CAIRE: Cultural Attribution of Images by Retrieval-Augmented Evaluation

Arnav Yayavaram*[1] Siddharth Yayavaram*[1] Simran Khanuja[2] Michael Saxon[3] Graham Neubig[2]
[1]BITS Pilani [2]Carnegie Mellon University [3]University of California, Santa Barbara
arnav.yayavaram@gmail.com, siddharth.yayavaram@gmail.com, skhanuja@andrew.cmu.edu, saxon@ucsb.edu, gneubig@cs.cmu.edu

Abstract

As text-to-image models become increasingly prevalent, ensuring their equitable performance across diverse cultural contexts is critical. Efforts to mitigate cross-cultural biases have been hampered by trade-offs, including a loss in performance, factual inaccuracies, or offensive outputs. Despite widespread recognition of these challenges,[1] an inability to reliably measure these biases has stalled progress. To address this gap, we introduce CAIRE*, a novel evaluation metric that assesses the degree of cultural relevance of an image, given a user-defined set of labels. Our framework grounds entities and concepts in the image to a knowledge base and uses factual information to give independent graded judgments for each culture label. On a manually curated dataset of culturally salient but rare items built using language models,* CAIRE *surpasses all baselines by* ***28%*** *F1 points. Additionally, we construct two datasets for culturally universal concepts, one comprising of T2I generated outputs and another retrieved from naturally-occurring data.* CAIRE *achieves Pearson's correlations of* ***0.56*** *and* ***0.66*** *with human ratings on these sets, based on a 5-point Likert scale of cultural relevance. This demonstrates its strong alignment with human judgment across diverse image sources. Our code is here,[2] and our tool will be open sourced upon publication.*

1. Introduction

Current text-to-image (T2I) models, despite their high-quality outputs, have serious issues in cultural representation and sensitivity. Previous research has reported that they produce culturally homogeneous outputs given under-specified prompts [23, 40] and their outputs are disproportionately biased toward Western cultures, failing to depict global diversity [42]. Further, they generate stereotypical, offensive and factually incorrect outputs when asked to be culturally inclusive [4, 8, 20, 48]. The proliferation of AI-generated content, particularly images, is reshaping our digital media ecosystem, with over 15 billion AI-generated images created since 2022.[3] Given this widespread adoption, it is imperative to ensure that these models are inclusive and unbiased towards users from diverse backgrounds.

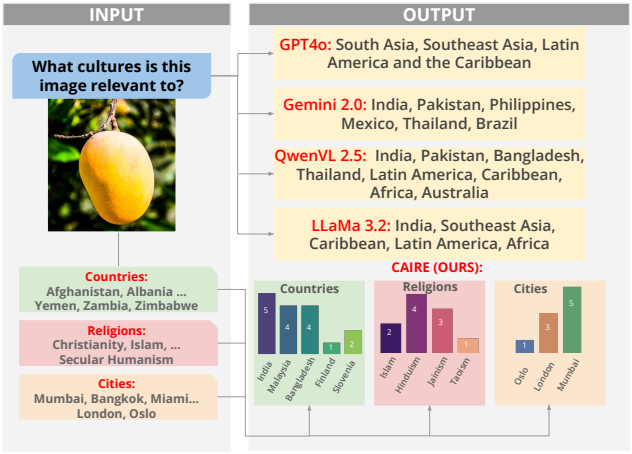


Figure 1. We propose CAIRE, a novel evaluation metric that assesses the *degree of cultural relevance* of an image, given a user-defined set of labels. Unlike existing methods that assume a definition of culture, we let the user specify the proxy of culture such as countries, religions, or cities to assess over as free-text labels.

Moving towards this ideal requires us to first develop robust evaluation metrics that can identify and quantify these cultural representation gaps. But how do we define *culture*? In the social sciences, culture is a complex concept that can refer to cultural heritage [5], social interactions [31], or ways of life [36]. It transcends basic categorizations, spanning ethnicities, communities, and even subtle neighborhood distinctions like those between upper and downtown Manhattan [6, 9], and is difficult to define concretely because it varies by context. Every individual and group lies at the intersection of multiple cultures (defined by their political, professional, religious, regional, class-based and other affiliations) and these are invoked according to the sit-

*Equal contribution

[1]A well-known case where a company had to redact image generation features due to such issues: https://tinyurl.com/yc5jjk64

[2]https://anonymous.4open.science/r/CAIRE-1618

[3]https://wired.me/culture/ai-image/

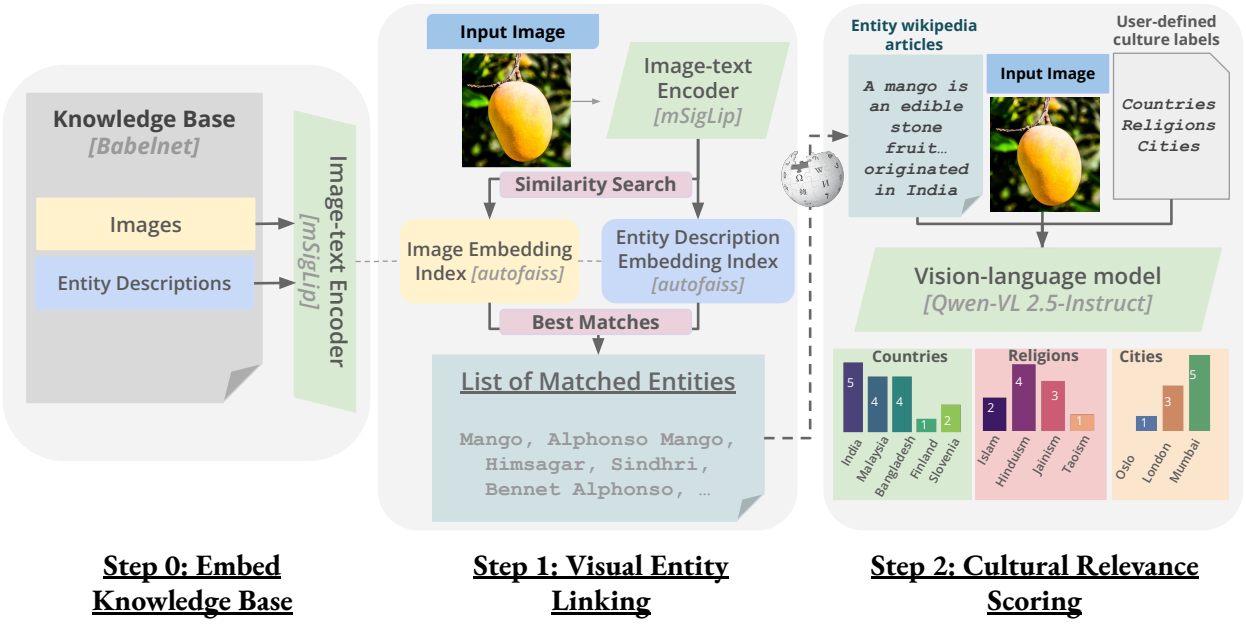


Figure 2. Overview of CAIRE. From an image-indexed multimodal knowledge base, we embed an input image to retrieve entities that are tied to Wikipedia articles. From the text of those Wikipedia articles and the query image, a vision-language model (VLM) generates an affinity score to each user-specified candidate culture label. A detailed description of our framework is in §3.

