CULTURE IN ACTION: EVALUATING TEXT-TO-IMAGE MODELS THROUGH SOCIAL ACTIVITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image (T2I) diffusion models achieve impressive photorealism by training on large-scale web data, but models inherit cultural biases and fail to depict underrepresented regions faithfully. Existing cultural benchmarks focus mainly on object-centric categories (e.g., food, attire, and architecture), overlooking the social and daily activities that more clearly reflect cultural norms. Few metrics exist for measuring cultural faithfulness. We introduce CULTIVate, a benchmark for evaluating T2I models on cross-cultural activities (e.g., greetings, dining, games, traditional dances, and cultural celebrations). CULTIVate spans 16 countries with 576 prompts and more than 19,000 images, and provides an explainable descriptor-based evaluation framework across multiple cultural dimensions, including background, attire, objects, and interactions. We propose four metrics to measure cultural alignment, hallucination, exaggerated elements, and diversity. Our findings reveal systematic disparities: models perform better for global north countries than for the global south, with distinct failure modes across T2I systems. Human studies confirm that our metrics correlate more strongly with human judgments than existing text–image metrics.

1 INTRODUCTION

The 2007 film Ratatouille earned 41 film awards including Best Feature at the 2008 Oscars (Wikipedia). Part of its appeal lies in the very realistic portrayal of the city of Paris, and of French culture and cuisine (SeattleTimes). To achieve this, creators visited places in Paris to soak in the culture and environment, including its highly distinctive visual aspects. Many other well-regarded films (animated ones like Luca and Coco, and live action ones like Amelie, Crouching Tiger Hidden Dragon, and Reservation Dogs) also devoted significant effort to ensuring they capture the true atmosphere and visuals of the places they portray. Such culturally accurate visual portrayals are important for many types of creative and marketing content beyond film, e.g., advertising.

The advancement of text-to-image generative models in theory offers the potential for automated or semi-automated creation of such content. However, recent Text-to-Image (T2I) models are trained on large-scale web-based data, which is WEIRD (Western, Educated, Industrialized, Rich, and Democratic) (Henrich et al., 2010). The resulting bias is particularly problematic for cultural activities, the social practices through which cultures express their values and meanings. Understanding culture is best achieved through everyday activities and social interactions since these practices embody the values and meanings of a society (Geertz, 2017; Hall, 1973). However, cross-cultural studies of T2I models are heavily understudied, with the most recent benchmarks mainly focusing on a few specific object-centric categories such as architectures, clothing, food, and landmarks (Rege et al., 2025; Chiu et al., 2024; Basu et al., 2025).

We examine how well T2I models portray different cultures, focusing on *activities* whose visual representation varies significantly across cultures. Unlike static cultural artifacts, **activities are contextual and compositional**. The same activity can have multiple valid cultural variants. For example, "eating at home in Iran" may involve sitting at a formal dining table or gathering on the floor around a traditional *sofreh*. This contextual nature makes activity evaluation fundamentally different from object recognition, where cultural artifacts have more limited shapes and attributes.

We address these challenges by introducing **CULTural acTIViTy (CULTIVate)**, a comprehensive benchmark for evaluating T2I models on culturally-grounded social activities. CULTIVate spans 16

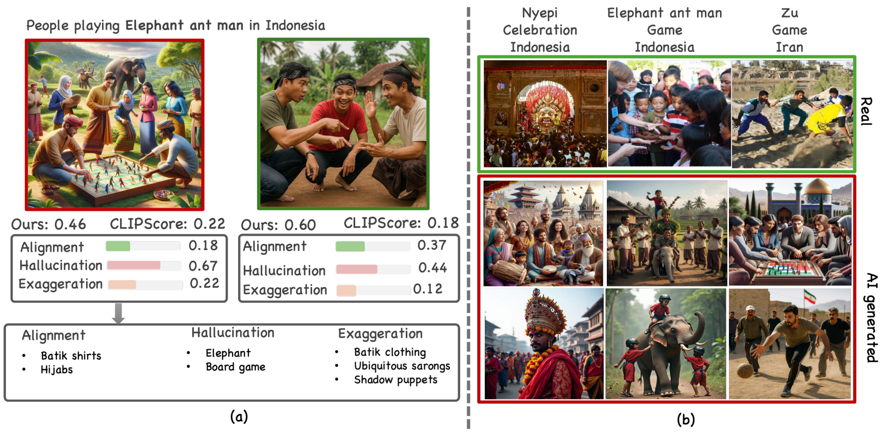


Figure 1: (a) Examples of good (aligned) and bad (hallucinated or exaggerated) aspects of images generated for three cultural activities; these aspects are automatically computed by our framework. (b) Contrasting real and generated images.

countries across 9 activity categories (eating, greetings, celebrations, religious practices, etc.), yielding 576 prompts that capture the contextual complexity missing from object-centric benchmarks. We evaluate 6 state-of-the-art T2I models, generating over 19,000 images and collecting 3,000 real reference images. As Fig 1 (a) illustrates, activity evaluation reveals complex and *new* failure patterns: models may generate wrong activities, include culturally correct but *hallucinated* elements, or produce heavily *exaggerated* scenes. This is in contrast to object/artifact-centric concepts (e.g., Eiffel tower, ceramic diyas), where the *major failure mode* is incorrect (i.e., wrong object) generation. These complex failure patterns raise critical questions: *Are current evaluation metrics effective for cultural assessment of activities? What characteristics must an effective metric have for this task*?

Our work explores *cultural faithfulness*. The most related prior evaluation work has relied predominantly on human surveys due to assessment complexity (Kannen et al., 2024; Bayramli et al.). Other recent works use image-text alignment as a proxy (Rege et al., 2025; Khanuja et al.), but these approaches use VLMs to directly score cultural faithfulness and **rely on VLM internal knowledge** that inherits similar cultural **biases**. For example, when models generate literal elephants for "elephant ant man game" (a rock-paper-scissors-like game in Indonesia), VLM-based metrics (e.g., CLIPScore (Hessel et al., 2021)) may reward this literal interpretation because due to their bag-of-words behavior (Yuksekgonul et al., 2022), lack of compositional understanding, and poor performance on implicit text-image alignment. This fundamental misalignment extends beyond individual cases: our analysis reveals that **image-text alignment metrics correlate positively with cultural *exaggeration***, while effective metrics should correlate **negatively according to human judgment**.

To cope with these challenges and answer the questions above, we introduce AHEaD diagnostic tools (**A**lignment, **H**allucination, **E**xaggeration, and **D**iversity) that **use external visual descriptors rather than biased VLM knowledge**. We decompose activities into *interpretable visual descriptors* that capture cultural elements across multiple dimensions. To ensure robust reference descriptor generation, we employ a *proposer-refiner* approach where *proposer* LLMs generate diverse candidates, and *refiner* filters duplicates and errors. AHEaD provides **interpretable insights**: *Alignment* measures cultural coverage with respect to reference descriptors; *Hallucination* quantifies incorrect or irrelevant elements in the image; *Exaggeration* measures over-representation compared to real images; *Diversity* captures **variation in cultural elements** rather than low-level visual attributes (e.g., color, texture). Unlike naive image-text alignment, our framework provides interpretable feedback at multiple levels (e.g., image sets, individual images, or specific elements within the image). This enables researchers to identify which cultural aspects are missing, over-represented, or faithfully depicted.

In more detail, applying our framework works as follows. First, we compute our AHEaD metrics (Alignment, Hallucination, Exaggeration, and Diversity) automatically, without requiring human annotations, which makes applying them to novel scenarios (e.g., new countries) scalable. The metrics compare different aspects of quality in the images generated by different text-to-image models, and can be used to quantitatively judge which T2I model to deploy when aiming to depict a particular country. Second, our framework outputs the aspects of social activities that are not represented well by T2I models. In particular, it can output the top-k and bottom-k of descriptors likely to be included mistakenly , and top-k and bottom-k of descriptors likely to be over-represented (exaggerated). This ability can be used to improve existing T2I models, e.g., by requesting models to add or remove particular concepts using the descriptors identified as problematic. Third, our framework computes correlations between the AHEaD metrics. These correlations can highlight trade-offs in adjustments to models and outputs. For example, we aim to answer the question, does boosting alignment reduce or boost hallucination and exaggeration?

We conduct comprehensive experiments on CULTIVate and reveal systematic limitations in current cultural evaluation. Our AHEaD metrics achieve 27% higher rank correlation with human judgments of cultural faithfulness compared to using the same MLLM backbone directly as a judge, and significantly outperform existing image-text alignment metrics. Importantly, analysis suggests that metrics are complementary; the best rank correlation with with human faithfulness is achieved when combined (Alignment, Hallucination, and Exaggeration). We also find consistent bias across all tested T2I models: they generate more culturally faithful images for Global North countries than Global South countries, with alignment score gaps of 4-8%.

To summarize, our contributions are: (1) CULTIVate benchmark: First cultural evaluation benchmark focused on social activities, spanning 576 prompts across 16 countries with over 19k images; (2) AHEaD diagnostic tools: Novel metrics using external descriptors that achieve 27% better human correlation strongest baseline; (3) Cultural bias analysis: Systematic demonstration of Global North bias across all tested T2I models; (4) Proposer-refiner framework: Robust descriptor generation enabling scalable cultural evaluation without human annotations.

2 Related Works

Image-Text Alignment Metrics. General-purpose metrics rely on low-level features (e.g. FID (Heusel et al., 2017), LPIPS (Zhang et al., 2018)) or global image-text alignment (e.g. CLIPScore (Hessel et al., 2021), VQAScore (Lin et al., 2024)). Some metrics require expensive human judgments (e.g. ImageReward (Xu et al., 2023), PickScore (Kirstain et al., 2023)). We show these correlate poorly with human judgment.

