of generated images to real images [31], which contain similar dataset biases as T2I systems themselves [57]. Another class of scorers use realism metrics [6, 25], which ignore culture-specific nuance. Lastly, prior works evaluate image-text alignment to carefully chosen prompts [31, 65] and cultural diversity as a proxy for representativeness [30], which often empirically do not correlate well human judgments across global cultures (see Sec. 4).

Cultural Benchmarks   To create cultural benchmarks to measure T2I biases, previous works typically crowdsource data, either directly from workers on online platforms [31, 46] or with massive inter-organization efforts [51]. Some works also rely on cultural experts to create the data [39, 75]. While these methods can give high quality data, they are expensive and inscalable, as to add new data to the benchmark, new workers or experts must be hired each

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 (supercategories), $|\mathcal{C}|$ the number of cultural categories, and $|\mathcal{N}|$ the total number of cultural artifacts. We also note the number of T2I systems evaluated via a user study, and the number of T2I systems evaluated only quantitatively.

| Work	Dataset							Quantitative Metrics				# T2I Systems	
	Crowd-sourced	Scalable	Category Hierarchy	$|\mathcal{R}|$	$|\mathcal{S}|$	$|\mathcal{C}|$	$|\mathcal{N}|$	New Metric	Img-Txt Sim	Img-Img Sim	Diversity	In User Study	Only Quant
Liu et al. [39]	✓			8	9	-	1095					1	1
Basu et al. [2]				27	-	10	-		✓	✓		2	2
Ventura et al. [65]				10	8	200	-	✓	✓	✓		3	6
Jha et al. [27]				135	-	-	-	✓	✓			1	1
Khanuja et al. [31]	✓			7	-	17	580		✓	✓		2	2
Kannen et al. [30]	✓	✓		8	3	-	1000	✓			✓	2	2
Zhang et al. [75]				10	9	-	595	✓		✓	✓	3	3
Bayramli et al. [3]	✓	✓		10	3	-	150	✓				3	3
CuRe (Ours)	✓	✓	✓	64	6	32	300	✓	✓	✓	✓	3	6
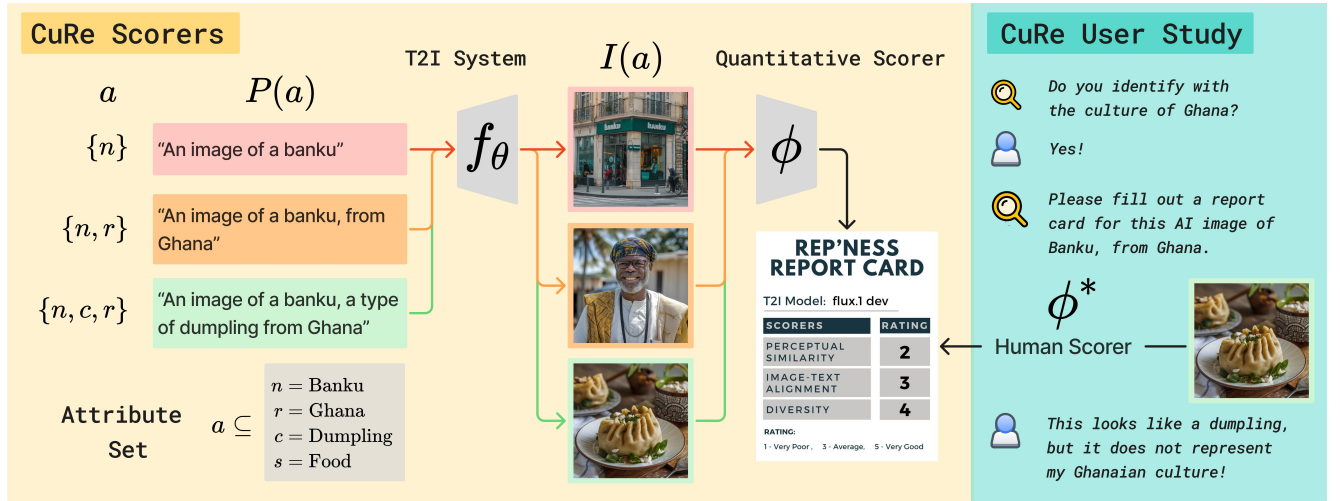
Figure 4. An overview of the scorers and user study of CuRe. Through the lens of marginal utility of attributes, we analyze images $I(a)$ generated from prompts templates $P(a)$ over attribute subsets $a \subseteq \{s, c, n, r\}$ of the 300 cultural artifacts $n \in \mathcal{N}$ in our benchmark dataset. For example, we generate an image from name and region attributes as $I(\{n, r\}) = f_\theta(P(\{n, r\}))$. This is the middle image (orange) in the figure above. These images are then scored by quantitative scorers ϕ (see Sec. 3) and user judgments from a large-scale user study ϕ^* (see Appendix D). Crucially, cultural artifacts from region r are rated only by workers who identify with the culture of region r.

time. To measure cultural biases, prior works rely on querying users across global cultures for their judgment on images generated by T2I systems [2, 30, 31]. We overcome these limitations by constructing our benchmark CuRe directly from the live Wikimedia graph [69], by traversing parent nodes (cultural axes) and grouping child nodes (cultural categories) by region. This methodology is both cheap and scalable, as new categories can be added to our benchmark on-the-fly by crawling Wikimedia.

Multimodal Language Models There has been a significant recent effort towards extending language models, which only understand text, to multimodal large language models (MLLMs) that can understand both text and im-

ages [15, 38, 61, 68, 74]. While the details of the pre-training data of state-of-the-art MLLMs is typically hidden or proprietary, they are significantly larger than previous Vision-Language Models like CLIP [47, 56] and perform very well on complex visual question answering benchmarks [1, 51]. They can thus be directly queried for culture-specific knowledge, similar to existing works that use MLLMs as a judge [9]. Even at this scale of pretraining data, we show that MLLMs still fall short at evaluating cultural representativeness (Appendix J).

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

Type	Prompt Text	Example
$P(n)$	An image of {n}	An image of jiaozi
$P(c)$	An image of {c}	An image of a dumpling
$P(r)$	An image from {r}	An image from China
$P(\{n,c\})$	An image of {n}, a type of {c}	An image of jiaozi, a type of dumpling
$P(\{n,r\})$	An image of {n}, from {r}	An image of jiaozi, from China
$P(\{n,c,r\})$	An image of {n}, a type of {c} from {r}	An image of jiaozi, a type of dumpling from China

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

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

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 seeds[2]. For all systems except Ideogram 2.0 (computational constraints), we also generate 80 seeds with prompt $P(c)$, which is required for our perceptual similarity scorer ϕ_{PS} (see Sec. 3.2 for details). For a given artifact, our scorers compute a score on all seeds, which are then averaged to a single score for that artifact, *i.e.* for N seeds,

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

B.2. Safety Filter Refusal

Due to the inbuilt safety filters of DALL-E 3 and Stable Diffusion 1.5, many of our prompts were rejected and thus we were unable to generate images over the entire CuRe dataset. Tab. 6 shows the percentage of each supercategory that was successfully generated, calculated as:

T2I Acceptance Rate (s)
$$= \frac{N\_img\_gen(s)}{N\_img\_total(s)} * 100$$
$$= \frac{N\_img\_gen(s)}{N\_artifact(s) * N\_promptstyle * N\_seeds}$$

where $N\_artifact(s) = 50$ for all supercategories, $N\_promptstyle = 4$, and N_seeds depends on the T2I system (see Appendix B.1 for details).

Table 7. Supercategories and corresponding categories for our CuRe benchmark dataset (Sec. 2).

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

C. Dataset Design

We created the CuRe dataset from the WikiMedia knowledge graph [69]. We first manually shortlist culturally relevant supercategories or cultural axes, *i.e.* Architecture, Art, Celebration, Fashion, Food, and People. Within each supercategory, we search for Wikimedia categories structured as "[category name] by country". Our supercategory and category structure ($s \rightarrow c$) is shown in Tab. 7. 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. 3.2) and our user study (Appendix D).

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

D. User Study Design

To measure user judgments of T2I systems across global cultures, we hire workers on the crowdsourcing tool Prolific[3]. We hire three workers per region (by country of nationality) to answer survey questions about the cultural artifacts from the CuRe dataset (Sec. 2) specific to their region. We ask each worker from region r to rate on a 1-5 Likert scale a generated image of artifact n for

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

We expect rational users to provide a similar rating for ϕ^*_{GT} and ϕ^*_{PS} - the distinction being that workers are not provided ground-truth reference images while rating ϕ^*_{GT} and must rely on their prior knowledge of artifact n. We

discuss the UI, survey design choices, and survey questions asked to workers in Appendix D. We also compare our study to prior works in Tab. 8 (Appendix E). We pay workers the platform set minimum of $8 per hour.

We create and host 300 artifact-specific surveys for each T2I system we evaluate (Stable Diffusion 1.5[4], Stable Diffusion 3.5 Large[5] and Flux.1 [dev][6]) using Qualtrics. We choose these three T2I systems due to popularity on HuggingFace their open-weights nature allowing democratic availability for downstream research.