uation, typically in contrast to other groups [1]. Despite this rich and nuanced complexity, most AI research defaults to using countries as proxies for culture – indeed, this is one of the most frequently cited limitations across these same papers [57]. Cultural AI work also struggles with capturing culture's inherently dynamic nature. While culture continuously evolves through social negotiation [33], existing benchmarks remain largely static collections of examples or facts [21, 24, 27, 44]. To address both the fluidity of cultural boundaries and their evolution over time, it is necessary to allow users to define culture on their own terms through natural language descriptions, creating a framework that can adapt as cultures themselves transform.

In this paper we introduce **C**ultural **A**ttribution of **I**mages by **Re**trieval (CAIRE), a family of evaluation metrics for visual cultural attribution. CAIRE evaluates images over user-defined cultural labels, by grounding them to entities and concepts in a knowledge-base (§3). Our approach takes as input an image and a list of free-text labels representing cultures of interest. It then outputs a score on a five point scale, indicating the relevance of the image to each label in the culture set (Figure 1). The framework operates in two key stages: First, it grounds the image in real-world concepts and objects, leveraging a massively multilingual and multicultural knowledge base. Second, it utilizes a vision-language model (VLM) to estimate cultural relevance in a retrieval-augmented evaluation setup, where all available information about the recognized concepts is leveraged to estimate this score. Importantly, our method is designed as a flexible framework, allowing users to integrate their preferred knowledge bases or VLMs to suit their specific needs and contexts.

All previous works which have attempted to evaluate geographical [17] or cultural diversity [23] of images assign a *single* country/region culture label to an image with a binary (relevant / not relevant) score. In contrast, CAIRE estimates the *degree* of relevance across *multiple* user-defined culture labels. Due to the absence of test sets with such annotations, we first construct a dataset (§5) of rare and culturally significant items, labeled using GPT-4o [34] and verified with Wikipedia. On this test set CAIRE surpasses baselines by **28%** F1 points. After validating CAIRE on this dataset, we evaluate its correlation with human judgments of cultural relevance across a broad range of universal concepts. CAIRE is adept at capturing annotators' judgments of the relevance of these images to their own cultures, with Pearson's correlations of **0.56** and **0.66** to these human ratings over generated and natural images, respectively, demonstrating its alignment with human judgment across diverse image sources (§6). In summary, we:

- Formalize the task of **visual cultural attribution**—providing graded assessments on how relevant an image is to a user-defined set of free-text culture labels.
- Build two **complementary test sets** using our formalism to test visual cultural attribution methods. The first covers obscure concepts spanning a variety of culture proxies

and is labeled using GPT-4o, while the second covers universal concepts and is annotated with human judgment on a five-point scale.[4]
- Introduce the **CAIRE framework** built using our design principles, which leverages vision-language models, massive KBs, and image-based retrieval to perform visual cultural attribution.

2. Task Formulation

We define *visual cultural attribution* as the task of assigning cultural relevance scores to an image, given a user-defined set of culture labels. Defining the relevance of an image with respect to a particular culture presents several challenges. An effective framework must:

1. **Allow flexible definition of culture labels:** Culture cannot always be represented by simple country or region-based labels; finer-grained community, ethnic, or social group labels must be considered. Thus it is necessary to allow labels to be defined flexibly using natural language.
2. **Provide graded judgments:** Cultural relevance is not binary; many cultural elements are shared across different groups to varying extents.

Formally, given an input image I and a set of cultural labels $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$ (such as the country or religion labels in Figure 1), we define the scoring function $f : I \times \mathcal{C} \rightarrow [1, 5]$ that outputs an integer *cultural relevance score* on a five-point scale. A higher score signifies high cultural relevance of the image to the culture. Prior work (§7) either assigns binary relevance scores to images, or only allows scoring over a fixed cultural proxy label set, thereby failing to adhere to our task formulation.

3. CAIRE Metric Framework

In this section, we detail how CAIRE performs visual cultural attribution using information about similar images to the input retrieved from a knowledge base (KB). CAIRE is based on the premise that cultural identity can be signaled by a broad range of elements, including objects, attire, architecture, symbols, and text. Thus, CAIRE processes each image in two steps: *a) Visual Entity Linking (VEL)*: retrieving relevant KB entries using image embedding-based queries [13], and *b) Cultural Relevance Scoring*: using VLM judgments over a user-specified set of candidate culture labels. An overview of this pipeline is in Figure 2.

3.1. Visual Entity Linking (VEL)

In this step CAIRE identifies entities and concepts present in the image to ground them in our KB. It uses these to retrieve Wikipedia articles which condition the LM in the cultural relevance scoring step. Our VEL method is inspired by recent works leveraging vision-language encoders like CLIP to directly perform image-image or image-text retrieval given the input image as the query and KB images and text descriptions as the keys [18, 45].

Creating a VEL system requires building an image-indexed KB using some vision-language encoder (*step 0* in Figure 2). In this paper index nodes in the massively multilingual BabelNet graph [32] by their images' and texts' separate mSigLIP embeddings [55]. We generate two FAISS [7] indices: one for the images and another for the text linked to each entity, with each index containing 6 million entries. Note that the framework design allows for a user to use custom KBs and vision-language encoders, making CAIRE generalizable across a wide variety of domains.

Now, when provided with an input image I, we again embed the input image using mSigLIP (*query*). We retrieve the top 20 most similar images to our query by searching through the KB image index. Each retrieved image corresponds to *at least* one entry in the knowledge graph, but may correspond to multiple too. For example, a picture of a mango matched to Figure 2 might link to both the `mango` and `fruit` nodes—we need to select the most precise and representative one to provide adequate information for a cultural assessment.

Each node contains lemmas—disambiguated single-sense WordNet definitions [11]—which we use to identify the best entry to source documents from. To do this, we perform image-text matching over the lemma definitions and our input to re-rank the retrieved KB IDs. We empirically found that this method performs well in a pilot study using the FOCI benchmark [14] of long-tail entity recognition, where we compared it to other KB disambiguation and lemma matching techniques, detailed in Appendix §A.3.

3.2. Cultural Relevance Scoring

Once we link an image to its associated entities in a KB, we leverage LLMs and VLMs augmented with the knowledge of these entities to estimate how relevant the image is to each of the culture labels in $\mathcal{C}$. In particular, we use the full text of relevant Wikipedia articles for each entity and the query image itself (for VLMs) to produce cultural assessments conditioned on candidate culture labels.