Cultural Evaluation. Cultural evaluations typically measure *realism*, *diversity*, and *cultural faithfulness* (Liu et al., 2024; Nayak et al., 2025; Jha et al., 2024; Liu et al., 2023). Most automated approaches focus on *diversity* using vision encoders for image-image similarity (Zhang et al., 2018; Khanuja et al.; Jha et al., 2024) or Vendi-score (Kannen et al., 2024; Friedman & Dieng). Vision encoders can exhibit geographical bias and focus on low-level variations (e.g., color/texture) rather than cultural content. Alternatively, (Basu et al., 2025) use VQA models with LLM-generated questions for geographical diversity. Importantly, **diversity metric do not provide insights on cultural faithfulness**, the focus of this work. For cultural *faithfulness*, recent approaches explore VLM image-text alignment; (Khanuja et al.; Basu et al., 2023) measure alignment with simple country prompts, while (Rege et al., 2025) measures alignment between hierarchical prompts. These approaches **rely on VLM internal knowledge** but VLMs inherit Western-centric biases and compositional scenes (Yuksekgonul et al., 2022). Since faithfulness remains challenging for automated methods, many works rely on human surveys which are expensive to obtain (Nayak et al., 2025; Liu et al., 2024; Jeong et al., 2025; Jha et al., 2024). We introduce the suite of automatic AHEaD metrics, which uses *external visual descriptors* to measure cultural alignment while penalizing for hallucination and over-representation.

Cultural Benchmarks. Cultural understanding has been extensively studied for image understanding tasks (Kalluri et al., 2023; Ramaswamy et al., 2023; Nayak et al., 2024; Astruc et al., 2024; Vayani et al., 2025; Liu et al., 2025; Yin et al., 2023). For T2I generation, existing benchmarks are primarily object-centric. (Kannen et al., 2024) covers 8 countries across 3 artifact categories, (Jha et al., 2024) includes 10 countries on food and architecture, and (Basu et al., 2023) covers 27 countries using parsed noun phrases. Our CULTIVate differs by focusing on social activities rather than artifacts. Activity scenes are contextual and compositional with multiple valid variants. This creates evaluation challenges: correctly generating objects is insufficient since cultural accuracy depends on appropriate interactions, context, and spatial arrangements. Models exhibit multiple failure modes,

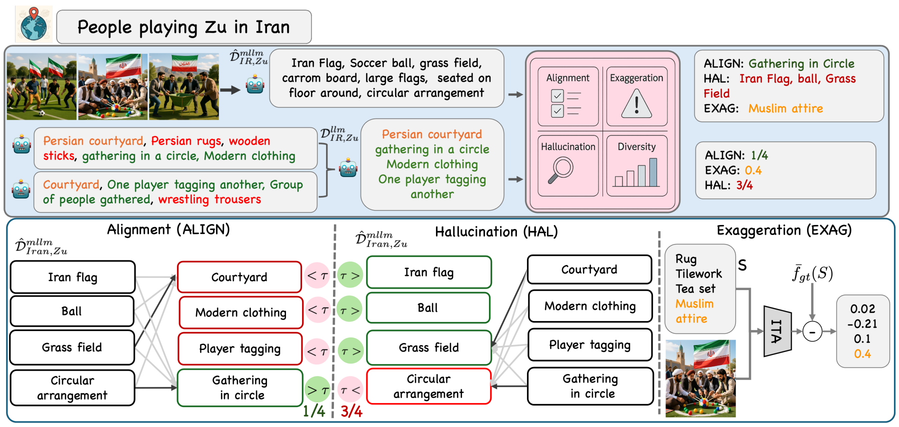


Figure 2: **(Top) Overview.** We extracted image descriptors $\hat{\mathcal{D}}^{\text{mllm}}$ with InternVL3, while reference descriptors $\mathcal{D}^{\text{llm}}$ are obtained via a proposer–refiner pipeline in data annotation stage (i.e. offline) without using images. Proposers generate diverse candidates, and the Refiner removes duplicates and filters incorrect ones. AHEaD measures cultural competence through alignment, hallucination, exaggeration, and diversity, providing not only quantitative scores but also interpretable feedback (i.e., what is aligned, missing, or exaggerated). **(Bottom) Cultural Faithfulness metrics.** Alignment measures whether expected descriptors are present (similarity above threshold τ), hallucination flags elements unsupported by references (e.g., circular arrangement), and exaggeration detects exaggerated cues overemphasized with respect to real-images (e.g., muslim attire)

including incorrect activity, correct activity but exaggerated scenes, or hallucinated cross-cultural elements. Concurrent work (Nayak et al., 2025) focuses on explicit vs implicit cultural expectations via *human evaluation*. CULTIVate complements this by specializing in social activities and proposing the first automated metrics for cultural alignment, hallucination, and exaggeration.

Knowledge probed from large language models. While LLM-based visual descriptors have been explored for fine-grained and cross-geography object recognition (Pratt et al., 2023; Menon & Vondrick, 2023; Saha et al., 2024; Buettner et al., 2024), this is the first work to use descriptors for evaluating cultural competence in T2I models.

3 Methodology

We argue that effective cultural evaluation requires more than simple alignment detection. Good metrics should correlate positively with cultural faithfulness while penalizing exaggeration and hallucination. Fig. 5a shows that VLM-based metrics fail this test– images with excessive cultural stereotypes score highly with VLMs but poorly with humans. We address these limitations through visual descriptors that provide transparent cultural criteria. Section 3.1 describes descriptor generation using cultural sources. Section 3.2 introduces AHEaD metrics that evaluate images against descriptors to measure cultural alignment and hallucination (Fig. 2).

3.1 Reference Descriptor Generation

Initially, during data annotation stage, we use LLMs to generate reference descriptors for each activity-country pair that capture cultural elements across five dimensions: *background* (e.g., Eiffel tower, geometric patterns), *attire* (e.g., traditional vs. modern clothing), *objects*, *actions/interactions* (e.g., greeting with a bow), and *spatial layout* (e.g., dancers in a circle).

We propose a two-stage approach inspired by self-consistency prompting (Wang et al., 2022). **(1) Proposer:** The proposer stage uses multiple LLMs (Gemini 2.5 Flash (Comanici et al., 2025) and GPT-4o (Hurst et al., 2024)) to generate diverse descriptor candidates, increasing coverage different of cultural elements. We specifically instruct each LLM to generate up to 10 elements per dimension, where descriptors can be mutually exclusive. **(2) Refiner:** The refiner stage filters candidates to remove duplicates and incorrect descriptors, improving precision and the proposer-refiner process improves AHEaD metric performance compared to single-model generation (Table 4). We emphasize that this stage is image agnostic and is performed once offline to obtain reference descriptors. In Sec. 3.2, we describe how AHEaD metrics are computed.

3.2 AHEaD Evaluation Metrics

Our goal is to evaluate the cultural faithfulness of T2I systems. A good image should capture the expected elements for an activity in a given culture, avoid introducing elements from other cultures, not exaggerate cultural elements, and display variety across plausible scenarios. We capture these aspects with four descriptor-based metrics.

Consider activity a, region r (in our setting, we experiment with countries as the regions), its corresponding prompt $T_{a,r}$, and LLM-generated reference descriptors $\mathcal{D}^{\text{llm}}_{r,a}$. For each (r, a) we generate N images $\{I_1, \ldots, I_N\}$ and extract predicted descriptors for each image I_j using an Multimodal LLM (i.e., InternVL3 (Zhu et al.)) denoted $\hat{\mathcal{D}}^{\text{mllm}}(I_j),\ j = 1, \ldots, N$.

Alignment. This metric measures cultural alignment: whether generated images reflect culturally expected elements in the GT descriptors. A GT descriptor $d \in \mathcal{D}^{\text{llm}}_{r,a}$ is considered aligned if its similarity with any $\hat{d} \in \hat{\mathcal{D}}^{\text{mllm}}_{r,a} = \bigcup_{j=1}^{N} \hat{\mathcal{D}}^{\text{mllm}}_{r,a}(I_j)$ exceeds a threshold τ.

$$\text{ALIGN}_{r,a} = \frac{1}{|\mathcal{D}^{\text{llm}}_{r,a}|} \left| \left\{ d \in \mathcal{D}^{\text{llm}}_{r,a} : \max_{\hat{d} \in \hat{\mathcal{D}}^{\text{mllm}}_{r,a}} \text{sim}(d, \hat{d}) > \tau \right\} \right| \tag{1}$$

where $\text{sim}(\cdot, \cdot)$ denotes sentence embedding similarity. τ is calibrated according to real images in CULTIVateBench (see appendix).