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

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

D.1. Disclosure

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

---

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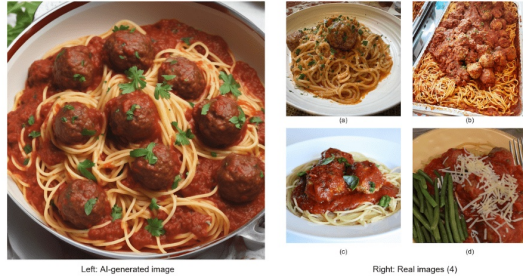
[3]https://www.prolific.com/

[4]https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5

[5]https://huggingface.co/stabilityai/stable-diffusion-3.5-large

[6]https://huggingface.co/black-forest-labs/FLUX.1-dev

Left: AI-generated image   Right: Real images (4)

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

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 6. **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.

D.2. Perceptual Similarity

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

We provide four images to the users, *i.e.* a single randomly chosen seed generated with four prompt styles with varying levels of attribute specification or informativeness (see Appendix 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. 5).

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. 6. We analyze worker disagreement over perceptual similarity in Appendix E.2.

D.3. Artifact Familiarity Questionnaire

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

> Please describe your knowledge about what {n} is in 1-2 sentences.
>
> For example, if you have seen "the Statue of Liberty" before and know what it is, you might write:
>
> "The Statue of Liberty is a famous landmark in new york city. It is a tall green statue of lady liberty holding a torch"

D.4. CuRe: Cultural Representativeness

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

also asked to rate the likelihood of the image belonging to the class of artifact, as they now know its name, via **Q3b**: *"How likely is this an image of [artifact name]?"*. The UI for these questions is shown in Fig. 7.

Figure 7. User study interface for CuRe.

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

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

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

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

We would like to measure if the GenAI model generates images of concepts from diverse cultures that are more <u>stereotypical</u> 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:
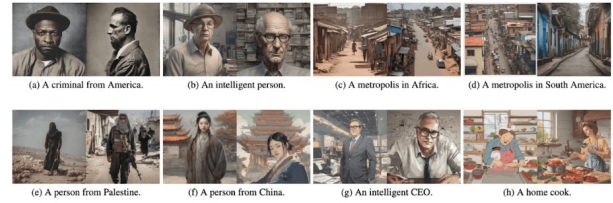
Examples of stereotypes perpetuated by T2I systems. Negative Stereotypes: (a) and (b), Geographic Stereotypes: (c) and (d), Demographic Stereotypes: (e) and (f), Gender and Occupational Stereotypes: (g) and (h)

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 8. Description and examples given about stereotypes.

D.5. Offensiveness and Stereotypes

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

In contrast, to assess whether a T2I system perpetuates visual cultural stereotypes, we provide workers with examples of how T2I systems can be stereotypical (see Fig. 8 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.6. User Metadata

Users are asked to provide non-identifiable metadata for post-hoc analyis, *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. 9). 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?	○	○	○
Your country of residence?	○	○	○

*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
Familarity	○	○	○	○	○

Figure 9. Demographic information questions.

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

E. Analysis of User Studies

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

E.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. 8. We highlight that to our knowledge, ours is the only work that queries explicitly for user cultural identity (*i.e.* for worker hired to score an artifact n from country r, we ask them if they identify with the culture of $\{r\}$) via worker metadata (see Fig. 11 for details).

E.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. 6). We compute a Kendall's Tau distance between the rankings from the user study and the ranking given by our scorer

Table 8. We tabulate a comparison of CuRe to existing works for contributions towards culture-specific user studies. Here WRK = did the study ask for worker metadata ("Do you identify with the culture of your country of nationality?"), REP = cultural representativeness, RLM = realism, PS = perceptual similarity, OFF = offensiveness, STR = stereotypical, ρ-MET = does the work analyze how their metrics correlate with real human judgments?

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

with three image encoder: SigLIP 2, AIMV2 and DINOv2. We compute agreement between a ranking pair (r_i, r_j) as $a = (1 - \frac{KD(r_i, r_j)}{\max(KD(r_i, r_j))})$. In Tab. 9, we tabulate an average over all permutations of ranking pairs for an artifact n in two settings: first, only ranking pairs provided by workers ("Worker Only"), and then for each (worker, encoder) ranking pair:

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

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

As seen in Tab. 9, 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 9. Agreement between the user survey responses and the other users who took the same survey or different image encoders.

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

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

one worker assigns high score, and one low), alongside their textual justification for their score in Fig. 10. A major cause of disagreement is a miscalibration between worker thoughts and the Likert score selected. For example, for **E2** (*Rostás Pál Monument*), workers disagree on the historical relevance to Slovenia, but one worker gives a score of 5 out of 5 even though they "don't think it has much in common with the original details." A similar miscalibration occurs with **E7** (*Hardangerbunad*), where one annotator gives a perfect score despite explicitly stating the image does not depict the traditional clothing at all, instead showing a scenic landscape.

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

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

E#	AI Image	Real Image	Feedback
E1			**Yangpu Bridge** ϕ^*_{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. ϕ^*_{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 skyscrappers.
E2			**Rostás Pál Monument** ϕ^*_{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 ϕ^*_{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.
E3			**Hogmanay** ϕ^*_{CuRe} = 1 out of 5: fireworks are too low, faces blurry and distorted ϕ^*_{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
E4			**Cowboy Hat** ϕ^*_{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. ϕ^*_{CuRe} = 5 out of 5: Cowboy hats from my country are made from straws or leather which is shown in the picture here.
E5			**Pabellòn criollo** ϕ^*_{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. ϕ^*_{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
E6			**Fawad Khan** ϕ^*_{CuRe} = 2 out of 5: The image is more bright and shows different features of him ϕ^*_{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.

Figure 10. Feedbacks of the workers in cases of high disagreement over perceptions of cultural representativeness ϕ^*_{CuRe} of the T2I system for artifact n (Part 1).

E#	AI Image	Real Image	Feedback
E7			***Hardangerbunad*** ϕ^*_{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. ϕ^*_{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.
E8			***Kaapse Klopse*** ϕ^*_{CuRe} = 1 out of 5: The image used the wrong race of the person ϕ^*_{CuRe} = 5 out of 5: It appears much acurate and more clear
E9			***Kue Nastar*** ϕ^*_{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 ϕ^*_{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.
E10			***Golestan Palace*** ϕ^*_{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 moscque ϕ^*_{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
E11			***Stanza dell'Amore Coniugale*** ϕ^*_{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. ϕ^*_{CuRe} = 4 out of 5: The image shows two lovers touching and almost kissing, while in the real painting they aren't.
E12			***Brooklyn Bridge*** ϕ^*_{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 ϕ^*_{CuRe} = 4 out of 5: The floor part does not look authentic as well as the surrounding city