Rather than prompting the model to choose cultural relevance labels directly conditioned on the image or text, we instead prompt it to independently rate each possible *candidate culture* in the user-defined set which may be a list of countries, ethnic groups, subcultures, philosophies, or whatever the user defines cultural groups to be. Querying the model using candidate culture labels allows us to sample scores estimating the *degree* of relevance, rather than simply predicting whether an image is *relevant / not relevant* for a particular culture.

[4] add b4 arxiv Both datasets and labels to be released upon publication

Following previous work on LLMs as judges [28], which rely on both numerical scoring [22] and token-likelihood [53] approaches to make judgments, we experimented with both 1–5 point scoring as well as log-likelihood based scoring. In the **numerical scoring** setup, we directly prompt the LM to score the cultural relevance of an image on a 1-5 scale, given the corresponding Wikipedia text, the query image when applicable, and a rubric designed to elicit consistent expert judgments. A full input example, including prompt and sample text, is shown in the Appendix (Figure 8).

To ensure well-formatted, unambiguous outputs, we perform constrained decoding, exclusively considering tokens that are directly attributable to a score on this scale (eg., "1", "2", ...). The numerical score corresponding to the maximum likelihood token in this set is taken to be the cultural affinity score for that (image, culture) pair.

In practice, we find that numerical scoring works better than converting raw log-likelihoods to a relevance rating, and is also practically feasible to correlate with human judgment. We describe, discuss, and compare both methods in Appendix §B.1.

4. Test Set Curation

Given our reformulation of the visual cultural attribution task to include dynamic definitions of culture and provide graded judgments on relevance (§2), it is necessary to have a test set that reflects this design. However, all test sets for geographical and cultural diversity of images thus far have provided for *single, binary* country annotations [17, 23] (§7). To bridge this gap, we create two complementary test datasets—**`specific`** and **`universal`**—that capture CAIRE's desiderata.

The `specific` set contains real images corresponding to a diverse set of culturally-specific (and sometimes quite obscure) entities. This set lets us 'sanity check' the reliability of CAIRE in capturing these rare entities. We construct labels for this set using GPT-4o, allowing us to assign labels across a diverse set of cultural proxies, beyond geographical regions (*desideratum 1* in §2).

The `universal` set contains images related to culturally universal concepts (eg., *food*, *ceremony*) annotated for cultural relevance by 200 annotators from 10 different countries. It contains two splits: `generated` and `retrieved`, containing T2I-generated and natural images corresponding to each concept, in each country. Each image is rated by each annotator on five-point scale for cultural relevance to their country. Since our data annotation platform provides only country-level metadata, we aggregate graded judgments from multiple countries, satisfying *desideratum 2* in §2 MS **:I don't get what this sentence means**

4.1. `specific` Concept Test Set

The `specific` set contains 68 concepts [19] produced using GPT-4o. We prompt GPT-4o to provide a list of concepts that are rare, yet significant to different proxies of culture, such as regions, religions, festival, ethnicities, philosophies, etc., as laid out in Adilazuarda et al. [1]. This produces a diverse set of entities across many candidate culture label sets, including *countries of the world, ethnicities of Africa, states of India, cities of Indonesia, world religions, Native American tribes, and Bronze Age civilizations.* Given one proxy, like *cities of Indonesia*, an image can have multiple gold labels indicating that its relevant to multiple cities of Indonesia. The complete set is provided in Appendix §C.6. Each generated concept and its labels are validated through an affirmative match to an existing Wikipedia article.

Next, we manually collect CC-licensed images from the web for each concept, ensuring that there is no exact match with images in the KB.[5] Samples from this test set with its respective culture proxies is shown in Figure 3. We avoid using images present in BabelNet (and thus our index) to avoid test set contamination.

Metric: Since GPT-4o doesn't reliably provide for numerical ratings that indicate the degree of relevance, and there is no objective basis by which we could assign such numerical labels, we instead frame this as a multi-label classification task, testing whether CAIRE can correctly identify whether or not a concept is relevant to all culture labels. F1 scores are reported for all baselines and CAIRE described in §5.

4.2. `universal` Concept Test Set

The curated `specific` test set above includes multi-cultural annotations for each concept, but doesn't allow us to measure the accuracy of CAIRE in estimating the *degree* of culture relevance, and how closely it matches human judgment. Hence, we curate a list of universal concepts that are common across multiple cultures (like *breakfast, weddings, rituals, etc.*) and collect both natural and generated images for selected concepts. Using these images, we run a human rating study on Prolific,[6] where we collect cultural relevance ratings for each image on a scale of 1 to 5. Our data curation process is detailed below:

- *Selecting Concepts*: We adopt 20 visually depictable and culturally universal concepts[7] provided by Bhatia et al. [3]. The full list is in Appendix §C.2.
- *Selecting Regions*: We recruit annotators from *China, India, United States, Brazil, Nigeria, Russia, Mexico,*

[5] The dataset will be open sourced along with the code

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

[7] From a set of 298 human universals available here: https://condor.depaul.edu/~mfiddler/hyphen/humunivers.htm

(a) `specific set (countries)`

(b) `universal-generated (countries)`

(c) `universal-retrieved (countries)`

(d) `specific set (religions)`

(e) `specific set (Bronze Age civilizations)`

(f) `specific set (cities of Indonesia)`

Figure 3. Examples from the evaluation set: (a) image from the `specific` set depicting Azulejos. The label set of countries consists of `Portugal`, `Spain`, `Brazil`, `Morocco`, `Mexico`. (b) T2I-generated image using the prompt "A realistic photo of a ritual in `India`," representing the universal-generated subset. (c) image retrieved from DataComp 1B, using the text query "A realistic photo of greetings in `Nigeria`", illustrating the universal-retrieved subset. (d-f) additional examples from the `specific` set corresponding to cultural proxies religion, Bronze Age civilizations, and cities of Indonesia, (Buddhism, Sumer, and Magelang respectively).

Egypt, Germany, and Indonesia, geographically diverse countries that have many available annotators.

- *Generated data* (`universal-generated`): We use Stable Diffusion 3 Medium (SD3) [10] to generate images. We use a simple prompt format: `A realistic photo of CONCEPT in COUNTRY`, for all combinations of concepts and countries listed above, producing 200 prompts (20 concepts × 10 countries) and 200 images per model.
- *Natural data* (`universal-natural`): We obtain mSigLIP embeddings for Datacomp-1B [12] and retrieve the top-8 images for the same 200 prompts used in `universal-generated`. We manually select a single high-quality image from the top-8 images, since retrieval may be noisy for certain prompts.

Human Annotation: We evaluate 400 images (200 generated from SD3 (`universal-generated`) and 200 retrieved from DataComp-1B (`universal-natural`)). Each image receives a cultural relevance rating on a 1-5 Likert scale by 20 annotators each across the 10 countries mentioned above. The definition of each rating is shown in Appendix C.3 (Table 12). Each image is assigned a cultural relevance score for each country by averaging the Likert scores given by annotators in said country.