Hallucination. High alignment exhibits high recall (coverage of expected elements); however, it does not take into account the existence of irrelevant elements in the image. For example, an image of "eating at home in Iran" may align with descriptors like *sofreh* or *table*, yet also include an incorrect item such as *chopsticks*. hallucinated $\hat{d}^{-}$ as a hallucination if $\max_{d \in \mathcal{D}^{\text{llm}}_{r,a}} \text{sim}(\hat{d}, d) \le \tau$.

$$\text{HAL}_{r,a} = \frac{1}{|\hat{\mathcal{D}}^{\text{mllm}}_{r,a}|} |\{ \hat{d} \in \hat{\mathcal{D}}^{\text{mllm}}_{r,a} | \max_{d \in \mathcal{D}^{\text{llm}}_{r,a}} \text{sim}(\hat{d}, d) \le \tau \}. \tag{2}$$

Exaggeration. Exaggeration quantifies whether models over-emphasize stereotypical descriptors. For each region r, an LLM produces a set S_r of exaggeration candidates. Given $d_j \in S_r$, let $f(I, d_j)$ be its image–text alignment score with generated image I, and let

$$\bar{f}_{gt}(d_j) = \frac{1}{n_{gt}} \sum_{i=1}^{n_{gt}} f(I^{i}_{gt}, d_j) \tag{3}$$

be the average score over real images. The exaggeration score for I is

$$\text{EXAG}(I) = \max_{s_j \in S_r} \left[f(I, s_j) - \bar{f}_{gt}(s_j) \right]. \tag{4}$$

Faithfulness. We define cultural faithfulness as a composite score that aggregates alignment, hallucination, and exaggeration. Intuitively, a faithful image should (i) cover expected cultural elements, (ii) avoid introducing incorrect ones, and (iii) not overemphasize cultural descriptors.

$$\text{FAITH}_{r,a} = g\big(\text{ALIGN}_{r,a}, \text{HAL}_{r,a}, \text{EXAG}_{r,a}\big), \tag{5}$$

where $g(\cdot)$ is an aggregation function. We simply use their arithmetic mean after adapting HAL and EXAG so that higher values indicate better performance (i.e,. 1-HAL and 1-EXAG)

Descriptor Diversity. Diversity measures how evenly descriptors are distributed across multiple generations. For n images of (r, a), let $q(d)$ be the relative frequency of a descriptor $d \in \mathcal{D}^{llm}_{r,a}$ being covered, with $\sum_d q(d) = 1$. Diversity is defined as normalized entropy:

$$\text{DDIV}_{r,a} = \frac{-1}{\log |\mathcal{D}^{llm}_{r,a}|} \sum_{d \in \mathcal{D}^{llm}_{r,a}} q(d) \log q(d). \tag{6}$$

Semantic Diversity. We define semantic diversity as the additional descriptor coverage obtained when generating multiple images instead of a single image. Specifically, let $ALIGN_k(r,a)$ denote the alignment score computed over k generated images for region–activity pair (r,a). Semantic diversity is defined as

$$\text{SDIV}_{r,a} = ALIGN_n(r,a) - \mathbb{E}[ALIGN_1(r,a)], \tag{7}$$

where $ALIGN_n(r,a)$ is the alignment across n generations and $\mathbb{E}[ALIGN_1(r,a)]$ is the expected alignment when considering only one image. Higher values indicate that additional generations introduce new descriptors, reflecting greater semantic variety.

4 Experimental Setup

4.1 CULTIVate Benchmark

We introduce CULTIVate, a benchmark for evaluating cultural competence of T2I systems through social and daily activities. CULTIVate contains **576 prompts** across **16 countries** and **9 activity supercategories**, generating **19,000+ images** from **6 T2I models** with comprehensive ground truth annotations. Constructing high-quality cross-cultural benchmarks with local/specific activities is nontrivial and requires expert knowledge across regions. In this work, we use a **scalable systematic** approach to create the benchmark by utilizing existing knowledge bases. We leverage two knowledge sources: (1) *CulturalAtlas*[1], which provides cultural practices across regions (countries), including greetings, religious customs, etiquette and communication. (2) *Wikipedia*, which offers fine-grained lists of activities (e.g., games and celebrations). Using GPT-4o, we parse both sources and extract non-overlapping sub-activities.

Activities and Coverage. We consider nine broad activity categories: *games, dances, greetings, celebrations, concerts, eating, religious, wedding, and funeral*. Activities are split into three groups: (1) *multi-variant categories* (dances, games, religious, greetings, celebrations) where we enumerate sub-activities (e.g., different types of dances relevant to a country), (2) *setting-based categories* (eating at home/restaurant, concerts indoor/outdoor), and (3) *single-activity categories* (wedding, funeral). We analyze cultural disparities by dividing countries into Global North (GN): USA, Spain, Italy, Germany, France, and (2) Global South (GS): Iran, Turkey, China, India, Indonesia, Philippines, Nepal, Nigeria, South Africa, Brazil, Mexico, following UN classification [2].

Image Generation. For each prompt, we generate 10 images (1 image for proprietary models due to the cost) using the template: *"A photorealistic photo of* {*sub-activity*} *in* {*country*}*."* We include six recent T2I models: three public (Stable Diffusion 3.5 (Esser et al., 2024), FLUX (BlackForestLabs, 2024), Qwen-Image (Wu et al., 2025)[3]) and three proprietary (DALL·E 3 (Betker et al., 2023), GPT-Image-1 (OpenAI, 2025), Nano-Banana (Google, 2025)). We set the random seeds $42 + i$ (for k-th image) in public models, generating more than 19,000 images.

Reference data. We adopt two complementary strategies for identifying what images of activities in a region (country) should portray: (1) Visual Descriptors. Inspired by prior use of LLMs for object descriptors (Menon & Vondrick, 2023), we extend the idea to activities. For each prompt, we generate up to 10 descriptors per cultural dimension—background, objects, attire, actions/interactions, and spatial relations. This produces diverse descriptors spanning both traditional and modern activity variants (details in Sec. 3.1); (2) Real Images. We collect 20 candidate images per prompt via Google search (10 using the English prompt, 10 using its translation into the language of the respective country), totaling $\sim$12k images. We then apply CLIPScore (Hessel et al., 2021) filtering and retain the top five (total of $\sim$3k) as representative real references which we use in our exaggeration metric. We also use real images for calibration and finding hyperparameters such as τ in Eq. 1.

[1] https://culturalatlas.sbs.com.au/

[2] `https://unctadstat.unctad.org/EN/Classifications/DimCountries_All_Hierarchy.pdf`

[3] We used distilled model: https://github.com/ModelTC/Qwen-Image-Lightning

Model	Region	N=1				N=10	
		ALIGN ↑	HAL ↓	EXAG ↓	FAITH ↑	DDIV ↑	SDIV ↑
SD-3.5-medium	GN	**0.31** ±0.01	**0.55** ±0.02	**0.05** ±0.02	**0.57** ±-	**0.68** ±0.03	**0.33** ±-
	GS	0.26 ±0.03	0.61 ±0.03	0.08 ±0.04	0.52 ±-	0.62 ±0.04	0.32 ±-
FLUX.1-dev	GN	**0.30** ±0.02	**0.56** ±0.03	**0.04** ±0.01	**0.57** ±-	**0.66** ±0.03	**0.32** ±-
	GS	0.25 ±0.03	0.63 ±0.04	0.06 ±0.02	0.52 ±-	0.60 ±0.04	0.30 ±-
Qwen-Image	GN	**0.36** ±0.02	**0.51** ±0.02	**0.06** ±0.01	**0.60** ±-	**0.68** ±0.02	0.28 ±-
	GS	0.30 ±0.03	0.56 ±0.04	0.10 ±0.03	0.55 ±-	0.63 ±0.04	**0.29** ±-
DALL·E 3†	GN	**0.36** ±0.01	**0.50** ±0.01	**0.10** ±0.03	**0.59** ±-	-	-
	GS	0.32 ±0.03	0.54 ±0.032	0.12 ±0.04	0.55 ±-	-	-
GPT-Image-1†	GN	**0.36** ±0.01	**0.49** ±0.01	**0.06** ±0.01	**0.61** ±-	-	-
	GS	0.30 ±0.03	0.55 ±0.03	0.07 ±0.02	0.56 ±-	-	-
Nano Banana†	GN	**0.40** ±0.01	**0.46** ±0.01	**0.10** ±0.03	**0.61** ±-	-	-
	GS	0.35 ±0.03	0.50 ±0.03	0.12 ±0.3	0.57 ±-	-	-

Table 1: **T2I models consistently generate more faithful images on GN countries.** N is number of images per prompt. Best values per model (GS/GN) is **bolded**. † N=1 only due to cost. EXAG values are small because the metric measures relative alignment of synthetic image to real images

4.2 Human Evaluation Setup

We conduct a controlled human study to assess cultural understanding across alignment, hallucination, exaggeration, and realism using Prolific[4] as our platform. Our evaluation covers 11 representative countries spanning all sociol regions, and both GN and GS, includes 1/2 prompt per each activity and 3 T2I models (Stable Diffusion 3.5, FLUX1, and Qwen-Image), totaling 381 forms and 2 annotators per each forms (762 total annotations). We also conduct human evaluation on real images for 3 countries (27 additional forms), bringing the total to 398 forms.

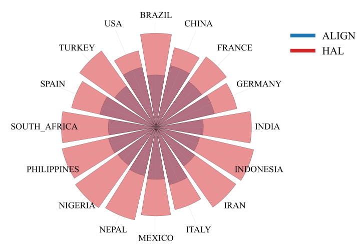
Annotations Collected. We collect the following data from our annotators, which we use as ground-truth (GT) labels: ***GT-FAITH*** = *How well does this image show* {*activity*} *in your country?* is the **main gold standard** that measures the overall faithfulness of an image according to the culture and activity. We further compare against ***GT-EXAG*** = *How exaggerated is the image?* and ***GT-HAL*** = *How incorrect is the image? (incorrect activity or incorrect element according to the activity and/or country)* Responses are collected on a 5-point Likert scale.

We also evaluate the quality of reference descriptors $\mathcal{D}^{\text{llm}}_{r,a}$ through human evaluation. We measure **precision** via binary (correct/incorrect) annotations for each descriptor, achieving 90% accuracy (Table 10). For **recall**, explicit evaluation is infeasible as no ground-truth descriptor set exists. We estimate recall through two complementary measures: (1) annotators rated overall descriptor coverage and quality on a 1-5 Likert scale, achieving an average of 4.5/5, and (2) only 26 out of 378 annotators indicated any missing descriptors. Together, these results demonstrate high precision and comprehensive coverage of our reference descriptors.