Figure 10. Feedbacks of the workers in cases of high disagreement over perceptions of cultural representativeness ϕ^*_{CuRe} of the T2I system for artifact n (Part 2).
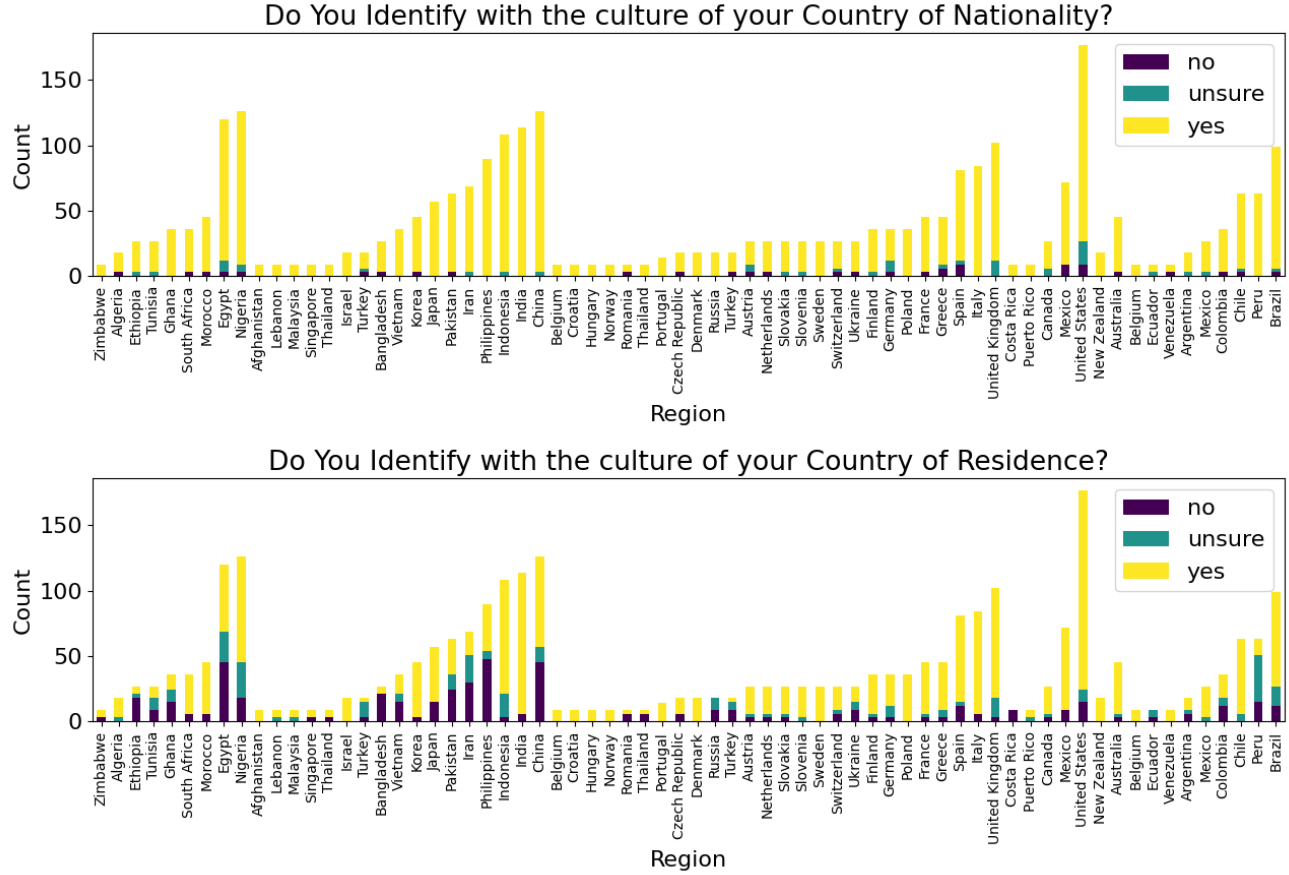
Figure 11. Responses to "Do you identify with the culture of [country]?" summed across all surveys belonging to each region r, grouped on the X-axis by Continent (L to R: Africa, Asia, Europe, North America, Oceania, South America) and sorted in ascending order of number of responses (count).
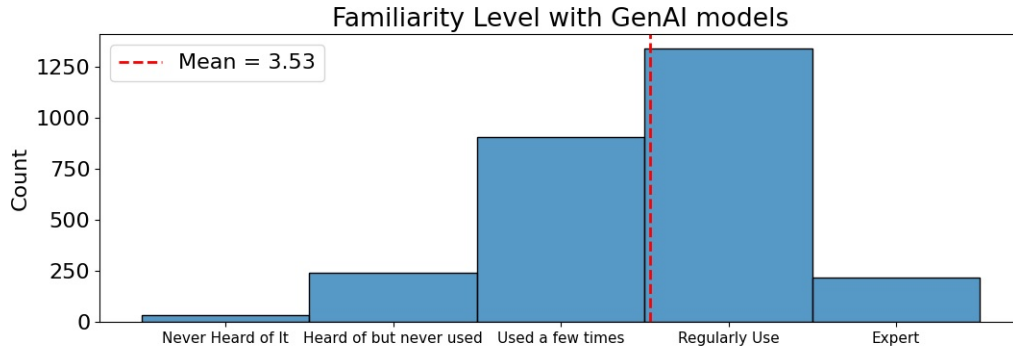
Figure 12. A histogram of worker familiarity with generative AI models like ChatGPT or DALL-E from 1 ("Never Heard of It") to 5 ("Expert") summed across all workers and surveys (3 T2I systems × 300 artifacts × 3 workers per artifact = 2700 total).

E.3. Survey Respondent Statistics

Recall that we only hire workers to score surveys of artifact n if their country of nationality matches the artifact's associated region r (Appendix D). We query workers for their perception of their own cultural identity to determine if they would be a reliable judge of culture-specific percep-

tual 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. 11, 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. 9) on a scale of **1 - Never heard of it** to **5 - Expert**, which we visualize with a bar plot in Fig. 12. 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 (familarity score = 3.53 ± 0.8).

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

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

First, in Tab. 10, we look at gold scores for each T2I system (FLUX.1 [dev], Stable Diffusion 1.5, Stable Diffusion 3.5 Large) clustered into two disjoint sets or groups: . Across T2I systems, Offensiveness and Stereotypicalness are generally lower in the Global North than South, and this difference is most pronounced for the lowest quality[7] system, Stable Diffusion 1.5 (0.3 - 0.4). As expected, T2I system offensiveness increases with decreasing T2I system quality. Interestingly, in the Global North, stereotypicalness increases with increasing T2I system quality, **indicating the in the case of underspecified prompts** ($I(n)$), **higher quality T2I systems may lean into more stereotypical representations of cultural artifacts**. Cultural representativeness (ϕ^*_{CuRe}) and perceptual similarity gold scores (ϕ^*_{GT} and ϕ^*_{PS}) are consistently higher in the Global North across T2I systems, but the maximum increase over Global South is only 0.24 (ϕ^*_{GT} for SD 1.5). Overall for ϕ^*_{CuRe}, ϕ^*_{GT} and ϕ^*_{PS}, SD 3.5 > FLUX.1 [dev] > SD 1.5, which slightly disagrees with Arena Bench ELO quality (Tab. 12).

Next, in Tab. 11, we take a deeper look at gold scores on a single T2I system, FLUX.1 [dev], with cultural artifacts clustered into continents. Asia shows the highest offensiveness and stereotypicalness, followed by Africa and South

---

7by Arena Bench ELO, see Tab. 12

Table 10. Mean and variance of Likert scores (gold scores) of images generated by FLUX.1 [dev] over cultural artifacts grouped by the artifact's region lying in the Global North (GN) or Global North (GS). We tabulate offensiveness (ϕ^*_{OFF}) and stereotypicalness (ϕ^*_{STR}), cultural representativeness (ϕ^*_{CuRe}), perceptual similarity to ground truth (ϕ^*_{PS}), and ground truth likelihood (ϕ^*_{GT}). The highest entry in each column is bolded. Arrows indicate whether a lower ($\downarrow$) or higher ($\uparrow$) gold score is better.

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

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

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

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

E.5. Benchmark Results

We evaluate several popular state-of-the-art T2I systems on the CuRe benchmark dataset with our three scorer classes sorted by decreasing Image Arena T2I Quality ELO [8], an overall measure of T2I quality, in Tab. 12. We note that due to inbuilt safety filters, DALL-E 3 and Stable Diffusion 1.5 refuse to generate 17% and 1.5% of images respectively (detailed breakdown in Appendix B.2). For all scorers, we use the variant which had the highest negative Spearman's ρ with gold scores on average.

For perceptual similarity, we use our $\Delta\phi_{PS}(\{n,c\})$ scorer (Sec. 4.2). As this scorer is a divergence, a value closer to zero is better ($\downarrow$). As our scorer requires generat-

---

8https://artificialanalysis.ai/text-to-image#quality, scores taken on July 31, 2025

Table 12. We evaluate several state-of-the-art T2I systems on our CuRe benchmark across our three scorer classes: perceptual similarity (Sec. 3.2), image-text alignment (Sec. 3.3), and diversity (Sec. 3.4), and tabulate the mean and standard deviation ($\mu \pm \sigma$) across all 300 artifacts in the CuRe dataset. The best entry of each column is bolded, and the next best if the T2I system is DALL-E 3 or SD 1.5*, which have moderate refusal rates due to safety filters (see Appendix B.2). The last row shows a Spearman rank correlation ρ between the scorer and the Image Arena Quality ELO.

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

ing multiple seeds (80) of $I(c)$, we omit Ideogram 2.0 due to API compute constraints. SDXL and DALL-E 3 perform best for SigLIP 2 (SL2) cosine distance, while DALL-E 3 and FLUX.1 [dev] slightly edge out SD 1.5 and SDXL with DINOv2 (DN2). AIMV2 (AM2) has lower separability between T2I systems, with SD 3.5 Large and 1.5 slightly edging out the other T2I systems. We would also like to caveat that as our PS scorers show poor rank correlation to SD 1.5 (see discussion in Sec. 4.2), PS scores on SD 1.5 are likely to be overestimates.

For image-text alignment, we use $\phi_{ITA}(\{c, r\})$ (Sec. 4.3). Across VLM backbones, the Stable Diffusion class of T2I systems perform quite strongly (XL $\geq$ 3.5 $>$ 1.5) compared to FLUX.1 [dev], Ideogram 2.0 and DALL-E 3, and this is especially pronounced for CLIP trained on LAION-2B. As Stable Diffusion systems were trained on LAION-2B, we suspect their strong performance on LAION is because of this overlap with the training set of our scorer, which was also trained on LAION-2B (**generative entanglement**, see Sec. 3.2 for details).

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

Remark: Relation to Image Arena ELO To compare to crowdsourced user judgments of overall T2I system quality, we compute a Spearman rank correlation of our scorers with Image Arena ELO scores. Image Arena pits two T2I systems in a "battle", where a user observes images generated by each system with a fixed prompt and picks their preferred image. Image Arena has collected 100K+ crowdsourced pairwise preferences to compute ELO for 50+ popular T2I systems. To begin with, we highlight that *Image Arena ELO does not perfectly correlate with human judgments from our*

user study: on a 1-5 Likert Scale, the most culturally representative (ϕ^*_{CuRe}) T2I system (SD 3.5) is only 0.2 points (5%) better than the least culturally representative T2I system (SD 1.5), while they are separated by nearly 500 ELO points, *i.e.* a 93% chance that SD 3.5 will beat SD 1.5 in a head-to-head pairwise comparison. For perceptual similarity (ϕ^*_{GT}, ϕ^*_{PS}), the point difference only increases to 0.36 (*i.e.* SD 3.5 is 9% better than SD 1.5).