Metric: We compute Pearson's correlation coefficient between the absolute cultural relevance scores assigned by human raters and those predicted by our tool for each image. This metric quantifies the degree of linear association between the predicted scores and human ratings, where a high correlation indicates that our tool not only ranks countries accurately but also assigns scores that align closely with human perceptions of cultural relevance.

5. Experimental Setup

Since CAIRE is the first work proposing a metric for assessing the degree of cultural relevance of an image, we benchmark its performance against simple multimodal encoder and VLM baselines. We set up these baselines as follows:

5.1. Baseline 1: Vision-Language Encoders

The first baseline uses embeddings of the input image and a set of textual *probe prompts* using either CLIP ViT-B/32 [38] or mSigLIP [55] to assign a culture label.

The probe prompts are constructed from the user-defined set of *candidate culture labels* and a set of five *relevance level* phrases by combining them into the string "`This image is {relevance level} to {culture candidate}`." Here, *relevance level* is one of: *Not relevant*, *Minimally relevant*, *Somewhat relevant*, *Relevant*, *Highly relevant*, following the template shown in Table 12.

We compute the cosine similarity between the input image and each of the probe prompt embeddings, selecting the label with the highest similarity score as the final rating.

5.2. Baseline 2: Vision-Language Models

We prompt three recent and popular vision-language models (VLMs) to provide a graded score on cultural rele-

vance: Llama-3.2-11B-Vision-Instruct [15], Qwen2.5-VL-7B-Instruct [50], and Pangea-7B-hf [54]. The prompt includes the target culture along with a description of the relevance levels, as defined in Table 12. We encourage the model to follow chain-of-thought reasoning to obtain best results. The full prompt template is provided in Figure 8.

5.3. Our CAIRE Implementations

We test LLama-3.2-11B, Qwen2.5-VL-7B, and Pangea-7B-hf [15, 50, 54] as components in CAIRE metrics. In the experiments involving just the LLM variant, we intentionally omit the original query image to evaluate the impact of including the image on model performance.

This is done to isolate the effect of image input and assess its contribution to the model's understanding. In contrast, the experiments with VLMs provide both the image and the context derived from the entity linking (VEL) stage. We experiment with two possible ways of augmenting the LLMs and VLMs with additional contextual information:

- *Wikipedia Text Augmentation:* We begin by incorporating the Wikipedia text content corresponding to the best matching entity ID, identified through lemma matching.
- *Top-K Entity Title Augmentation:* We provide a list of the Wikipedia titles of the top 20 entities retrieved during the entity linking. Unlike the first method, we do not include detailed descriptions of these entities and only supply their titles as supplementary context.

5.4. Evaluation

Visual cultural attribution entails producing a cultural relevance score on a five-point scale; however there is no objective basis by which the examples in the `specific` set should be assigned numerical relevance scores. Thus, to evaluate performance on the `specific` set we convert all metrics' outputs into binary judgments by assigning scores greater than 4 on the five-point scale (*relevant, highly relevant*) to positive label 1, and the others 0. We report F1 scores for each metric based on these in Table 1.

For the `universal` human-labeled test set where we collect ratings on a scale of 1-5, we directly measure Pearson's correlations between the baseline's score and average human ratings, which can be found in Table 2.

6. Results and Analysis

Table 1 presents the F1 scores for the CAIRE metrics and baselines over the gold-labeled `specific` test set. Table 2 presents the Pearson correlation coefficients of each model against the human-labeled cultural relevance judgments in the `universal` set. For compact presentation, all scores are multiplied by 100. Our analysis of CAIRE's performance against the baselines should answer six questions:

Model	Img.	Wiki.	Top-20	F1	Δ_{CAIRE}
Vision-Language Encoders (*Baseline-1*)					
CLIP	✓			3.3	-
mSigLIP	✓			*12.8*	-
Vision-Language Models (*Baseline-2*)					
Llama-3.2-11B-Vis.-Ins.	✓			40.8	-
Qwen2.5-VL-7B-Ins.	✓			*41.0*	-
Pangea-7B-hf	✓			20.3	-
CAIRE					
Llama-3.2-11B-Vis.-Ins.	✓	✓		47.4	(+6.6)
	✓		✓	47.9	(+7.1)
Qwen2.5-7B-Ins.		✓		**68.9**	(+27.9)
			✓	52.1	(+11.1)
Qwen2.5-VL-7B-Ins.	✓	✓		65.5	(+24.5)
	✓		✓	54.8	(+13.8)
Pangea-7B-hf	✓	✓		42.9	(+22.6)
	✓		✓	29.0	(+8.7)

Table 1. F1-scores on the `specific` set. Δ_{CAIRE} represents the improvement CAIRE provides over the naive baseline with that same LM. *Img.* indicates the input image, *Wiki.* represents the wikipedia content of the top-matched entity, while *Top-20* refers to the names of the top-20 matched entities.

Does CAIRE correctly attribute specific entities to their respective cultures? The `specific` set checks this using particularly difficult real-world example concepts. Table 1 shows that CAIRE consistently outperforms the open-source LM and VLM baselines at this binary cultural relevance classification task. Among all VLMs (*Baselines-2*), Qwen2.5-VL-7B-Instruct achieves the highest F1-score of 41.0 without additional context, representing the strongest baseline result. Integrating this VLM into a CAIRE metric significantly improves performance.

How well does CAIRE model image-level human opinions? The `universal` set images provide natural subjective annotator opinions of an image's relevance to their own culture. Thus, correlation between metric judgments and human labels over this set characterizes how well the metric models these opinions. These results are presented in Table 2. The strongest CAIRE metric on the `specific` set, Qwen2.5–VL-7b-Instruct, is also most performant on the `universal` sets, achieving Pearson's correlations of **0.56** and **0.66** on average across all countries. Note that the performance difference between the VLM baselines and CAIRE are not as pronounced on this set—this may be due to the relative commonality of concepts present in the `universal` images to the `specific` ones (Figure 3).