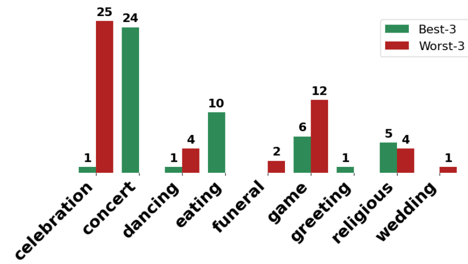
Quality Control. We ensure domain expertise by constraining recruitment via nationality and residence requirements (verified by Prolific). Compensation is set to $8/hour. To ensure reliability, we implement multiple quality control measures. We include attention checks, such as selecting a pre-mentioned number. We use repeated questions to test consistency across responses. We require free-text rationales where annotators must describe what is incorrect in the image. We conduct direct discussions with annotators when facing inconsistent scoring and explanations.

Correlation Metric & inter-rater Agreement. use Spearman's rank correlation to measure how well our proposed metrics align with human judgments. Spearman's ρ evaluates the strength of monotonic relationships between ranked variables, where values near 1/-1 indicates a strong positive/negative correlation. Following Kannen et al. (2024) we use Krippendorff's Alpha (Krippendorff, 2018), appropriate for ordinal Likert scores to measure inter-rater agreement in A.4 Per-country. We observe comparable agreement compared to related works (Nayak et al., 2025; Kannen et al., 2024), exceeding their maximum per-country agreement.

[4] https://www.prolific.com/



(a) Models consistently score better (higher on ALIGNment, lower on HALlucation) for GN. Interestingly, they perform strongly on China



(b) Frequency of activity supercategories appearing among the best-3 (green) and worst-3 (red) subactivities across 16 countries.

Figure 3: Analysis of performance by country (left) and activity (right).

Backbone	Method	GT-FAITH		
		GS (n=231)	GN (n=150)	Overall (n=381)
PickScore Kirstain et al. (2023)	I–T Alignment	**0.20**	-0.02	**0.15**
ImageReward Xu et al. (2023)		-0.03	-0.13	-0.08
CLIPScore Hessel et al. (2021)		0.08	-0.01	0.04
VQAScore Lin et al. (2024)		0.15	**0.16**	0.14
CuRe Rege et al. (2025)		0.13	0.08	0.10
Qwen2.5–VL	MLLM	0.13	0.08	0.10
	FAITH (Ours)	**0.42** (+0.29)	**0.38** (+0.30)	**0.42** (+0.32)
InternVL3	MLLM	0.19	0.18	0.20
	FAITH (Ours)	**0.46** (+0.27)	**0.47** (+0.29)	**0.47** (+0.27)
–	MLLM (GPT-4o)	0.49	0.46	0.48
	VIEScore(GPT-4o) (Ku et al., 2024)	0.37	0.27	0.35
	Human	0.59	0.57	0.58

Table 2: **Comparison with baselines on Cultural Faithfullness on expanded human evaluation (11 countries).** ITA metrics do not capture cultural nuances effectively. Our Faithfulness metric achieves substantially higher Spearman correlation with human cultural-faithfulness judgement. The best metric for each section is **bolded**. Values in parenthesis show improvement with respect to MLLM baseline. We include human–human correlation for reference. InternVL3 is 'InternVL3-14B' and QwenVL2.5 is 'Qwen2.5-VL-7B-Instruct'.

5 RESULTS

We benchmark 6 pre-trained text-to-image (T2I) models using CULTIVate, and evaluate the proposed metrics against prior works using three main human ground truth labels (see Sec. 4.2).

5.1 HOW DO DIFFERENT T2I MODELS PERFORM FOR DIFFERENT COUNTRIES?

T2I models consistently generate more faithful content for Global North (GN) countries. Table 1 shows a consistent bias against GS, with all models performing consistently better on GN than on GS, e.g., Qwen-Image achieves $0.36/0.51$ on GN (ALIGN) vs. $0.30/0.57$ on GS (HAL). Lower ALIGN, along with higher HAL and EXAG and lower DDIV/SDIV scores on GS, suggest that models not only make more mistakes (e.g., depicting the wrong activity or showing a scene from the wrong country) but also generate exaggerated contents with less cultural concepts included. The trend is illustrated in Fig. 3a.

T2I systems perform worst on more culturally-grounded activities. Fig. 3b show often each activity category appeared among the best-3 and worst-3 performing sub-activities across the 16 countries.Models perform best on the least culturally-grounded categories (e.g., concerts, eating) while they make more mistakes on more culturally-grounded activities (e.g., celebrations).

5.2 WHAT METRICS ARE EFFECTIVE FOR CULTURAL FAITHFULLNESS?

Image-Text Alignment methods are ineffective for cultural understanding. Table 2 shows all ITA methods achieve near-zero or negative correlation with human scores (e.g., ImageReward: -

Backbone	Method	GT-FAITH		
		GS (n=231)	GN (n=150)	Overall (n=381)
Qwen2.5-VL	MLLM baseline	0.13	0.08	0.10
	ALIGN	0.41	0.32	0.39
	ALIGN + HAL	0.37	0.37	0.39
	FAITH (ALIGN+HAL+EXAG)	**0.42**	**0.38**	**0.42**
InternVL3	MLLM baseline	0.19	0.18	0.20
	ALIGN	0.40	0.40	0.41
	ALIGN + HAL	0.42	0.46	0.44
	FAITH (ALIGN+HAL+EXAG)	**0.46**	**0.47**	**0.47**
Human	–	0.59	0.57	0.58

Table 3: **Cultural Faithfullness ablation.** Cultural Faithfullness captures best through combination of ALIGNment, EXAGgeration, and HALlucination

Ref. Desc. Generator	LLM (Proposer/ Refiner)	Spearman	Kendall
Proposer	GPT-4o / –	0.28	0.20
Proposer	Gemini 2.5-Flash / –	0.30	0.22
Proposer-Refiner	GPT-4o + Gemini 2.5-Flash / GPT-4o	**0.33**	**0.24**

Table 4: **Proposer–Refiner improves descriptor quality.**

τ	Spearman	Kendall
0.29	0.21	0.15
0.39	0.27	0.20
0.52	**0.33**	**0.24**

Table 5: **Threshold (τ) ablation.**

0.03/-0.13/-0.08 for GS/GN/overall). Results are improved when using MLLMs as a judge, asking the same question as we ask our human annotators to obtain the GT-FAITH. However, MLLM still significantly under-perform our FAITH metric (e.g., Qwen2.5-VL: 0.10 vs FAITH: 0.42; InternVL3: 0.20 vs FAITH: 0.47).

Capturing exaggeration and hallucination are complementary to alignment. Table 3 shows the best correlation with human faithfulness is achieved when all three metrics are combined. This confirms our key finding: effective cultural faithfulness metrics must penalize exaggeration and hallucination, not just measure alignment.

Note that inter-rater agreement (A.4) is moderate and varies by country, consistent with related work (Nayak et al., 2025; Kannen et al., 2024), reflecting the subjectivity of cultural evaluation. We do not evaluate diversity metrics, as diversity is not part of faithfulness and requires collection-level assessment while our annotations are image-level.

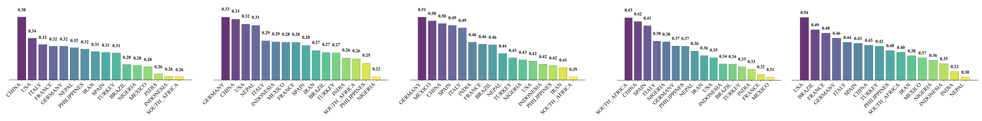
Ablations. In Table 10 (appendix), we show descriptor accuracy using the relevant/irrelevant selections by annotators for a subset of countries; the average is 90.27%. We seek to boost results with the refiner (see Sec. 3.1). In Tab. 4, we see improved results on alignment with human scores, when using the two-stage proposer-refiner, over proposer alone. Further, in Table 5, we show the effect of the threshold parameter τ we use in Sec. 3.2 to compute descriptor matches. We use values corresponding to 25-th, 50-th, and 75-th percentile, and find that the latter performs best.

5.3 What aspects of the activities are depicted best/worst by T2I models?

Fig. 4 shows performance by region (country), for each of five dimensions (groupings) of descriptors. The best-performing country by dimension varies, but USA, China and Germany are consistently among the best. South Africa, Nigeria and India are consistently in the bottom half, except for Interaction (South Africa and Nigeria, both African countries) and Spatial (India). In appendix, we show further results using top descriptors for individual countries, activities, and metrics.

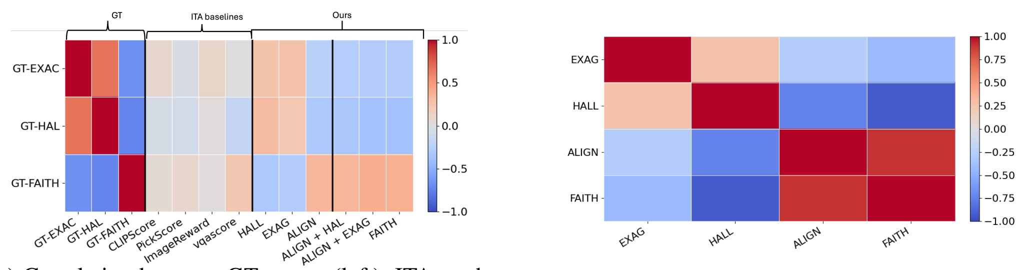
5.4 How do the metrics relate to each other?

To improve the performance of T2I models, a user might want to know how improving upon one metric will affect others. We aim to answer this question by computing correlations between the metrics, shown in Fig. 5b. We see that alignment is negatively correlated with both exaggeration and hallucination. The same trend is observed using human scores; see Fig. 5a which also demonstrates visually the much stronger alignment of our metrics with human scores.