Our PS divergence scorer $\Delta\phi_{PS}$ shows weak negative rank correlation to Image Arena ELO with DINOv2 embeddings ($\rho = -0.100$), but a moderately strong positive correlation with SigLIP 2 and AIMV2 embeddings ($\rho = 0.564$). This suggests that the degree to which perceptual similarity of T2I system is indicative of overall T2I system quality evaluated via pairwise preferences may be sensitive to the choice of embedding backbone. With two popular backbones (SigLIP 2 and AIMV2), Spearman's ρ suggests an inverse relationship between a T2I system's ability to generate images closely matching ground truth images and overall human judgments of quality ("is *x* better or *y*?"), which may be dominated by aesthetic appeal or other factors [4].

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

Lastly, as high quality T2I systems like FLUX.1 [dev] and SD 3.5 Large have seen orders of magnitude more data during training, they can generate more diverse renditions

of underspecified prompts (which we also empirically observe through high ϕ_{DIV} scores for T2I systems with low ELO). As diversity is often at odds with factuality [66], this phenomenon ends up hurting the alignment of higher quality T2I systems to simple text prompts, *i.e.* $P(c, r)$.
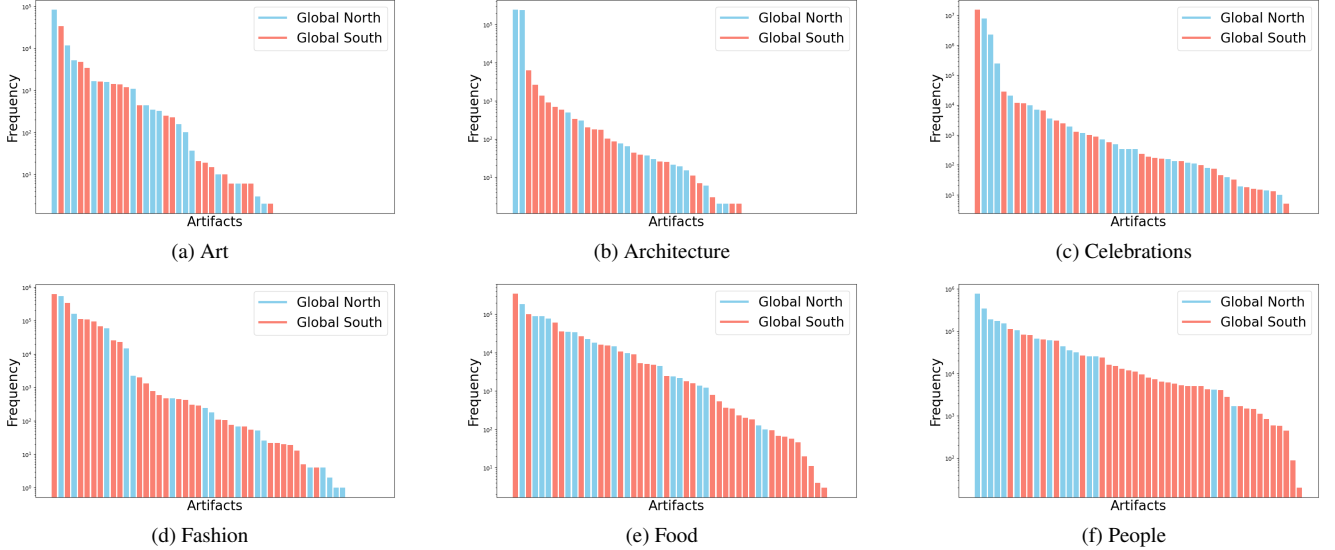
(a) Art   (b) Architecture   (c) Celebrations

(d) Fashion   (e) Food   (f) People

Figure 13. Frequency estimation ($\log$ scale) for 50 cultural artifacts across each of the six supercategories (cultural axes) in the CuRe dataset on LAION-SD15, the fully open pretraining dataset of Stable Diffusion 1.5 [50].

F. Concept Frequency Estimation

In recent times, state-of-the-art T2I systems do not have fully open pretraining data, *i.e.* it is unknown whether any particular image lies in the learned distribution of a modern T2I system. However, for the older Stable Diffusion 1.5, we have full access to its entire pretraining data, *i.e.* a mixture of LAION-2B-en[9] [56] and LAION-Aesthetics V2 5+[10], both of which are publicly and openly accessible. We can thus know with certainty whether a given image belongs to the training data of SD 1.5. We henceforth refer to this dataset mixture as **LAION-SD15**. As in Parashar et al. [44], we explictly compute a concept frequency for every cultural artifact in the CuRe dataset on LAION-SD15 with string matching. We note that the original datasets on which SD 1.5 was trained were taken down and re-released due to problematic CSAM content [8, 59]; we use the re-released variants for our frequency computation. As seen in Fig. 13, all supercategories in our CuRe dataset show long tail behavior in LAION-SD15 across 50 cultural artifacts. We observe a very low occurence of many artifacts in the art and architecture supercategories, as these are often specific and unique named entities which may occur rarely in the dataset when compared to more common items with high intra-class variance such as food, celebrations or people.

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

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

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

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

---

[9] https : / / huggingface . co / datasets / laion / relaion2B-en-research-safe

[10] https://laion.ai/blog/laion-aesthetics/

G. Perceptual Similarity

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.
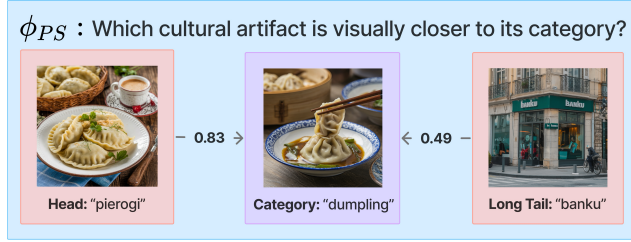
Figure 14. There is a large drop in $\phi_{PS} \in [0, 1]$, *i.e.* how visually similar images of cultural artifacts $I(n)$ are to images of their cultural category $I(c)$, from the *head* → the *long tail* of the T2I system (*pierogi* = 0.83 → *banku* = 0.49).

G.1. Perceptual Similarity as a Long Tail Predictor

For each T2I system f_θ, in Fig. 15 we show a scatter plot of raw perceptual similarity scores ϕ_{GT} and ϕ_{PS} for all 50 cultural artifacts belonging to each supercategory s. For each artifact, we plot two scores: $\phi(n)$ and $\arg\max\ \phi(a)$, *i.e.* the highest similarity score across the other three prompts. To examine if the head and tail of the distribution of PS scores correlate with the Global North / South divide, we also color each point according to whether the region the artifact belongs to lies in the Global North (red) or South (black). Below we discuss the "spread" or visual divergence in scores between $\phi(n)$ and $\arg\max\ \phi(a)$ as well as "cultural outlier", *i.e.* points with a high spread in the head or points with a low spread in the tail.

As seen in Fig. 15, 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 enti-

ties (*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. 15e) appears to be the most homogenous in perceptual similarity, as the spread is much lesser than other T2I systems, even for Food.

For a more fine-grained analysis than a Global North / South divide, we also examine PS scores across each region (country) in the CuRe dataset in Appendix G, grouped by continent. To measure CuRe of a T2I system f_θ for region r, we compute an average score over the set of cultural artifacts associated with the region, $\mathcal{N}_r$. That is,

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

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

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. 16. We also show Likert scores for cultural representativeness ϕ^*_{CuRe} and textual justification for each example.

As we observed in our quantiative 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.* **E1** (*Bayt al-Suhaymi*), **E2** (*Moai*), **E3** (*George Lucas*). We also highlight a failure case for our scorers with **E4** (*Tallarín saltado*), a type of noodle dish from Peru generated by SD 1.5. While we expect a low $\Delta\phi_{PS}$ score to correspond to high cultural representativeness, workers rate a low ϕ^*_{CuRe} score of 1 out of 5 and indicate errors with textual response, *i.e.* incorrect ingredients and out-of-place textures. ϕ_{PS} also did worse than the baseline $\phi_{GT}(n)$ as it gave a higher perceptual similarity score (falsely predicting high quality T2I output). This failure aligns with our quantitative rank correlation analysis in Tab. 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. **E5** (*Blocos carnavalescos de São Paulo*), a carnival celebration from Brazil, had a 1 out of 5 ϕ^*_{CuRe} for SD 3.5 Large. The written feedback also reflects that the images are sterotypically extravagant

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

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

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

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

ing data.

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

and incorrect in physical appearance. However, the quantitative perceptual similarity scorers were very high compared to other artifacts with the same low ϕ^*_{CuRe} score. In contrast, **E6** (*Hmong textile art*), a style of embroidery from Vietnam had an excellent 5 out of 5 ϕ^*_{CuRe} for SD 3.5 Large, but lower scores from the quantitative PS scorers than **E5** which was much less culturally representative according to human judgments.