Does CAIRE's model of human opinions generalize across diverse cultures? Most columns in Table 2 con-

Model	Brazil		China		Egypt		Germany		India		Indonesia		Mexico		Nigeria		Russia		USA		Avg.	
	(N)	(G)	(N)	(G)	(N)	(G)	(N)	(G)	(N)	(G)	(N)	(G)	(N)	(G)	(N)	(G)	(N)	(G)	(N)	(G)	(N)	(G)
Vision-Language Encoders (*Baseline-1*)																						
CLIP	-0.04	0.38	0.04	0.16	0.09	0.07	-0.17	-0.09	-0.05	0.12	0.28	0.33	-0.06	0.07	-0.02	0.20	-0.10	0.08	-0.02	0.11	-0.02	0.11
mSigLIP	0.07	-0.02	-0.01	-0.15	0.22	0.08	0.15	-0.04	-0.13	-0.01	0.16	-0.20	0.09	0.05	0.20	-0.02	0.15	0.21	0.12	0.03	0.10	-0.01
Vision-Language Models (*Baseline-2*)																						
Llama-3.2-11B-Vis.-Ins.	0.54	0.18	0.57	0.37	0.31	0.29	0.65	0.27	0.59	0.34	0.56	0.33	0.45	0.11	0.68	0.56	0.57	0.01	0.52	0.28	0.55	0.31
Qwen2.5-VL-7B-Ins.	0.59	0.51	0.59	0.37	0.54	0.52	0.56	0.54	0.76	0.72	0.64	0.60	0.52	0.39	0.62	0.61	0.69	0.33	0.66	0.63	0.61	0.52
Pangea-7B-hf	0.56	0.39	0.60	0.50	0.54	0.58	0.62	0.52	0.72	0.66	0.56	0.47	0.50	0.40	0.46	0.60	0.61	0.39	0.64	0.59	0.58	0.51
CAIRe																						
Llama-3.2-11B-Vis.-Ins.	0.64	0.47	0.61	*0.30*	0.52	0.49	0.70	0.58	0.69	0.70	0.65	0.44	0.57	0.33	0.47	0.49	0.74	0.53	0.65	0.61	0.63	0.49
Qwen2.5-7B-Ins.	0.57	0.41	0.66	0.34	*0.22*	0.43	0.48	0.34	0.62	0.67	0.67	0.47	0.55	0.36	0.42	0.44	0.66	0.29	0.48	0.42	0.53	0.42
Qwen2.5-VL-7B-Ins.	0.67	0.63	0.68	0.53	0.60	0.50	0.65	0.58	0.74	0.68	0.70	0.53	0.58	*0.46*	0.61	0.55	0.74	0.56	0.59	0.62	**0.66**	**0.56**
Pangea-7B-hf	0.70	0.49	0.64	0.45	0.61	0.56	0.62	0.51	0.52	0.65	0.58	0.55	0.59	*0.39*	0.64	0.64	0.61	0.47	0.61	0.47	0.61	0.52

Table 2. Country-wise Pearson's correlation with human judgment. **(N)** represents natural data, while **(G)** represents generated data. The highest average scores are in **bold**, with the highest and lowest country-wise values underlined and *italicized*, respectively. A detailed analysis of the results is provided in §6.

tain the correlation scores averaged over all concepts within specific countries. The lowest and highest values of correlation for each model are marked using underlines and *italics*. Generally, CAIRe is more performant for India, Germany and Russia compared to Mexico, Egypt and Nigeria. This can be attributed to distributional inequalities in mSigLIP's training data and how well-populated the KB is for all these regions. This may also be a result of disparities in the representativeness of the presented images in the test set. We also see that competent VLMs like Qwen display lesser variation across countries (0.5-0.7), while LLaMa displays higher variance. Overall, this suggests that with better models and larger KBs, CAIRe can consistently improve with time.

How effective is CAIRe in visual entity linking, i.e., matching images to the right entities? The VEL system successfully matches 86% of instances in the `specific` set, to their exact labeled concepts, as verified through manual evaluation. In instances where an exact match is not achieved, the retrieved concepts often exhibit semantic or cultural similarity to the query image, with only a minimal number of cases being entirely incorrect. We present a few of these examples in §C.1, with most inaccuracies arising from concepts that are absent in the knowledge base.

To estimate an upper bound on how much VEL performance impacts the overall F1 score in the case of the `specific set`, we incorporate gold-standard Wikipedia pages as context in our cultural relevance scoring, allowing us to quantify retrieval-induced loss (detailed results in §C.5: Table 10).

Which variant of CAIRe performs best? Qwen2.5-7B-Instruct outperforms all other methods on the `specific` set, despite not using the image. A possible explanation is that for difficult entities where Qwen-VL may disagree with the retrieved content, errors are more likely to occur. Since Qwen2.5-7B-Instruct relies solely on retrieved text, this may contribute to its slightly superior performance. Additionally, methods that incorporate the full content of the most relevant Wikipedia page generally perform better than those that only provide the top 20 retrieved entities without any description. While passing 20 entities may help the model develop a broader understanding of the target entity, it can also introduce entities with conflicting cultural relevance, potentially leading to confusion. For the `universal` set, Qwen2.5-VL-7B-Instruct consistently achieves the highest weighted average Pearson correlation across all cultures for both natural and generated images, and also the lowest variance across countries, indicating its dominance compared to other VLMs.

How does performance vary between natural and generated images? Per Table 2, it's clear that CAIRe is more effective in evaluating the cultural relevance of natural images as compared to generated ones. This is expected since arises our KB comprises exclusively natural images and mSigLIP is also trained on naturally-occurring data. Furthermore, the system demonstrates strong performance on the challenging `specific` set comprising of natural images, reinforcing this observation. In the case of generated images, the depicted entities do not always correspond to real-world concepts or objects, even when the model is explicitly prompted to generate "realistic" images. However, as generative model capabilities advance, both in terms of realism and in accurately depicting real-world entities, this performance gap is expected to diminish.

7. Related Work

Culturally-diverse image datasets Several studies have explored the need for culturally diverse image datasets to better assess biases in text-to-image (T2I) models. Recent work [52] demonstrated that incorporating culturally

and linguistically diverse data in training can enhance the fairness and accuracy of visual representations. The GeoDE dataset [39] contains 61,940 images of common objects from geographically diverse regions to evaluate object recognition systems otherwise trained on western-centric web-scraped images. Jha et al. [20] developed ViSAGe, a dataset of T2I generated outputs, human-annotated for visual stereotypes. Khanuja et al. [25] introduce a new task and test set to evaluate whether image-editing models are capable of localizing images for a target culture. Finally, Bhatia et al. [3] create a benchmark that introduces two challenging tasks to test for cultural inclusion in vision-text models: retrieval across universals and cultural visual grounding. Unlike most of these works that have treated geographical regions as a proxy for culture and assigned single culture labels to each image, our datasets include multi-culture labels with a human rating on how relevant an image might be to each culture in the set.

Evaluating for fairness and diversity Prior work has attempted to assess biases and fairness in T2I model outputs. DIG In [17] and Decomposed-DIG [46] evaluate whether T2I models generate geographically diverse outputs for single-object prompts, revealing representational disparities. Basu et al. [2] highlighted the Western-centric bias in these models, while Ventura et al. [47] examined embedded cultural perspectives, advocating for improved evaluation frameworks.