(a) Objects (b) Background (c) Spatial (d) Interaction (e) Attire

Figure 4: Country alignment ranked using each of the five descriptor dimensions. (Zoom to 250%.)



(a) Correlation between GT scores (left), ITA methods (middle), and our scores (right).

(b) Correlations among our proposed metrics.

Figure 5: An effective cultural faithfulness metric should negatively correlate with exaggeration and hallucination.

6 Conclusion

We developed a framework for evaluating generation of images of social activities in different countries. We propose a suite of metrics that can be computed without human involvement, yet show much higher agreement with human assessment than prior metrics. Using our framework, we conduct analysis on sixteen countries and six text-to-image models. We show performance on Global North countries exceeds that of Global South. and demonstrate specific failure modes using our descriptor dimensions. We hope our work equips future researchers with the tools to scalably improve and test performance on this task which has broad applicability, e.g., in the entertainment industry.

References

Guillaume Astruc, Nicolas Dufour, Ioannis Siglidis, Constantin Aronssohn, Nacim Bouia, Stephanie Fu, Romain Loiseau, Van Nguyen Nguyen, Charles Raude, Elliot Vincent, Lintao XU, Hongyu Zhou, and Loic Landrieu. OpenStreetView-5m: The many roads to global visual geolocation, 2024.

Abhipsa Basu, R. Venkatesh Babu, and Danish Pruthi. Inspecting the geographical representativeness of images from text-to-image models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5113–5124. IEEE, 2023. ISBN 9798350307184. doi: 10.1109/ICCV51070.2023.00474.

Abhipsa Basu, Mohana Singh, and Venkatesh Babu Radhakrishnan. Geodiv: Measuring concept diversity of images across geographical regions. In *CVPR 2025 Workshop Vision Language Models For All*, 2025.

Zahra Bayramli, Ayhan Suleymanzade, Na Min An, Huzama Ahmad, Eunsu Kim, Junyeong Park, James Thorne, and Alice Oh. Diffusion models through a global lens: Are they culturally inclusive? URL http://arxiv.org/abs/2502.08914.

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

BlackForestLabs. Flux: A powerful tool for text generation. https://blackforestlabs.ai/, 2024.

Kyle Buettner, Sina Malakouti, Xiang Lorraine Li, and Adriana Kovashka. Incorporating geo-diverse knowledge into prompting for increased geographical robustness in object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13515–13524, June 2024.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, et al. Culturalbench: A robust, diverse, and challenging cultural benchmark by human-ai culturalteaming. *arXiv preprint arXiv:2410.02677*, 2024.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. Culturalbench: a robust, diverse and challenging benchmark on measuring (the lack of) cultural knowledge of LLMs, 2025. URL https://openreview.net/forum?id=n1X2n7MJ8L.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. URL http://arxiv.org/abs/2210.02410.

Clifford Geertz. *The interpretation of cultures*. Basic books, 2017.

Google DeepMind / Google. Nano banana (gemini 2.5 flash image), 2025. URL https://ai.google.dev/gemini-api/docs/image-generation. Image editing / generation model.

Edward T Hall. *The silent language*. Anchor, 1973.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83, 2010.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Suchae Jeong, Inseong Choi, Youngsik Yun, and Jihie Kim. Culture-TRIP: Culturally-aware text-to-image generation with iterative prompt refinment, 2025. URL http://arxiv.org/abs/2502.16902.

Akshita Jha, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Rida Qadri, Chandan Reddy, and Sunipa Dev. ViSAGe: A global-scale analysis of visual stereotypes in text-to-image generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12333–12347. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.667.

Tarun Kalluri, Wangdong Xu, and Manmohan Chandraker. Geonet: Benchmarking unsupervised adaptation across geographies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15368–15379, 2023.

Nithish Kannen, Arif Ahmad, Marco Andreetto, Vinodkumar Prabhakaran, Utsav Prabhu, Adji Bousso Dieng, Pushpak Bhattacharyya, and Shachi Dave. Beyond aesthetics: Cultural competence in text-to-image models. 2024.

Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. An image speaks a thousand words, but can everyone listen? on image transcreation for cultural relevance. URL http://arxiv.org/abs/2404.01247.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Klaus Krippendorff. *Content analysis: An introduction to its methodology*. Sage publications, 2018.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12268–12290, 2024.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pp. 366–384. Springer, 2024.

Bingshuai Liu, Longyue Wang, Chenyang Lyu, Yong Zhang, Jinsong Su, Shuming Shi, and Zhaopeng Tu. On the cultural gap in text-to-image generation, 2023. URL http://arxiv.org/abs/2307.02971.

Shudong Liu, Yiqiao Jin, Cheng Li, Derek F Wong, Qingsong Wen, Lichao Sun, Haipeng Chen, Xing Xie, and Jindong Wang. Culturevlm: Characterizing and improving cultural understanding of vision-language models for over 100 countries. *arXiv preprint arXiv:2501.01282*, 2025.

Zhixuan Liu, Peter Schaldenbrand, Beverley-Claire Okogwu, Wenxuan Peng, Youngsik Yun, Andrew Hundt, Jihie Kim, and Jean Oh. Scoft: Self-contrastive fine-tuning for equitable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10822–10832, 2024.

Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=jlAjNL8z5cs.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Aishwarya Agrawal, et al. Benchmarking vision language models for cultural understanding. *arXiv preprint arXiv:2407.10920*, 2024.

Shravan Nayak, Mehar Bhatia, Xiaofeng Zhang, Verena Rieser, Lisa Anne Hendricks, Sjoerd van Steenkiste, Yash Goyal, Karolina Stańczak, and Aishwarya Agrawal. CulturalFrames: Assessing cultural expectation alignment in text-to-image models and evaluation metrics, 2025.

OpenAI. Gpt image 1, 2025. URL https://platform.openai.com/docs/models/gpt-image-1. Model documentation.

Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15691–15701, October 2023.

Vikram V Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. Geode: a geographically diverse evaluation dataset for object recognition. *Advances in Neural Information Processing Systems*, 36:66127–66137, 2023.

Aniket Rege, Zinnia Nie, Mahesh Ramesh, Unmesh Raskar, Zhuoran Yu, Aditya Kusupati, Yong Jae Lee, and Ramya Korlakai Vinayak. CuRe: Cultural gaps in the long tail of text-to-image systems, 2025. URL http://arxiv.org/abs/2506.08071.

Oindrila Saha, Grant Van Horn, and Subhransu Maji. Improved zero-shot classification by adapting vlms with text descriptions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17542–17552, 2024.

SeattleTimes. French find "ratatouille" ever so palatable. https://www.seattletimes.com/nation-world/french-find-ratatouille-ever-so-palatable/.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kukreja, et al. All languages matter: Evaluating lmms on culturally diverse 100 languages. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19565–19575, 2025.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Wikipedia. List of accolades received by ratatouille. https://en.wikipedia.org/wiki/List_of_accolades_received_by_Ratatouille.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Da Yin, Feng Gao, Govind Thattai, Michael Johnston, and Kai-Wei Chang. Givl: Improving geographical inclusivity of vision-language models with pre-training methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10951–10961, 2023.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Jinguo Zhu, W Wang, Z Chen, Z Liu, S Ye, L Gu, H Tian, Y Duan, W Su, J Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. *URL https://arxiv. org/abs/2504.10479*, 9.

A APPENDIX

A.1 USAGE OF AI

In this section we elaborate on LLM usage in this study. LLMs were used throughout this research as writing assistants, for text polishing, and for literature review through LLM agents and available tools. AI coding assistants[5] were used to assist with programming. However, LLMs were not used blindly and served only as assistants to improve accuracy and efficiency. This paper introduces a benchmark on social activities. As described in the main paper, LLMs (GPT-4o) were utilized to parse online knowledge bases (CulturalAtlas and Wikipedia) to identify activities across countries. Furthermore, the descriptor-based metrics rely on LLM-generated descriptors. However, a proposer-refiner approach was incorporated to improve quality, and descriptors were evaluated through human evaluation (see Table 10).

A.2 LIMITATIONS

Cultural Bias in LLMs. AHEaD uses LLM-generated descriptors as reference points for measuring the cultural competence of T2I models. Since LLMs are trained on web text, we acknowledge that they may encode biases toward Western societies. To mitigate this, we adopt a Proposer–Refiner strategy, which improves descriptor quality and increases agreement with human ground-truth scores. Human evaluation showed 90%. Compared to common alternatives, such as human surveys or real images, our approach is scalable and less costly. Real images collected from the web are themselves biased, while surveys are subjective and expensive. Unlike VLM-based image–text alignment methods or raw image references, our **descriptors are explainable and allow direct inspection of model errors**, rather than being opaque scores.

A.3 CALIBRATION OF THRESHOLD

We propose ALIGN and HAL to measure *how well images cover expected activity/cultural cues* and which *visual elements are incorrect*. Since these metrics are ratio-based, we must set a similarity threshold τ to decide whether a descriptor counts as a *hit* (aligned) or *miss* (hallucinated).

We calibrate τ using real reference images rather than synthetic generations to avoid leakage, since synthetic data may reflect biases of the very T2I models under evaluation. Real images, while noisy, contain culturally faithful content without "wrong" or "exaggerated" elements, making them suitable for calibration. Concretely, we compute descriptor–descriptor similarities between LLM-provided ground-truth descriptors and MLLM-extracted descriptors from real images, then consider candidate thresholds at the lower quartile (Q1), median, and upper quartile (Q3). As shown in Fig.6, Q3 offers the best trade-off by reducing false positives while maintaining recall. Table5 further confirms that Q3 yields the most robust alignment scores across regions.