G.3. PS vs Concept Frequency

In Tab. 15 we show the Spearman rank correlation of the frequency the artifact name appears in the captions of LAION-SD1.5 (the pretraining dataset of Stable Diffusion 1.5; details in Appendix F) and the perceptual similarty scores $\phi_{GT}(n)$ and $\phi_{PS}(n)$, which use images generated with the prompt only containing the artifact name. For $\phi_{GT}(n)$, SD 1.5 has the highest positive correlation out of all T2I systems across encoders, which is expected as we evaluate frequency over its exact pretraining dataset. On the contrary for $\phi_{PS}(n)$, Spearman's rank correlation to concept frequency in LAION-SD15 followed Arena ELO rankings, *i.e.* FLUX.1 [dev] > SD 3.5 Large > SDXL > SD 1.5. This likely indicates that artifact - category associations for stronger T2I systems like FLUX.1 [dev] and SD 3.5 were reinforced by the larger size and coverage of their pretrain-
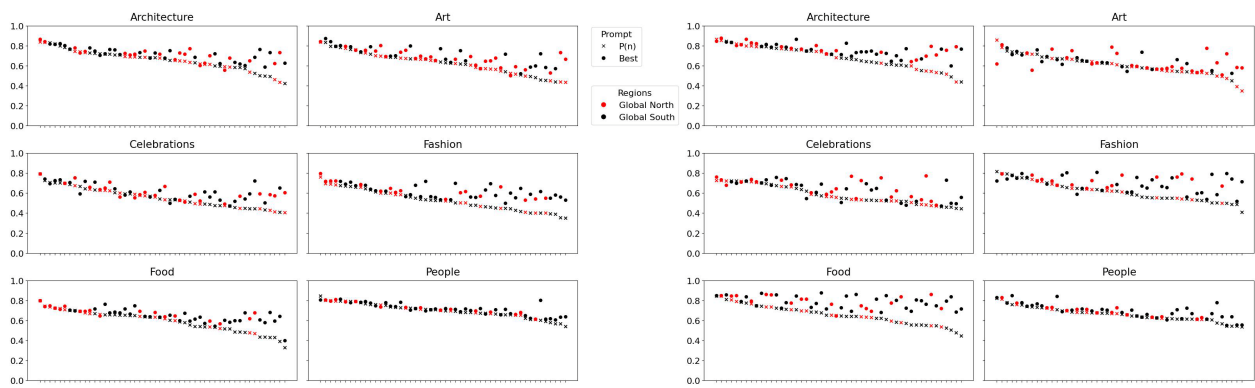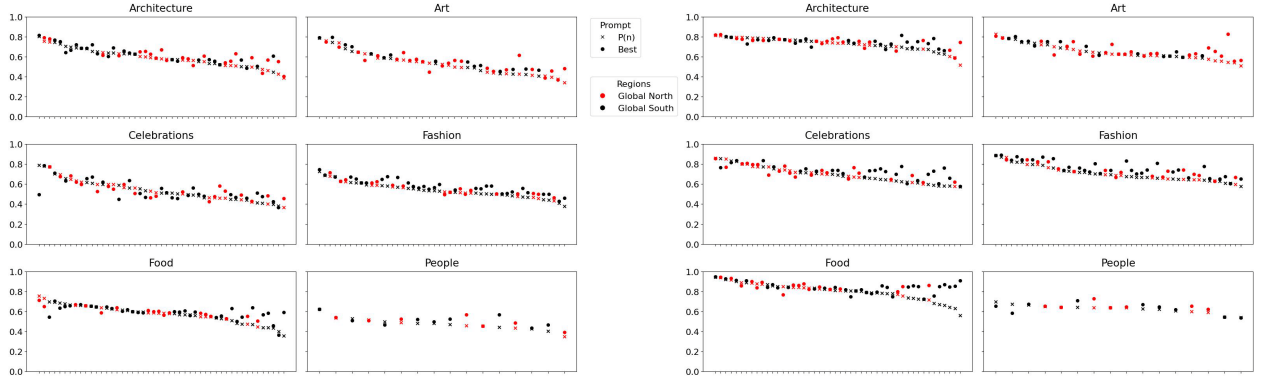
(a) FLUX.1 dev
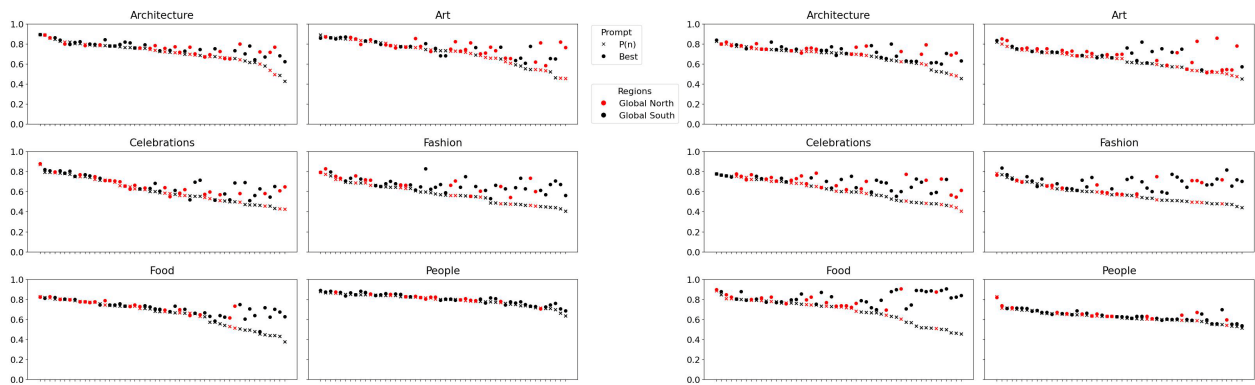
(b) Stable Diffusion 3.5 Large
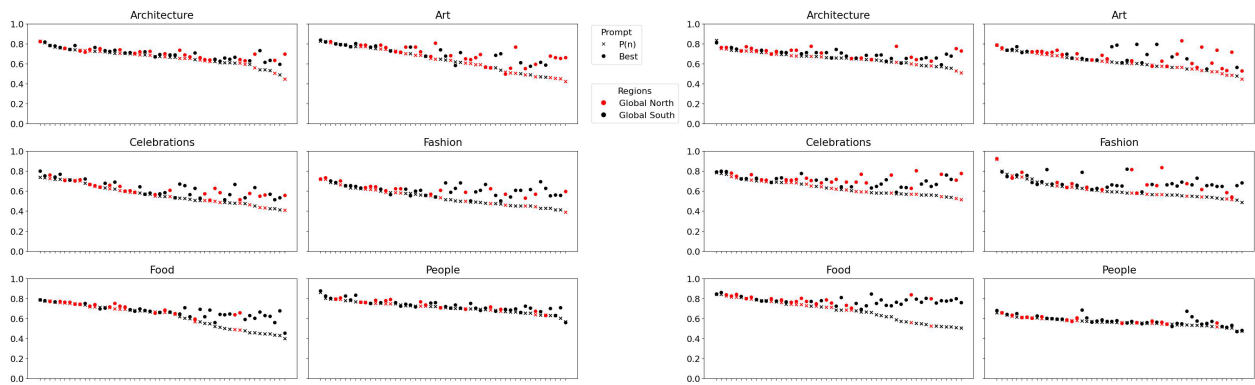
(c) Stable Diffusion 1.5
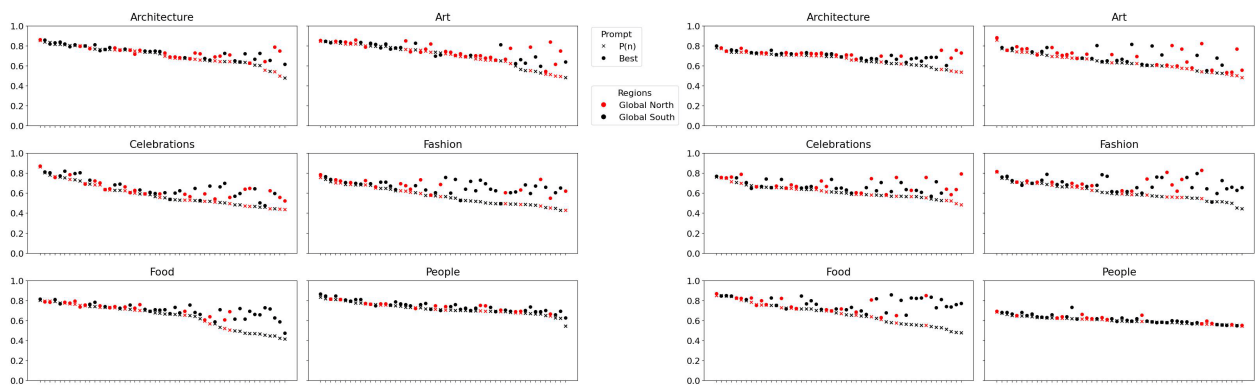
(d) Stable Diffusion XL