CUBE [23] assesses cultural diversity using a manually curated reference set limited to eight countries and three domains with 300k artifacts. In contrast, our framework leverages a knowledge base spanning 6M concepts, enabling broader assessments. Unlike prior work that primarily evaluates batch-level diversity and focuses only on generated images, our approach assesses cultural relevance at the individual image level and works with both generated and natural images. Finally, while most existing evaluations assign a single country label per image, we introduce a more flexible and inclusive method, allowing multiple cultural proxies to define and assess cultural relevance.

Work on multilingualism in text-to-image models overlaps with themes of cultural evaluation. Saxon and Wang [42] introduced an evaluation of cross-lingual capabilities of a slate of T2I models using image similarity over cultural universals. This line of work has faced challenges for a lack of techniques like CAIRE to perform cultural attribution [41, 43].

8. Conclusion

From our formulation of the visual cultural attribution task, we have produced two test sets that capture important desiderata: specific accuracy over rare entities and faithful modeling of human opinions over diverse cultures.

Our CAIRE family of visual cultural attribution metrics is the first framework in this space that permits assessments over free-text culture labels (open vocabulary) and provides numerical scores rather than binary judgments. By leveraging a massively multilingual, multimodal knowledge base and state-of-the-art vision-language encoders, CAIRE can account for rare entities, and using those retrieved KB entries as input to LMs for final judgment, we robustly provide these numerical scores. Our results demonstrate that CAIRE beats naive LM-based approaches on both test sets, and satisfies our need for a robust visual cultural attribution tool.

With CAIRE, we significantly expand the scope of concepts we can evaluate for cultural diversity, since we use the KB images as our reference set for grounding. Given its modularity and leverage of state-of-the-art LMs, CAIRE can continue to evolve going forward and enable research into cultural issues in multimodality.

9. Limitations & Ethical Considerations

Biases from using Wikipedia Content. Wikipedia is a valuable resource for training language models due to its breadth and structured content, but it also introduces biases. Its coverage reflects the interests of its editor community, leading to variations in topic emphasis and linguistic framing. Additionally, its reliance on verifiable sources can favor well-documented perspectives over emerging viewpoints. These factors influence LM outputs, reinforcing the need for careful interpretation.

Potential Reinforcement of Cultural Stereotypes. While CAIRE aims to provide a nuanced evaluation of cultural relevance, it relies on knowledge bases and vision-language models that may contain biases. If not carefully calibrated, the framework could inadvertently reinforce existing stereotypes by overemphasizing certain cultural elements while under-representing others.

Subjectivity in Cultural Attribution. Defining cultural relevance is inherently subjective, and different users may have differing perspectives on whether an image aligns with a particular culture as shown in §subsection D.2. While CAIRE allows for flexible cultural definitions, the subjectivity in user-defined labels and scoring necessitates caution in interpreting results, especially in sensitive contexts.

References

[1] Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Ashutosh Dwivedi, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. Towards measuring and modeling" culture" in llms: A survey. *arXiv preprint arXiv:2403.15412*, 2024. 2, 4

[2] Abhipsa Basu, R Venkatesh Babu, and Danish Pruthi. Inspecting the geographical representativeness of images from text-to-image models. *arXiv preprint arXiv:2305.11080*, 2023. 8

[3] Mehar Bhatia, Sahithya Ravi, Aditya Chinchure, Eunjeong Hwang, and Vered Shwartz. From local concepts to universals: Evaluating the multicultural understanding of vision-language models. *arXiv preprint arXiv:2407.00263*, 2024. 4, 8, 15

[4] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, page 1493–1504, New York, NY, USA, 2023. Association for Computing Machinery. 1

[5] Janet Blake. On defining the cultural heritage. *International & Comparative Law Quarterly*, 49(1):61–85, 2000. 1

[6] Mary Bucholtz and Kira Hall. Identity and interaction: A sociocultural linguistic approach. *Discourse studies*, 7(4-5): 585–614, 2005. 1

[7] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *arXiv preprint arXiv:2401.08281*, 2024. 3

[8] Sean Dudley and Al Kuslikis. Opportunity and risk: Artificial intelligence and indian country. *Tribal College: Journal of American Indian Higher Education*, 36(2), 2024. 1

[9] Penelope Eckert. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual review of Anthropology*, 41(1):87–100, 2012. 1

[10] 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. 5

[11] Christiane Fellbaum. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer, 2010. 3

[12] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024. 5

[13] Jingru Gan, Jinchang Luo, Haiwei Wang, Shuhui Wang, W. He, and Qingming Huang. Multimodal entity linking: A new dataset and a baseline. *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 3

[14] Gregor Geigle, Radu Timofte, and Goran Glavaš. African or european swallow? benchmarking large vision-language models for fine-grained object classification, 2024. 3, 12

[15] Aaron Grattafiori. The llama 3 herd of models, 2024. 6

[16] Agrim Gupta, Piotr Dollár, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation, 2019. 12

[17] Melissa Hall, Candace Ross, Adina Williams, Nicolas Carion, Michal Drozdzal, and Adriana Romero Soriano. Dig in: Evaluating disparities in image generations with indicators for geographic diversity, 2024. 2, 4, 8

[18] Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12065–12075, 2023. 3

[19] 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. 4

[20] Akshita Jha, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Rida Qadri, Chandan K. Reddy, and Sunipa Dev. Visage: A global-scale analysis of visual stereotypes in text-to-image generation, 2024. 1, 8

[21] Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. Kobbq: Korean bias benchmark for question answering. *Transactions of the Association for Computational Linguistics*, 12:507–524, 2024. 2

[22] Jaylen Jones, Lingbo Mo, Eric Fosler-Lussier, and Huan Sun. A multi-aspect framework for counter narrative evaluation using large language models. *arXiv preprint arXiv:2402.11676*, 2024. 4

[23] 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. *arXiv preprint arXiv:2407.06863*, 2024. 1, 2, 4, 8

[24] Amr Keleg and Walid Magdy. Dlama: A framework for curating culturally diverse facts for probing the knowledge of pretrained language models. *arXiv preprint arXiv:2306.05076*, 2023. 2

[25] 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, 2024. 8, 12

[26] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models?, 2024. 12

[27] Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*, 2023. 2

[28] Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*, 2024. 4

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 12

[30] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 12

[31] Leila Monaghan, Jane E Goodman, and Jennifer Robinson. *A cultural approach to interpersonal communication: Essential readings*. John Wiley & Sons, 2012. 1

[32] Roberto Navigli and Simone Paolo Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, 193:217–250, 2012. 3

[33] Elinor Ochs. *Linguistic resources for socializing humanity.* Cambridge University Press, 1996. 2

[34] Aaron Hurst OpenAI. Gpt-4o system card, 2024. 2

[35] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 12

[36] Talcott Parsons. The system of modern societies, 1971. 1

[37] Tianhao Qi, Hongtao Xie, Pandeng Li, Jiannan Ge, and Yongdong Zhang. Balanced classification: A unified framework for long-tailed object detection. *IEEE Transactions on Multimedia*, 26:3088–3101, 2023. 12