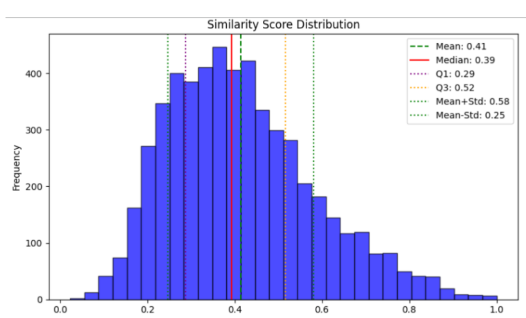


Figure 6: **Threshold τ calibration for ALIGN**

[5]https://cursor.com/

A.4 Implementation Details and Baselines

Evaluation baselines. The goal of this paper is to evaluate the cultural faithfulness competence (GT-ALIGN) of T2I models, where automated evaluation methods remain extremely limited. Existing works rely heavily on human annotations Kannen et al. (2024); Nayak et al. (2025); Basu et al. (2023), while a few recent approaches Khanuja et al.; Rege et al. (2025) approximate cultural faithfulness using image–text similarity. Accordingly, we compare against commonly used and state-of-the-art ITA metrics, including CLIPScore (Hessel et al., 2021), VQAScore (Lin et al., 2024) with "CLIP-FlanT5-xxl" (the strongest publicly available ITA setup), PickScore (Kirstain et al., 2023), and ImageReward (Xu et al., 2023). Following prior ITA practice, we use each model's generation prompt—"A photorealistic image of activity in country"—as the reference for evaluation. We also benchmark against CuRe (Rege et al., 2025), the only metric explicitly designed for cultural faithfulness. For fair comparison, we adopt CuRe's recommended SigLIP2 (Tschannen et al., 2025) configuration and compute mean image–text similarity using the prompts "An image of activity" and "An image from country," omitting their parent-category prompt since this information is already embedded in our activity descriptions (e.g., "people playing tag game").

Across all settings, we find that ITA methods and CuRe exhibit weak correlation with human cultural judgments, whereas our proposed metrics achieve substantially higher and more stable agreement across different MLLM backbones (InternVL3 and QwenVL2.5). We attribute the limitations of existing VLM-based ITA methods to: (1) bag-of-words behavior that misses compositional cultural nuance (Yuksekgonul et al., 2022), (2) reliance on Western-centric training data that introduces cultural biases, and (3) inability to distinguish authentic cultural representation from stereotypical exaggeration. For instance, CLIPScore rewards images containing literal elephants for the "elephant ant man" game—an Indonesian rock–paper–scissors variant—due to keyword matching rather than cultural understanding. To address these issues, AHEAD uses externally generated cultural descriptors instead of VLM embeddings, enabling interpretable evaluation of ALIGN, HAL, and EXAG that aligns more faithfully with human cultural judgment. This is the first work to evaluate cultural HAL and EXAG, and we study both descriptor–descriptor methods (Sec. 3.2) and MLLM-as-Judge baselines using InternVL3 and QwenVL2.5, which answer the same cultural assessment questions posed to human annotators (full prompts in Appendix A.7).

Implementation Details We first use GPT4-o and Gemini 2.5 Flash (best LLMs in cultural understanding (Chiu et al., 2025)) offline once to in the data annotation phase to produce "reference LLM descriptors", these are used as noisy reference to evaluate cultural faithfulness. To minimize the LLM-bias we developed proposer-refiner to combine descriptors of different LLMs which is refined by removing duplicate and incorrect descriptors (results in Table 4). We set the temperature to 0.2 for proposers and 0.1 for the refiner. AHEaD uses an MLLM to extract descriptors, we mainly use InternVL3 ("InternVL3-14B") as MLLM in our pipeline and also test our pipeline with QwenVL2.5("QwenVL2.5-7B"). We set temperature 0 for MLLMs to ensure high precision and reproducibility, and use `all-MiniLM-L6-v2` as the sentence embedding model for similarity computation.

Inter-Rater Agreement We consider "GT-ALIGN" for interrater agreement as the main goal of this work is to measure cultural faithfulness and GT-EXAG/GT-HAL are even more subjective. To assess the reliability of our human annotations, we compute country-level agreement scores for the cultural relevance ratings. Each image is annotated by two independent raters who are originally from the corresponding country. Across the eleven countries in our study, Krippendorff's Alpha Krippendorff (2018) ranges from 0.15 to 0.62. We also compute Cohen's Kappa McHugh (2012) between the two annotator groups and observe a mean value of 0.50. These agreement levels are consistent with previously reported values for cross-cultural image evaluation. CulturalFrames Nayak et al. (2025) reports country-level Alpha values between 0.24 and 0.42, and CUBE Kannen et al. (2024) reports values between 0.09 and 0.58. Our scores are therefore comparable to prior work and also achieve a higher maximum value, which indicates that our annotation protocol yields reliable judgments.

We observe variation across countries, with a standard deviation of 0.13 for Krippendorff's Alpha. Such variation is expected because cultural faithfulness assessments are subjective and depend strongly on cultural and geographic context. Interestingly, the average agreement among Global North countries is 0.28, which is lower than the Global South average of 0.35, even though text-to-image models tend to perform better on Global North regions. We hypothesize that higher-quality outputs may cause annotators to focus more on aspects unrelated to cultural content, such as image quality or visual artifacts, or to rely more heavily on subjective interpretations.

A.5 Additional Results on AHEaD

HAL can effectively detect hallucinations. Table 6 demonstrates HAL's correlation with human scores. shows our proposed HAL metric achieves the best correlation with humans on both GT-HAL and GT-FAITH, demonstrating the effectiveness of our method compared to strong LLMs, specifically InternVL. Although InternVL is used in our pipeline to extract image descriptors, our method outperforms InternVL by 11%. Further, we observe that our HAL metric achieves the most negative correlation with GT-FAITH. This **confirms our hypothesis that hallucination has a strongly negative correlation with faithfulness and can be used to design strong metrics.** In particular, we use InternVL to extract image descriptors.

Method	Backbone	GT-HAL↑			GT-FAITH↓		
		GS	GN	overall	GS	GN	overall
MLLM	InternVL3	0.22	0.24	0.23	-0.20	-0.24	-0.21
	QwenVL2.5	0.29	0.30	0.29	-0.31	-0.36	-0.33
HAL	InternVL3	**0.31**	0.39	0.35	**-0.39**	**-0.44**	**-0.41**
	QwenVL2.5	0.30	**0.42**	**0.36**	-0.33	-0.35	-0.36
Human	–	0.40	0.38	0.39	-	-	-

Table 6: **Correlation with humans on Hallucination.** Our Hallucination metric achieves the highest correlation with human ground truth scores compared to existing MLLM-based approaches, including InternVL which serves as the backbone for MLLM descriptor extraction. Best scores per column are **bolded**.

EXAG can effectively detect hallucinations. We are the first to measure exaggeration. metric achieves the highest correlation with human ground truth scores compared to existing MLLM-based approaches overall, while it achieves **balanced** scores across GN/GS. We further test three different 3 of measuring HAL in Table 11: (1) using LLM GT descriptors, stereotype candidates (ours), and MLLM descriptors. We observe that using LLM/MLLM descriptors is ineffective. This shows that "over-representation" any element (e.g., people or regular objects) is not considered as exaggeration. Over exaggeration is only related to certain culturally specific visual elements.

A.6 Raw human scores

Table 13illustrates the raw human scores. We observe that overall

A.7 Prompts

In this section, we include prompts used in this project.

Method	Backbone	GT-HAL↑				GT-FAITH↓			
		Flux1	Qwen	SD3.5	Avg.	Flux1	Qwen	SD3.5	Avg.
MLLM	InternVL3	0.32	0.07	0.20	0.20	-0.30	-0.10	-0.18	-0.18
	QwenVL2.5	**0.38**	0.11	**0.37**	0.26	-0.39	-0.19	-0.31	-0.30
HAL	InternVL3	0.36	**0.37**	0.24	0.32	**-0.51**	**-0.33**	-0.30	**-0.38**
	QwenVL2.5	0.35	0.31	0.32	**0.33**	-0.38	-0.21	**-0.37**	-0.32
Human	–	0.38	0.34	0.40	0.37	-	-	-	-

Table 7: **Hallucination Per T2I on expanded human evaluation**. Spearman Correlation.

Method	Backbone	GT-EXAG↑			GT-FAITH↓		
		GS	GN	Overall	GS	GN	Overall
	InternVL3	0.24	0.26	0.25	-0.24	-0.21	-0.25
EXAG (MLLM)	QwenVL2.5	0.34	**0.33**	**0.34**	**-0.31**	**-0.25**	**-0.30**
EXAG (ITA)	VQAScore	**0.36**	0.16	0.29	-0.27	-0.14	-0.22
Human	–	0.31	0.39	0.34	-	-	-

Table 8: **Correlation with humans on Exaggeration on expanded human evaluation.** Best scores per-column are bolded. We explore two approaches: EXAG(MLLM) use MLLM for predicting exaggeration, while EXAG(ITA) uses VQAScore and exaggerated candidates from Sec. 3.2. Human evaluation incldues 11 countries with 381 samples (231/150 for GS/GN)

Method	Model	Flux1	Qwen	SD3.5	Avg.
	VQAScore	0.14	-0.02	0.143	0.09
	PickScore	0.06	-0.05	0.03	0.01
	ImageReward	-0.09	-0.24	-0.19	-0.17
Image-Text Alignment	CLIPScore	0.04	-0.28	0.02	-0.05
	CuRe	0.17	0.10	-0.01	0.09
	GPT-4o	0.53	0.27	0.48	0.43
MLLM	InternVL3	0.38	-0.14	0.14	0.13
	QwenVL2.5	0.12	0.02	0.11	0.09
ALIGN (InternVL3)		**0.51**	0.30	0.31	**0.38**
Human	–	0.65	0.40	0.55	0.55

Table 9: Detailed results per T2I model, using Spearman correlation with GT-FAITH human scores

China	France	Iran	Nigeria	USA	India	Brazil	Avg.
89.80	90.54	85.21	91.62	91.44	91.61	91.68	90.27

Table 10: **LLM generated descriptors validation by humans.**

Method	Model	Flux1	Qwen	SD3.5	Avg.
LLM GT Descriptors	EXAG (VQAscore)	-0.276	-0.256	-0.220	-0.251
Stereotype Cand.	EXAG (VQAscore)	0.349	**0.184**	0.230	**0.183**
MLLM-Desc.	EXAG (VQAscore)	-0.221	-0.092	0.063	-0.083
Human	–	0.231	0.097	**0.482**	0.270

Table 11: **Exaggeration per T2I (Spearman)**. Correlation across different text-to-image generators.