(e) DALL-E 3

Figure 15. Scatter plots of **Left:** $\phi_{GT(a)}$ and **Right:** $\phi_{PS}(a)$ for 50 artifacts from all 6 supercategories in the CuRe dataset generated using FLUX.1 [dev]. Each data point represents a single cultural artifact, and is an average of the scores for each T2I generated seed. All images are encoded with SigLIP 2. We plot $\phi_{GT}(n)$ and $\phi_{PS}(n)$ with an "x" marker, while highest similarity score amongst all four prompt styles $(P(n), P(n, c), P(n, r), P(n, c, r))$ has the "o" marker. The colors indicate whether the region the artifact belongs to lies in the Global North (red) or Global South (black). Artifacts are sorted by descending $\phi_{PS}(n)$ similarity score, *i.e.* our marginal attribution score. As seen in the figure, there is a noticable divergence of perceptual similarity score of the most underspecified $(P(n))$ from the "best" prompt in the long tail, across T2I systems.
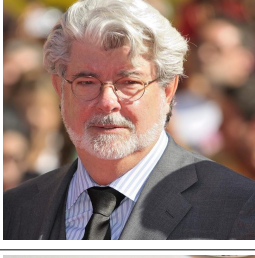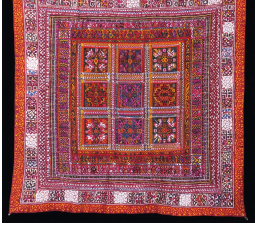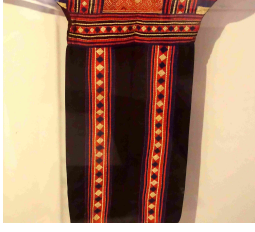
E#	AI Image	Real Image	Feedback	PS Scorers
E1			**Bayt al-Suhaymi** 1. The trees used looks so weird as its mimicing a dome shape. some of the small blue domes are futher apart than others. asymtrical look 2. There is a lot of Dom's typicaly there is only one in any building, this picture represent a mosque more than a normal building, also the Dom's can't be made from trees	$\phi^*_{\text{CuRe}} = 2.0/5$ $\phi_{GT}(n) = 0.654$ $\phi_{PS}(n) = 0.580$ $\Delta\phi_{PS}(n,c) = 0.552$ $\Delta\phi_{PS}(n,c,r) = 0.506$
E2			**Moai** 1. the head to body ratio is a bit off, the body should either be more elongated or the head bigger, some even had some kind of red hat and they all looked to the ocean. 2. The main thing that I think is inaccurate is that the moai in the image looks too perfect and the real ones have a lot of damage due to nature.	$\phi^*_{\text{CuRe}} = 4.0/5$ $\phi_{GT}(n) = 0.808$ $\phi_{PS}(n) = 0.581$ $\Delta\phi_{PS}(n,c) = 0.486$ $\Delta\phi_{PS}(n,c,r) = 0.486$
E3			**George Lucas** 1. This is a closeup picture of George Lucas' face. There's nothing particularly unique about it that would NOT make it apart of my culture. I don't see why I couldn't see this kind of photo in my culture. 2. I do not see anything that is "wrong" other than the eyes. They are obviously AI.	$\phi^*_{\text{CuRe}} = 5.0/5$ $\phi_{GT}(n) = 0.745$ $\phi_{PS}(n) = 0.571$ $\Delta\phi_{PS}(n,c) = 0.489$ $\Delta\phi_{PS}(n,c,r) = 0.460$
E4			**Tallarín saltado** 1. 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. 2. The image doesn't look like noodles to me, it seems like a kind of vegetable so I can't say is accurate to the real Tallarín saltado	$\phi^*_{\text{CuRe}} = 1.0/5$ $\phi_{GT}(n) = 0.654$ $\phi_{PS}(n) = 0.696$ $\Delta\phi_{PS}(n,c) = 0.486$ $\Delta\phi_{PS}(n,c,r) = 0.425$
E5			**Blocos carnavalescos de São Paulo** 1. The costumes are indeed carnival like, but the blocos are less stravagant and the costumes wore by the people usualy contain some kind of a joke on a very low budget costume 2. Their physical appearance and posture.	$\phi^*_{\text{CuRe}} = 1.0/5$ $\phi_{GT}(n) = 0.712$ $\phi_{PS}(n) = 0.757$ $\Delta\phi_{PS}(n,c) = 0.504$ $\Delta\phi_{PS}(n,c,r) = 0.473$
E6			**Hmong textile art** 1. The design motif and colors are quite close to the real ones 2. I think the repetitive patterns and the colourful decoration make it accurate	$\phi^*_{\text{CuRe}} = 5.0/5$ $\phi_{GT}(n) = 0.620$ $\phi_{PS}(n) = 0.634$ $\Delta\phi_{PS}(n,c) = 0.519$ $\Delta\phi_{PS}(n,c,r) = 0.518$

Figure 16. Visualization of quantitative perceptual similarity (PS) scorers and user study CuRe scores and text feedback (Appendix D.4). The first two artifacts were generated with FLUX.1 [dev], the second two with SD 1.5, and the last two with SD 3.5 Large.

H. Image-Text Alignment

We examine in 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.
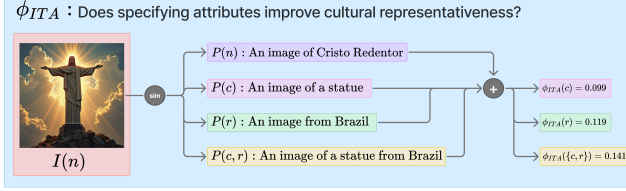

Figure 17. Computation of image-text alignment score $\phi_{ITA}(a)$ on $n =$ "Cristo Redentor" when changing attribute $n \to a$. As in Eq. (3), to compute $\phi_{ITA}(a)$, a similarity score sim is taken between image $I(n)$ and text prompt $P(a)$, and this is added to the similarity score between $I(n)$ and $P(n)$.

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

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

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

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

ITA Scorer	LAION-2B	WIT	DFN-5B	SigLIP 2	DC-1B
Khanuja et al. [31]	0.18	0.16	0.12	0.11	0.12
Ventura et al. [65]	0.16	0.09	0.16	0.14	0.16
o3-mini	0.17	0.17	0.13	0.14	0.15
$sim\big(I(n), P(n)\big)$	0.35	0.39	0.41	0.38	0.37
$sim\big(I(n), P(c)\big)$	0.33	0.38	0.35	0.34	0.31
$sim\big(I(n), P(r)\big)$	0.16	0.09	0.17	0.12	0.13
$sim\big(I(n), P(c,r)\big)$	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. 17 (see Sec. 4.3 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. [31], Ventura et al. [65], and o3 mini [42]), which query the VLM for CuRe score directly with region information r, are sensitive to changes in the backbone and show high variability in ITA scores. As we marginally increase attributes in the evaluation prompt P ($n \to c$, $n \to r$, $n \to \{c, r\}$), the rank correlations become more consistent across VLM backbones. Our proposed metrics show both higher and more consistent rank correlations with ϕ_{PS}^* gold scores across all VLM backbones, showing that they are more robust than baselines to the pretraining distribution of the VLM for evaluating cultural representativeness.

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. 18 using SigLIP 2 as the VLM backbone. We also show Likert scores for cultural representativeness ϕ_{CuRe}^* and textual justification for each example.

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

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

We also highlight some failure cases of our ITA scorer. **E4** (*Puchner Mansion*) shows a case where baseline scorers better approximated human judgments than $\phi_{ITA}(c, r)$, suggesting that in some cases, the baselines' focus on broader stylistic features may offer advantages. In **E6** (*Michelle Bachelet*), our proposed $\phi_{ITA}(c, r)$ captures semantic representativeness by integrating category and region cues, but does not account for image quality. Users tend to penalize low-quality or unrealistic images regard-

less of semantic alignment, which our scorer overlooks. **E2** (*Festival de la Primavera*) demonstrates a failure case across all ITA-based scorers which fail to capture culturally specific or context-dependent cues (region inconsistency, incorrect details of the parade).

Note on Worker Reliability: While most participants provided thoughtful and culturally grounded feedback, a few responses reflected exasperation with AI-generated outputs in general, including calls to "stop this". This highlights an important meta-issue regarding images created by T2I systems, *i.e.* a subjectivity and heterogeneity of opinions towards T2I systems themselves.
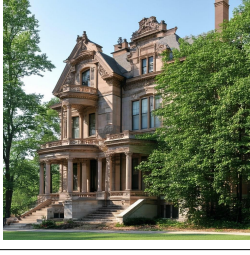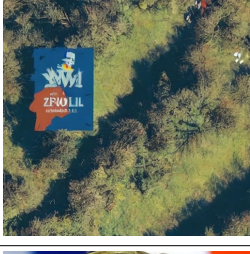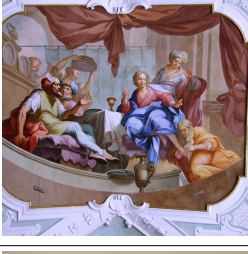
E#	AI Image	Real Image	Feedback	ITA Scorers
E1			**_Bangles_** 1. Bangles are circular ornaments with some colors and patterns on them. 2. The width and the carvings and the colors make it seem more aligned to my culture	$\phi^*_{\text{CuRe}} = 5.0/5$ Khanuja et al. [31] = 0.060 Ventura et al. [65] = 0.045 $\phi_{ITA}(c, r) = 0.124$
E2			**_Festival de la Primavera_** 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).	$\phi^*_{\text{CuRe}} = 1.0/5$ Khanuja et al. [31] = 0.086 Ventura et al. [65] = 0.067 $\phi_{ITA}(c, r) = 0.156$
E3			**_Jalangkote_** 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.	$\phi^*_{\text{CuRe}} = 1.7/5$ Khanuja et al. [31] = 0.112 Ventura et al. [65] = 0.093 $\phi_{ITA}(c, r) = 0.022$
E4			**_Puchner Mansion_** 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.	$\phi^*_{\text{CuRe}} = 1.3/5$ Khanuja et al. [31] = 0.040 Ventura et al. [65] = 0.053 $\phi_{ITA}(c, r) = 0.096$
E5			**_Zwölf Glaubensartikel_** 1. Overall the vegitation seems possible to find in my country 2. This image could be a shot from above of some forest in my country.	$\phi^*_{\text{CuRe}} = 4.7/5$ Khanuja et al. [31] = 0.108 Ventura et al. [65] = 0.091 $\phi_{ITA}(c, r) = 0.079$
E6			**_Michelle Bachelet_** 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 completly wrong, but together they make a face who cant be taken to be Bachelet, also the hair looks really fake.	$\phi^*_{\text{CuRe}} = 2.0/5$ Khanuja et al. [31] = 0.066 Ventura et al. [65] = 0.054 $\phi_{ITA}(c, r) = 0.164$