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5, 12

[39] 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. 8

[40] Royi Rassin, Aviv Slobodkin, Shauli Ravfogel, Yanai Elazar, and Yoav Goldberg. Grade: Quantifying sample diversity in text-to-image models. *arXiv preprint arXiv:2410.22592*, 2024. 1

[41] Michael Saxon and William Yang Wang. Disparities in text-to-image model concept possession across languages. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, page 1870, New York, NY, USA, 2023. Association for Computing Machinery. 8

[42] Michael Saxon and William Yang Wang. Multilingual conceptual coverage in text-to-image models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4831–4848, Toronto, Canada, 2023. Association for Computational Linguistics. 1, 8

[43] Michael Saxon, Yiran Luo, Sharon Levy, Chitta Baral, Yezhou Yang, and William Yang Wang. Lost in translation? translation errors and challenges for fair assessment of text-to-image models on multilingual concepts. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 572–582, Mexico City, Mexico, 2024. Association for Computational Linguistics. 8

[44] Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. Kmmlu: Measuring massive multitask language understanding in korean. *arXiv preprint arXiv:2402.11548*, 2024. 2

[45] Wenxiang Sun, Yixing Fan, Jiafeng Guo, Ruqing Zhang, and Xueqi Cheng. Visual named entity linking: A new dataset and a baseline. *arXiv preprint arXiv:2211.04872*, 2022. 3

[46] Abhishek Sureddy, Dishant Padalia, Nandhinee Periyakaruppa, Oindrila Saha, Adina Williams, Adriana Romero-Soriano, Megan Richards, Polina Kirichenko, and Melissa Hall. Decomposed evaluations of geographic disparities in text-to-image models. *arXiv preprint arXiv:2406.11988*, 2024. 8

[47] Mor Ventura, Eyal Ben-David, Anna Korhonen, and Roi Reichart. Navigating cultural chasms: Exploring and unlocking the cultural pov of text-to-image models. *arXiv preprint arXiv:2310.01929*, 2023. 8

[48] Yixin Wan, Di Wu, Haoran Wang, and Kai-Wei Chang. The factuality tax of diversity-intervened text-to-image generation: Benchmark and fact-augmented intervention, 2024. 1

[49] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yolov10: Real-time end-to-end object detection, 2024. 12

[50] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 6

[51] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 12

[52] Andre Ye, Sebastin Santy, Jena D Hwang, Amy X Zhang, and Ranjay Krishna. Computer vision datasets and models exhibit cultural and linguistic diversity in perception. *arXiv preprint arXiv*, 2310:4, 2023. 7

[53] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: evaluating generated text as text generation. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2021. Curran Associates Inc. 4

[54] Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. Pangea: A fully open multilingual multimodal llm for 39 languages. *arXiv preprint arXiv:2410.16153*, 2024. 6

[55] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. 3, 5

[56] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR, 2021. 15

[57] Naitian Zhou, David Bamman, and Isaac L Bleaman. Culture is not trivia: Sociocultural theory for cultural nlp. *arXiv preprint arXiv:2502.12057*, 2025. 2

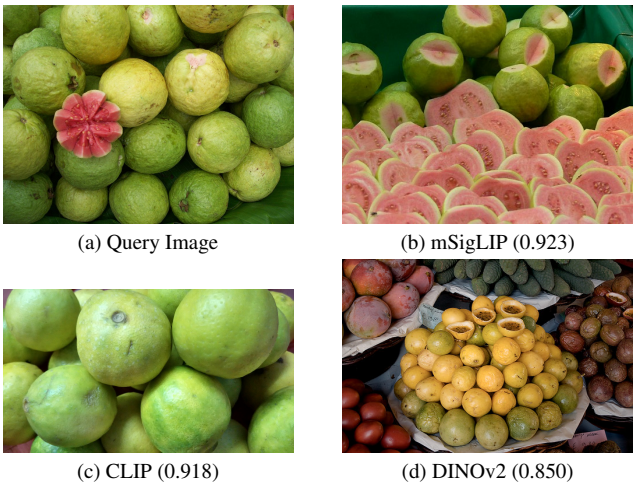
A. Details on Visual Entity Linking

A.1. Off-the-shelf Object Detection

In our approach to Visual Entity Linking (VEL), we initially employed a straightforward methodology by leveraging off-the-shelf object detection models to identify entities within images. We utilized Detectron 2 [51] from Facebook research along with YOLOv10 [49]. The underlying premise was to recognize object names through these detection tools and subsequently map them to corresponding knowledge base (KB) identifiers, akin to traditional textual entity linking. To this end, we experimented with state-of-the-art object detection models capable of recognizing fine-grained object categories in both open and closed-world settings, where KB identifiers served as predefined classification labels. COCO [29] and LVIS [16] classes are 2 of the most commonly used sets of classes. Additionally, we utilized Vision language models to detect entities. Specifically, we used LlaVA-NeXT [30] and Idefics2-8B [26].

However, these methods yielded suboptimal performance, particularly for long-tail entities, a challenge commonly observed in object detection [37]. Given that culturally rare or niche objects often fall within the long-tail distribution, this approach was deemed impractical for robust entity linking.[8] The results are presented on the *"transcreation"* dataset introduced by [25].

A.2. Encoder Models for Building KB Index



(a) Query Image (b) mSigLIP (0.923) (c) CLIP (0.918) (d) DINOv2 (0.850)

Figure 4. Retrieval Comparison Across Encoders

We evaluated several image-text encoder models to compute both image-image and image-text similarity. After qualitatively analyzing the outputs from different variants, we found that mSigLIP performed the best. Its unique loss function improves image-text alignment, making it superior to other encoders such as CLIP [38] and DINOv2 [35] for our task. We conduct a qualitative analysis of retrieval performance using a small subset of images. This subset comprises all fruit images from BabelNet, with each encoder provided the same set of potential matches for retrieval. Figure 4 presents a comparative evaluation of retrieval performance across different encoders, demonstrating mSigLIP's superior semantic image understanding and similarity scores.

A.3. Visual Entity Linking

In this section, we detail further approaches to perform visual entity linking.

Image Query + Text Keys: We compute the similarity between input image embeddings and KB text embeddings to retrieve the top-20 entities, ranking them by similarity score. We build three different text-embedding indices to individually search over for top-matching keys, given the query image. The first comprises of lemma embeddings of each entity, the second comprises of gloss embeddings of each entity,[9] and the third comprises of Wikipedia page embeddings.[10]

Image Query + Image Keys: We experiment with bypassing text-based disambiguation and rank each KB ID based on its occurrence frequency among the top-20 retrieved images, a method we refer to as frequency-based matching.