Activity	Example Subactivities (across countries)
Eating	Home, Restaurant
Greeting	Namaste (India), Prostrating (Nigeria), Three-kiss (Iran), Cheek kiss (France)
Dancing	Samba (Brazil), Flamenco (Spain), Bharatanatyam (India), Dragon Dance (China)
Game	Kabaddi (India), Ayoayo (Nigeria), Pétanque (France), Baseball (USA), Mahjong (China)
Celebration	Nowruz (Iran), Carnival (Brazil), Bastille Day (France), Thanksgiving (USA), Chinese New Year (China)
Religious	Tazieh (Iran), Candomblé ceremony (Brazil), Catholic mass (Mexico), Temple aarti (India)

Table 12: **A subset of examples of subactivities in CULTIVate.** Highlights distinctive cultural practices across countries.

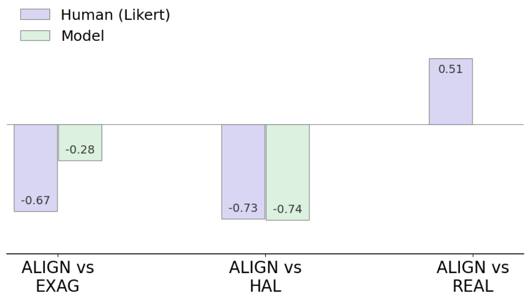


Figure 7: **Detailed results using different correlations with GT-FAITH**. ALIGN and HAL show consistent negative correlation in both humans and models (-0.74), validating the accuracy of our descriptor-based metrics. ALIGN and EXAG are also negatively correlated, though values differ: humans penalize exaggeration as misalignment, whereas ALIGN counts it as aligned if it matches a ground-truth descriptor. This highlights the need for EXAG to capture exaggeration effects not reflected in alignment alone, especially since its computation depends on noisy ground-truth images.

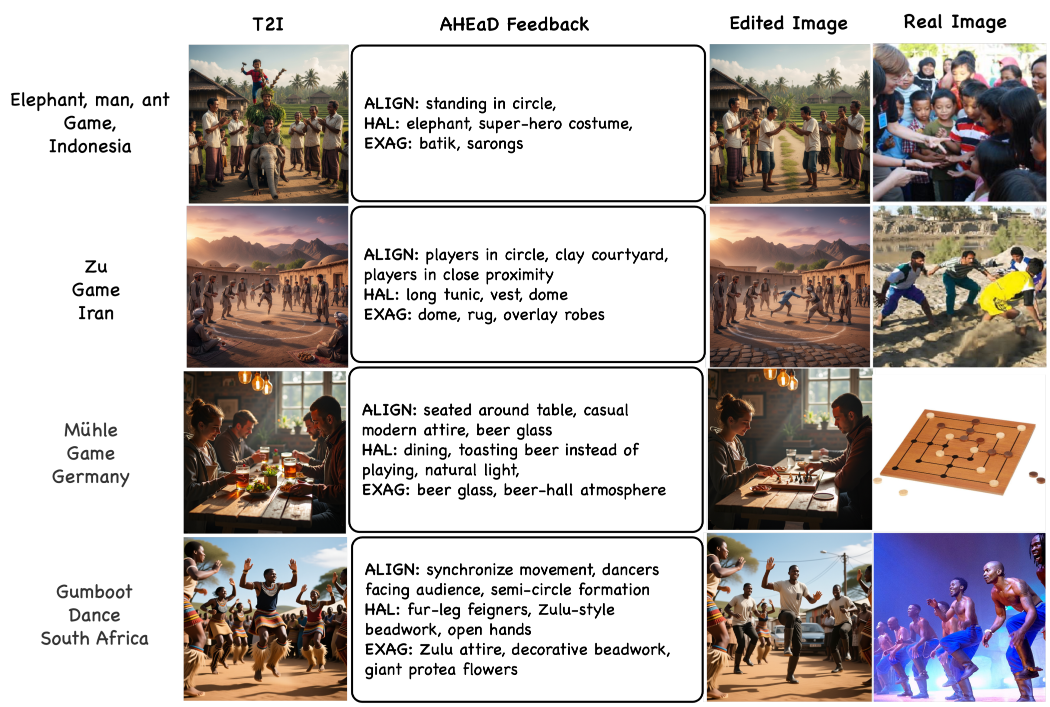


Figure 8: **Illustration of descriptor effectiveness in guiding image editing for improved generation.** (a) **Initial T2I-generated images** (top to bottom: Nano-Banana, Nano-Banana, FLUX, Qwen-Image). (b) **Generated feedback by AHEaD**: We use AHEaD feedback along with reference descriptors $\mathcal{D}^{\text{llm}}$ to create clear instruction prompts (prompt in Table.31) (c) **Edited images**: Nano-Banana is utilized to edit images according to instruction prompts generated in (b). (d) **Real images**.

Region	**GT-FAITH**	**GT-EXAG**	**GT-HAL**	**GT-IMAGE-REALISM**
FRANCE	3.09 (0.73)	2.63 (0.70)	2.29 (0.74)	3.25 (0.39)
BRAZIL	3.61 (0.65)	2.27 (0.47)	1.79 (0.61)	3.46 (0.20)
CHINA	2.95 (0.84)	2.66 (0.52)	2.19 (0.52)	3.17 (0.44)
INDIA	3.27 (0.94)	2.38 (0.73)	1.97 (0.56)	3.56 (0.65)
MEXICO	2.91 (1.00)	2.41 (0.86)	2.38 (0.88)	2.98 (0.65)
GERMANY	3.25 (0.79)	2.25 (0.59)	2.04 (0.84)	3.10 (0.52)
NIGERIA	3.53 (0.73)	**1.97** (0.57)	1.92 (0.52)	**3.75** (0.57)
TURKEY	2.79 (1.22)	2.40 (0.79)	2.58 (1.18)	3.33 (0.55)
USA	**3.96** (0.45)	2.19 (0.47)	**1.59** (0.42)	3.48 (0.52)
IRAN	2.52 (0.76)	3.03 (0.78)	2.78 (0.53)	3.15 (0.61)
SPAIN	3.03 (1.10)	2.25 (0.74)	2.43 (0.68)	2.95 (0.45)
GS	3.04 (0.97)	2.44 (0.74)	2.28 (0.83)	**3.31** (0.59)
GN	**3.28** (0.89)	**2.31** (0.65)	**2.13** (0.75)	3.15 (0.50)

Table 13: **Human evaluation absolute scores.** Scores shows mean (standard deviation) Likert scores. GT-FAITH, GT-HAL, GT-EXAG, and GT-IMG-REALISM evaluate cultural faithfulness, hallucination, exaggeration and realism of generated images. Results are on 11 countries, 9 activities (1 or 2 sub-activity), 3 T2I images, and 2 annotator per each image/form.

LLM Descriptor Generator — System Prompt

System: You are an expert in cross-cultural visual representation. Your task is to generate precise visual descriptors capturing how a typical scene of a given activity appears in a specific country. Descriptors must cover both traditional and modern variations and represent common culturally accurate scenes.

Rules: 1. The output must strictly follow this JSON structure:
`"descriptors":["token":"...", "style":"traditional|modern|neutral"]`
2. Use culturally-aware terminology (e.g., samovar, sari) when appropriate; use broader cultural phrases when high specificity is unnecessary.
3. Focus only on the core activity scene (not before/after events).
4. Capture multiple common variations where they exist.
5. If the dimension has no representative descriptors, return an empty list.

Table 14: LLM descriptor generator — System Prompt

LLM Descriptor Generator — Setting & Background

Goal: Describe the environment — the physical location, architecture, and design elements that define the atmosphere of the scene.
Guidelines: **INCLUDE**:
– Location and architectural style (indoors/outdoors; temple interior, city street)
– Art and design (calligraphy, geometric tiles, minimalist décor)
– Major furnishings (communal tables, floor cushions, rugs)

EXCLUDE: people, clothing, handheld objects, specific actions.