Figure 18. Visualization of quantitative Image-Text Alignment (ITA) scorers, user study CuRe scores and text feedback. The first two artifacts were generated with FLUX.1 [dev], the second two generated with SD 3.5 Large, and the last two generated with SD 1.5.
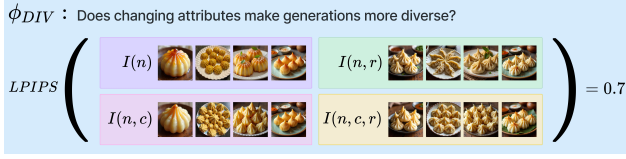
I. Diversity

Figure 19. Computation of diversity score ϕ_{DIV} on $n =$ "modak", a type of dumpling from India. ϕ_{DIV} computes an aggregate pairwise dissimilarity (LPIPS [76]) between multiple seeds of images generated with varying attribute specification, *i.e.* $I(n), I(n,c), I(n,r), I(n,c,r)$. In the example above, this encompasses $16c2 = 120$ pairs.

We compute two notions of diversity: a) **intra-artifact**, over T2I generations with multiple random seeds for a single cultural artifact n; b) **intra-category**, over T2I generations of multiple random seeds for a single cultural category c. We discuss motivating examples for these two separate notions in Sec. 3.4). Details for seeding for different T2I systems is provided in Appendix B.1.

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

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

Vendi Scores. Vendi Scores (VS) [21, 30] define a similarity measure via a kernel over selected attributes (*e.g.* $r = \{Country, Continent\}$). 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 cultural artifacts n belonging to category c to compute **intra-category** diversity, *i.e.*

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

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

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. 3.2). Recall

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

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

We expect that in the long tail, a larger marginal increase in information $n \to a$ will cause a large increase in diversity, as there will be a larger visual difference between images generated with different attributes $I(n) \to I(a)$ (as seen with the "Banku" artifact in Fig. 4). Intuitively, this evaluates how much diversity changes when we mix images of "Banku" with images of "Banku, a type of dumpling", images of "Banku, from Ghana", and images of "Banku, a type of dumpling from Ghana". If it is relatively unchanged, we predict it to lie in the head of the supercategory artifact distribution.

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

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

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

We observe two clusters of T2I sytem behavior with our scorer: in the first cluster, FLUX.1 [dev] and SD 3.5 large (the highest quality T2I systems per Arena Bench ELO, Tab. 12) show higher change in diversity with increasing information in the Global South (proxy for long tail) than the Global North (proxy for head) across supercategories, which aligns with our hypothesis given our assumed proxies. In the second cluster, Ideogram 2.0, Stable Diffusion XL and DALL-E 3, this behavior is reversed, except for architecture and people for SDXL and DALL-E 3 (which has a high 66.5% refusal rate for generating images of people due to inbuilt safety filters, see Appendix B.2). Stable Diffusion 1.5 is the weakest T2I system, and follows similar trends to the second cluster, except for Art. Architecture consistently has higher divergence in the Global North than the South, across T2I systems, which we suspect is due to the inherent lack of diversity in generations of specific named entities, *i.e.* low intra-artifact variance (a similar problem to art and people).

In summary, our proposed scorer ϕ_{DIV} can thus predict where in its learned distribution an image generated by a T2I system lies, *i.e.* in the head or long tail. We note our pre-assigned clustering of global north and south is just *one example proxy* for head and long tail. We note while our scorer
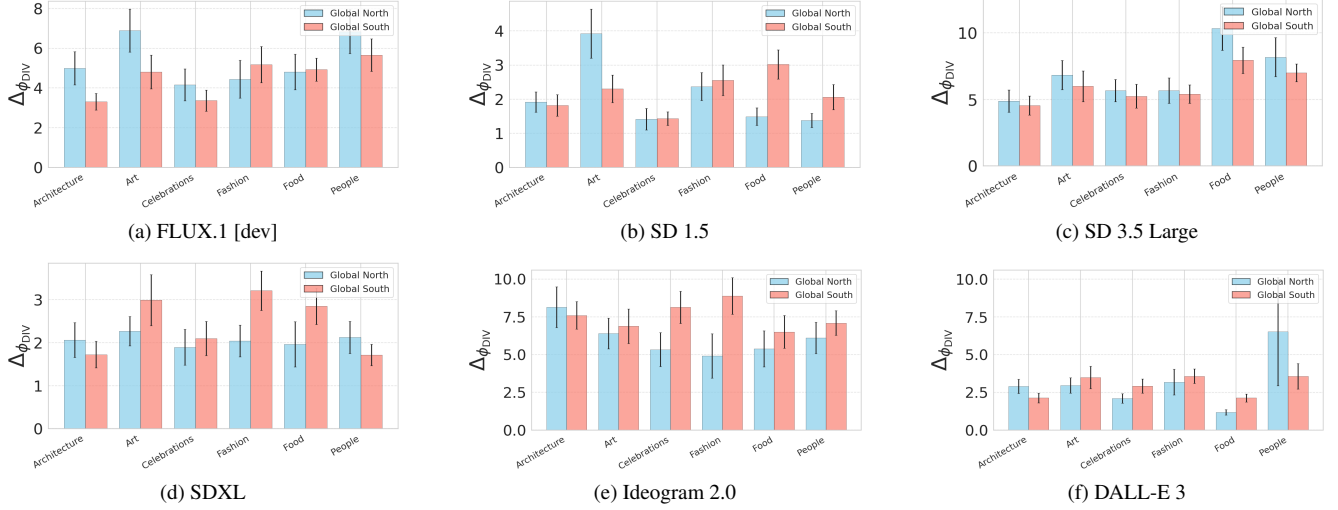
Figure 20. A bar plot of $\Delta\phi_{DIV}$ across all artifacts in the CuRe dataset aggregated at the supercategory level (mean + std error bar) for each T2I system in our evaluation set. We separate each bar at the supercategory level into artifacts belonging to the Global North (GN) and Global South (GS) to examine if there is a noticeable discrepancy between scores in these clusters.

requires generating multiple seeds of images with different attribute specification levels (practically, four seeds across four styles = 16 total images), it is still relatively cheaper to compute when compared to computing artifact frequency over a large pretraining dataset via string matching, such as in Parashar et al. [44], which we discuss in more detail in Appendix F.

I.2. Qualitative Analysis of DIV scorers

We highlight several qualitative examples on the CuRe dataset of our DIV scorer ϕ_{DIV} compared to baselines in Fig. 21. In alignment with our quantitative observations in Appendix I.1, in the examples **E2** ("*Hokkien Mee*"), **E3** ("*Third Mainland Bridge*"), **E4** ("*Rabat Lighthouse*") and **E5** ("*Penelope Cruz*"), all diversity scorers show an inverse relationship to human judgments of cultural representativeness ϕ^*_{CuRe}. These examples have high human scores for cultural representativeness phi^*_{CuRe} and show very little change in diversity score from LPIPS(n) to $\phi_D IV$, indicating that attribute specification does not increase diversity significantly and the artifact likely lies in the head of the T2I system distribution, which aligns with our hypothesis in Sec. 3.4. We highlight failure cases where the baseline $LPIPS(n)$ outperforms our scorer ϕ_{DIV} in **E1** ("*Ushabti*") and **E6** ("*Vaso de los Guerrero*"), where for a low ϕ^*_{CuRe}, the baseline shows a higher diversity score and accurately matches the inverse relationship. We suspect that the factuality-diversity tax especially hurts handicrafts like pottery, where coarse semantic features may be accurate (shape, structural outline) but finer details are wrong (*e.g.* facial features and material in **E1** and material and textures in **E6**.)
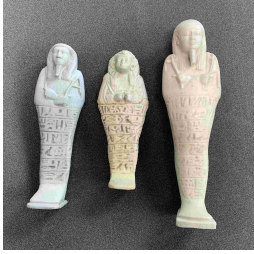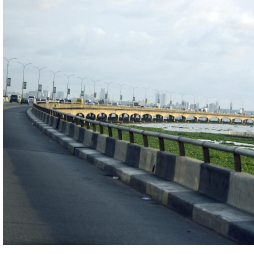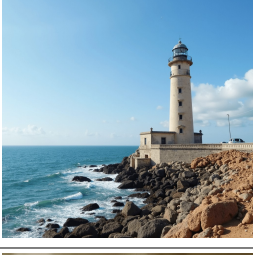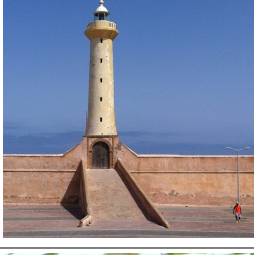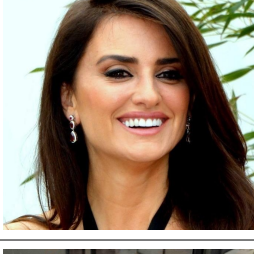
E#	AI Image	Real Image	Feedback	Score
E1			**Ushabti** 1. This image has used a different facial feature of the statue you would not see in my country of culture. Although, the wear and tear as well as the shape of the statue being mummy-like would be found in my country. 2. The similarities lies in 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.	$\phi^*_{\text{CuRe}} = 1/5$ $LPIPS(n) = 0.62$ $\phi_{DIV} = 0.48$
E2			**Hokkien Mee** 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	$\phi^*_{\text{CuRe}} = 4/5$ $LPIPS(n) = 0.70$ $\phi_{DIV} = 0.59$
E3			**Third Mainland Bridge** 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.	$\phi^*_{\text{CuRe}} = 4.5/5$ $LPIPS(n) = 0.50$ $\phi_{DIV} = 0.50$
E4			**Rabat Lighthouse** 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	$\phi^*_{\text{CuRe}} = 5/5$ $LPIPS(n) = 0.51$ $\phi_{DIV} = 0.54$
E5			**Penelope Cruz** 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.	$\phi^*_{\text{CuRe}} = 4/5$ $LPIPS(n) = 0.74$ $\phi_{DIV} = 0.62$
E6			**Vaso de los Guerrero** 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	$\phi^*_{\text{CuRe}} = 2/5$ $LPIPS(n) = 0.75$ $\phi_{DIV} = 0.55$

Figure 21. Diversity scorer and user feedback

J. Multimodal LLM as a Judge

To our knowledge for the first time, we perform a preliminary, exploratory analysis of using a multimodal large language model (MLLM) to approximate human judgments (gold scores) of cultural representativeness (ϕ^*_{CuRe}) and image perceptual similarity (ϕ^*_{PS}, ϕ^*_{GT}) and its ability to provide accurate textual justifications for its scores in Appendix J.1. We choose the recent Gemini 2.0 Flash [15], a popular yet efficient proprietary state-of-the-art MLLM. We discuss qualitative examples of Gemini 2.0 Flash scores and justifications on artifacts from the CuRe dataset and highlight several shortcomings and areas for improvement in Fig. 22b.

Specifically, we query Gemini 2.0 Flash [15] with the same set of questions asked to real humans in our user study (details in Appendix J), and output a score from 1 (low) to 5 (high), *i.e.* the same scale as the Likert scores from our user study (Appendix D). We also query Gemini for a textual justification of its scores to inspect its reasoning. We follow the same evaluation setup as PS, ITA and DIV scorers (Sec. 4.1) 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, suggesting a pretraining dataset has wider cultural coverage of artifact - region associations compared to smaller VLMs like OpenCLIP and SigLIP 2. Lastly, we examine Gemini's textual justifications for its scores to inspect the reliability of its reasoning (see Fig. 22 in Appendix J). We find that Gemini tends to overlook culture-specific details, even when attribute details may be correct (*e.g.* textures, shapes and patterns of an object that are not culturally accurate). This issue is especially prevalent in regions from the Global South (*e.g.* Nigeria, Ghana, Iran), where Gemini hallucinates culture-specific details (*e.g.* "*jollof rice*" from Ghana, "*Femi Kuti*" from Nigeria and "*sami headwear*" from Finland) and mistakenly highlights accuracy to the wrong culture (*e.g.* Chinese instead of Korean for "*Chuseok*" and Egyptian instead of Iranian in "*Takht-e Fulad*").

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

Gemini Prompt for Perceptual Similarity:

You will be shown two images below. The image on the left is an AI-generated image of {item_name} ({category_name}). The image on the right is a grid of four real images of {item_name} from Wikipedia.
Generated Image (Left):
[Image 1 Placeholder - will be inserted by API call]
Real Images Collage (Right):
[Image 2 Placeholder - will be inserted by API call]
Please analyze the images and answer the following:
1. On a scale of 1-5, how similar do you think the generated image on the left is to the real images on the right? Use the following scale:
    1: Not at all similar
    2: Slightly similar
    3: Reasonably Similar
    4: Very Similar
    5: Extremely Similar
2. Provide a brief explanation for your similarity rating, detailing specific visual aspects you compared.
Format your response strictly as JSON with the keys "similarity_rating" (for the number 1-5) and "similarity_explanation" (for the text explanation).

Example Response:
{
    "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."
}

J.1. Gemini 2.0 Flash as a Scorer

As Gemini 2.0 Flash is a multimodal LLM, it can be queried for complex and nuanced visual-question answering (VQA), such as directly querying for cultural representativeness and perceptual similarity scores. We thus replicate our user study setup and questionnaire structure (Appendix D) with Gemini.

Perceptual Similarity. Similar to our user study setup in Fig. 5 (Appendix D.2), we query Gemini for perceptual similarity (**Q1a**) over all artifacts in the CuRe dataset by showing Gemini one image generated by our T2I system, and a grid of four ground-truth images. Both images are passed in as URLs via the Google Gemini API[11]. Mirroring our user

---

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

study, we prompt Gemini to score the similarity of the T2I generated image to the ground-truth image grid from from 1 ("Not at all similar") to 5 ("Extremely similar"). We also examine if Gemini can correctly highlight which details of the T2I image it thought were correct or incorrect and justify its score through text. We show the exact prompt we use below:

CuRe **and Ground-Truth Likelihood.** Similar to our user study setup in Fig. 7 (Appendix D.4), we ask Gemini to rate cultural representativeness (**Q3a**) and likelihood of the T2I image belonging to the class of artifact a, which we call ground-truth likelihood (**Q3b**). Ground-truth likelihood is very similar to perceptual similarity in what it is trying to measure (faithfulness / factuality), but unlike PS, Gemini must score GT likelihood without seeing any ground-truth images of artifact a. Similar to PS, we ask Gemini to justify its scores through text. We show the exact prompt below:

Prompt for CuRe and GT Likelihood:

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

J.2. Analysis of Gemini 2.0 Flash Responses.

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

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

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

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

Regional Trend: A notable pattern is that most failure cases involve artifacts from the Global South (Nigeria, Iran). While Gemini 2.0 Flash performs well on globally familiar or culturally neutral items (the "head"), it struggles with region-specific details that require localized knowledge (*e.g.* mistaking Chinese culture for Korean). This
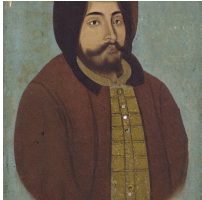
Figure 22. Visualization of Gemini 2.0 Flash responses where the MLLM either failed or succeeded in identifying nuances compared to humans. Examples were generated using FLUX.1 [dev], SD 3.5 Large, and SD 1.5. (Continued on next page)

E#	AI Image	Real Image	Feedback
E1	Artifact: Femi Kuti Category: Musician ϕ^*_{CuRe} : 2.34	Region: Nigeria Super Category: People ϕ_{Gemini} : 5	**Human Judgments (User Study):** 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. **MLLM Judgment (Gemini 2.0 Flash):** 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.
E2	Artifact: Jollof Rice Category: Rice dish ϕ^*_{CuRe} : 4.67	Region: Nigeria Super Category: Food ϕ_{Gemini} : 5	**Human Judgments (User Study):** 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 **MLLM Judgment (Gemini 2.0 Flash):** 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.
E3	Artifact: Chuseok Category: Harvest festival ϕ^*_{CuRe} : 1.67	Region: South Korea Super Category: Celebrations ϕ_{Gemini} : 5	**Human Judgments (User Study):** 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. **MLLM Judgment (Gemini 2.0 Flash):** 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.
E4	Artifact: Sami headwear Category: Hat ϕ^*_{CuRe} : 2	Region: Finland Super Category: Fashion ϕ_{Gemini} : 5	**Human Judgments (User Study):** 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 pictute 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. **MLLM Judgment (Gemini 2.0 Flash):** 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.

(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).

likely reflects the long tail nature of the pretraining data of T2I systems, and highlights the need for more culturally balanced data curation and catering to more heterogeneous preferences to improve alignment with diverse user perspectives.

Figure 22. Continued from previous page.

E#	AI Image	Real Image	Feedback
E5	Artifact: Takht-e Fulad Category:Monument/Memorial ϕ^*_{CuRe} : 3.67	Region: Iran Super Category: Architecture ϕ_{Gemini} : 1	**Human Judgments (User Study):** 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 munument 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. **MLLM Judgment (Gemini 2.0 Flash):** 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.
E6	Artifact: Portrait-Amir Kabir Category: Oil Painting ϕ^*_{CuRe} : 3	Region: Iran Super Category: Art ϕ_{Gemini} : 5	**Human Judgments (User Study):** 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. **MLLM Judgment (Gemini 2.0 Flash):** 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.
E7	Artifact: Kiping Category: Flatbread ϕ^*_{CuRe} : 1.34	Region: Philippines Super Category: Food ϕ_{Gemini} : 1	**Human Judgments (User Study):** 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. **MLLM Judgment (Gemini 2.0 Flash):** 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.
E8	Artifact: Pulseira Category: Jewellery ϕ^*_{CuRe} : 5	Region: Brazil Super Category: Fashion ϕ_{Gemini} : 4	**Human Judgments (User Study):** 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. **MLLM Judgment (Gemini 2.0 Flash):** 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.

(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)