Generate up to `{max_items}` **descriptors for**:
`{concept}`

Table 15: LLM descriptor generator — Setting & Background

LLM Descriptor Generator — Objects
Goal: Identify the core objects central to the activity. **Guidelines**: - Ensure descriptors accurately represent objects common in the activity scene within the given country. - **INCLUDE**: essential tools, vessels, foods (samovar, board game, hot pot). - Use visually descriptive categories (e.g., "bowls of noodle soup") instead of abstract labels ("Chinese food"). **EXCLUDE**: people, animals, clothing, architecture, actions, background décor. **Generate up to** {`max_items`} **descriptors for**: {`concept`}

Table 16: LLM descriptor generator — Objects

LLM Descriptor Generator — Attire
Goal: Describe typical clothing, accessories, and appearance features. **Guidelines**: - Use specific garment names only when culturally essential (e.g., sari). - Otherwise, use broader cultural categories (e.g., traditional West African attire). - Include both traditional and modern clothing variations unless the concept is strictly historical. **INCLUDE**: garments, headwear, accessories, ceremonial markings, uniforms. **EXCLUDE**: tools, furniture, actions, gestures. **Generate up to** {`max_items`} **descriptors for**: {`concept`}

Table 17: LLM descriptor generator — Attire

LLM Descriptor Generator — Interaction & Gesture
Goal: Capture actions, gestures, and social dynamics central to the activity. **Guidelines**: **INCLUDE**: – Key person–object actions (pouring tea from samovar) – Social gestures (sharing food, group dancing) – Culturally typical postures and formations (kneeling rows) **EXCLUDE**: static object descriptions, clothing, setting details. Focus on actions and interactions. **Generate up to** {`max_items`} **descriptors for**: {`concept`}

Table 18: LLM descriptor generator — Interaction & Gesture

LLM Descriptor Generator — Spatial Arrangement
Goal: Describe layout and spatial organization of people and objects. **Guidelines**: **INCLUDE**: - Positioning of people relative to key objects or surfaces - Culturally meaningful configurations (eating at a table vs. around a sofreh) - Ensure descriptors cover common variations in the activity across the country. **EXCLUDE**: clothing details, object descriptions, actions. **Generate up to** {`max_items`} **descriptors for**: {`concept`}

Table 19: LLM descriptor generator — Spatial Arrangement

LLM Refiner Prompt
System: You refine candidate visual descriptors for evaluating the cultural alignment of AI-generated images representing a specific concept or activity in a given country. Your job is to select, clean, and filter descriptors based on cultural accuracy and relevance. **Task**: Select and refine descriptors according to the concept, country, and descriptor dimension. **Dimensions**: - Setting — venues, architecture, décor - Objects — central objects in the activity - Attire — clothing, accessories, headwear - Interaction — gestures, postures, social relations - Spatial Layout — positioning patterns **Rules**: 1. Keep only culturally accurate descriptors. 2. Create a diverse set covering typical variations. 3. Do not invent new descriptors. 4. Merge duplicates or overly specific items. 5. Remove unrelated descriptors. 6. Keep phrases concise (1–4 words). 7. Descriptors must match the assigned dimension. 8. Output up to {`max_items`} descriptors. 9. If none are valid, return an empty list. **Output Format**: `["token":"item","style":"traditional|modern|neutral"]` **Input**: Concept: {`prompt`} in {`country`} Dimension: {`dimension`} Candidate Descriptors: {`candidate_descriptors`}

Table 20: LLM Refiner Prompt

MLLM Descriptor Exctractor (System Prompt)
As an expert on cross-cultural visual representation, your task is to generate precise visual descriptors to evaluate the cultural alignment and accuracy of AI-generated images. **Goal**: Capture visual elements of a typical scene of an activity in a specific country, covering both traditional and modern variations. **Rules**: 1. Output strictly in JSON: `"descriptors":["token":"...", "style":"traditional|modern|neutral"]` 2. Use culturally-aware terms (e.g., samovar, sari) when precise, or broader cultural terms when sufficient. 3. Focus on the core activity scene—not before or after actions. 4. Capture common variations (e.g., eating at a table vs. sitting on the floor). 5. If nothing distinctive exists, return an empty list.

Table 21: System Prompt for descriptor generation.

MLLM Descriptor Extractor (Setting & Background Prompt)
Goal: Describe the environment (location, architecture, design, furnishings). **INCLUDE**: - Indoors/outdoors (temple interior, busy street, simple home) - Art & design (calligraphy, tiles, minimalist decor) - Major furnishings (floor cushions, rugs, communal tables) **EXCLUDE**: clothing, handheld objects, actions.

Table 22: MLLM descriptor detector (Setting & Background).

Objects Prompt
Goal: Identify key objects, tools, foods, vessels central to the activity. **INCLUDE**: essential items (samovar, board game, noodle bowls, shared hot pot). **EXCLUDE**: animals, clothing, architecture, actions, background décor.

Table 23: Prompt: Objects.

Attire Prompt
Goal: Describe typical clothing, accessories, and appearance. **Rules**: - Use specific garment names only when culturally essential (e.g., sari). - Otherwise, use broader cultural categories. - Always include both traditional and modern possibilities. **INCLUDE**: garments, headwear, accessories, ceremonial markings, uniforms. **EXCLUDE**: tools, furniture, actions, gestures.

Table 24: MLLM descriptor detector (Attire).

Interaction & Gesture Prompt
Goal: Capture actions, gestures, and social dynamics. **INCLUDE**: - Person and/or object actions (pouring tea from a samovar) - Social gestures (sharing food, group dancing) - Group formations (kneeling rows, circle formations) **EXCLUDE**: static objects, clothing, setting.

Table 25: Prompt: Interaction & Gesture.

MLLM Descriptor Detector (Spatial Arrangement)
Goal: Describe the physical layout and positioning of key objects. **INCLUDE**: - Relative positions (sitting around sofreh, standing in line) - Culturally significant layouts (table seating vs. floor seating) **EXCLUDE**: clothing, object details, gestures.

Table 26: MLLM descriptor extractor (Spatial Arrangement).

Prompt

System: You are a helpful assistant that identifies culture-specific visual elements that a text-to-image model may exaggerate when depicting a given activity in a given country. These elements should represent unrealistic overuse, oversized forms, or out-of-place insertions of cultural cues. Return only concise visual stereotypes based on observable features.

Task: Generate up to {{`num_elements`}} candidate exaggerated elements for the activity and country below. Elements must be concise noun phrases describing culture-specific items (e.g., "Persian rugs", "giant flags"). Focus on attire, objects, symbols, decorations, architecture, and interaction styles. Avoid generic or non–culture-specific items (e.g., "trees", "crowds"). Do not describe scenarios or contexts—only list the exaggerated element itself.

ACTIVITY: {{`concept`}}
COUNTRY: {{`country`}}

User: Provide the final list as a single line of noun phrases separated by the "—" character. Do not include numbering, explanations, or extra text.

Example format:

```
Persian rugs | oversize lanterns | giant religious symbols
| people in traditional robes
```

Table 27: EXAG Candidate Generation Prompt

ALIGN Likert Scoring Prompt

System: You are a helpful assistant scoring how well an image depicts an activity as it would typically appear in the specified country. Provide a rating from 1 to 5 using the guidelines below. Only output the score in the required format.

Scoring Scale:
1 = Not well at all (completely wrong or unrelated)
2 = Slightly relevant
3 = Somewhat accurate
4 = Mostly accurate
5 = Exactly as expected for this country

Task:
PROMPT: {{`prompt`}}
COUNTRY: {{`country`}}

User: Respond with ONLY the score in the exact format below:
score is $< num >$

Do not include any other text, explanation, or formatting.

Table 28: ALIGN MLLM-as-a-Judge Prompt

HAL MLLM-as-a-Judge Prompt

System: You are a helpful assistant scoring whether an image contains incorrect elements with respect to the mentioned activity and country. Rate on a scale from 1 to 5 using the rules below. Only output the score in the required format.

Scoring Scale:

1 = Not at all. All elements seem appropriate.

2 = Slightly. One minor element seems foreign.

3 = Somewhat. A few elements feel unusual.

4 = Mostly. Many elements appear incorrect or from another culture.

5 = Strongly. The image is completely wrong, most elements are wrong, the activity is wrong, or the scene belongs to another country.

Task:

PROMPT: {{`prompt`}}

COUNTRY: {{`country`}}

User: Respond with ONLY the score in the exact format below:

score is $< num >$

Do not include any other text, explanation, or formatting.

Table 29: HAL MLLM-as-a-Judge prompt

EXAG MLLM-as-a-Judge Prompt

System: You are a helpful assistant scoring how exaggerated an image is in terms of cultural elements (e.g., attire, objects, background, art/design). Rate on a scale from 1 to 5 using the rules below. Only output the score in the required format.

Scoring Scale:

1 = Appropriate and balanced.

2 = Slightly decorative but within cultural norms.

3 = Several features feel idealized or overly emphasized.

4 = Many elements feel exaggerated and unrealistic.

5 = Strongly stereotypical, misleading, or culturally irrelevant.

Task:

PROMPT: {{`prompt`}}

COUNTRY: {{`country`}}

User: Respond with ONLY the score in the exact format below:

score is $< num >$

Do not include any other text, explanation, or formatting.

Table 30: EXAG MLLM-as-a-Judge Prompt

Image Editing Instruction Prompt
Task: Edit the image to correctly show {{`activity`}} in {{`country`}} by following the instructions below. **Remove**: - All hallucinated elements from the HAL list - All exaggerated elements from the EXAG list - Any objects, clothing, poses, or background features that belong to the wrong culture, historical period, or activity **Add/Preserve**: - ALIGN list that must remain present - Add correct interaction from REF DESCP list if mentioned in HAL/EXAG - ADD 1-3 different types of attire from REF DESCP if mentioned in HAL/EXAG - ADD 1 correct background from REF DESCP if mentioned in HAL/EXAG **Input**: ACTIVITY: {{`activity`}} COUNTRY: {{`country`}} HAL DESCRIPTORS: {{`HAL`}} EXAG DESCRIPTORS: {{`EXAG`}} ALIGN DESCRIPTORS: {{`ALIGN`}} REFERENCE DESCRIPTORS: {{`REF DESCP`}}

Table 31: Image editing instruction prompt template. AHEaD feedback (HAL, EXAG, ALIGN) combined with reference descriptors $\mathcal{D}^{\text{llm}}$ guides image editing to improve cultural accuracy.