All of our methods are summarized below:

1. `Lemma (T)`: Image-to-text similarity with KB lemmas.
2. `Gloss (T)`: Image-to-text similarity with KB glosses.
3. `Wikipedia (T)`: Image-to-text similarity with Wikipedia content.
4. `Lemma (V-T)`: Vision-based similarity followed by text matching.
5. `Frequency-Based (V-T)`: Most frequently occurring IDs among retrieved images.

To identify the best-performing method among the above alternatives, we conduct a quantitative evaluation using the Fine-grained Object Classification (FOCI) benchmark [14]. The FOCI benchmark includes nine datasets, with four subsets of ImageNet (IN-Food, IN-Plant, IN-Animal, IN-Artifact). We select this subset from the FOCI benchmark as we are able to map each ImageNet ID to the corresponding BabelNet ID using the structure of ImageNet-21k. This is enabled by ImageNet being built using WordNet cate-

[8] A comparative analysis of the outputs produced by various object detection models and VLMs utilized in our study can be found here: https://hub.zenoml.com/project/954d212c-eca6-4ad6-a461-7cb623641732/Object Captioning. We present the results for 5 different systems; Using Detectron2, yolo10x, LlaVA-NeXT, Idefics2-8B, and an Entity-Attribute-Relation (EAR) pipeline.

[9] https://www.babelnet.org/synset?id=bn:00015267n&orig=dog&lang=EN

[10] https://www.tensorflow.org/datasets/community_catalog/huggingface/wikipedia

Datasets	IN-Food				IN-Plant				IN-Animal				IN-Artifact			
Methods	@1	@5	@10	@20	@1	@5	@10	@20	@1	@5	@10	@20	@1	@5	@10	@20
Idefics-2B	11.58	10.95	10.72	11.91	4.60	4.20	4.68	5.03	8.51	6.86	8.30	9.10	7.48	8.19	8.82	9.43
Lemma (V-T)	**23.65**	**50.65**	**57.93**	**62.07**	16.15	42.14	51.30	**57.34**	22.47	52.45	62.64	**70.11**	**11.61**	**22.02**	**25.51**	**27.88**
Freq-Based (V-T)	22.71	45.79	52.02	57.85	**26.32**	**46.21**	**53.24**	56.24	**31.84**	**56.55**	**63.43**	68.52	10.34	19.99	23.16	25.98
Lemma (T)	4.68	13.69	18.02	22.78	3.20	9.13	12.82	17.10	5.67	15.66	21.69	28.17	1.63	4.15	5.65	7.22
Gloss (T)	3.32	8.13	11.54	15.16	2.39	7.32	10.82	15.17	4.15	12.46	17.66	23.73	1.09	2.85	3.80	4.97
Wikipedia (T)	8.40	20.48	25.93	31.83	4.99	14.52	20.50	27.83	6.84	20.76	28.60	37.20	4.38	10.72	13.84	17.03

Table 3. Retrieval accuracy at different thresholds across the ImageNet datasets for various disambiguation methods. The highest accuracy for each dataset under each setting is highlighted.

gories and hierarchy. Dataset statistics are listed in Table 4. Each category has 10 associated images.

Dataset	No. of Categories
IN-Food	563
IN-Plant	957
IN-Animal	1314
IN-Artifact	2630

Table 4. ImageNet dataset Statistics

FOCI is a rigorous benchmark that reframes image classification as a series of multiple-choice questions (MCQs). Leveraging CLIP-based similarity, it constructs challenging answer choices by selecting visually and semantically similar instances that are frequently confused with one another. We conduct our evaluation in an open-ended setting, where the model is provided only the input image, without predefined answer choices, making the task significantly more challenging.

Our evaluation aims to:

- Benchmark our system against state-of-the-art VLMs for VEL.
- Compare the effectiveness of various retrieval and disambiguation techniques.
- Identify retrieval performance thresholds beyond which VEL performance saturates (e.g., Recall@X).

To enable quantitative evaluation, we map each category in the IN datasets to BabelNet entities and consider a prediction correct if it retrieves the exact BabelNet ID. This facilitates a direct comparison of different disambiguation strategies. Additionally, by evaluating retrieval at varying levels, we can determine the point at which performance saturates and establish the optimal number of retrieved entities per example.

Results on the FOCI benchmark shown in Table 3 in an open setting reveal that image-to-image retrieval combined with image-text disambiguation (`Lemma (V-T)`; `Gloss (V-T)`; `Frequency-based (V-T)`) consistently surpasses direct image-text retrieval methods (`Lemma (T)`; `Gloss (T)`; `Wikipedia (T)`). We outperform all state-of-the-art VLMs presented as baselines in the FOCI work itself, demonstrating the robustness of our framework in generally identifying objects belonging to the long tail and linking them to appropriate KB entities.

A.4. VEL Example

Figure 5. VEL Query Image:
Pysanka (Slavic Decorated Egg)

This section aims to provide a detailed example of the best performing entity linking techniques, illustrating various disambiguation techniques along with their respective variations.

We begin by retrieving the top 20 images from our knowledge base (KB) based on the query image shown in Figure 5. Figure 6 illustrates the 5 highest similarity matches.

Note: While we present step-by-step results using only 5 retrieved images for clarity, our actual framework utilizes 20 images.

Each retrieved image in Figure 6 corresponds to a list of BabelNet IDs as shown in Table 5. We extract and aggregate these IDs, retaining only the unique ones before proceeding to the disambiguation stage. The unique IDs and their corresponding lemmas are presented below:

After aggregation, these IDs are treated uniformly, irrespective of their initial image similarity. Our approach incorporates three primary disambiguation techniques.

`lemma matching`: We compute the cosine similarity between the query image embeddings and the lemma text embeddings for each BabelNet ID in the retrieved set. The

Sim: 0.9232446 Sim: 0.92118037 Sim: 0.92117214 Sim: 0.9197761 Sim: 0.9105648

Figure 6. Top-5 Image Retrieval Results for Query Image.

BabelNet ID	Lemma
bn:00068196n	Romania
bn:00029497n	Easter
bn:00078872n	Ukraine
bn:00538675n	Folklore of Romania
bn:02889635n	Etymology of Ukraine
bn:03096581n	Pysanka
bn:00029503n	Easter egg

Table 5. Unique BabelNet IDs and their lemmas.

IDs are ranked based on their similarity scores. Table 6 presents the top-5 ranked IDs alongside their corresponding lemmas and similarity scores. As observed, the highest-ranked lemma aligns with the gold label of the query image.

BabelNet ID	Lemma	Score
bn:03096581n	Pysanka	0.5243
bn:00029497n	Easter	0.5161
bn:00029503n	Easter egg	0.5157
bn:00538675n	Folklore of Romania	0.5108
bn:00068196n	Romania	0.5037
bn:00078872n	Ukraine	0.4985
bn:02889635n	Etymology of Ukraine	0.4942

Table 6. Ranked BabelNet IDs

BabelNet ID	Lemma	Frequency
bn:03096581n	Pysanka	13
bn:00029503n	Easter egg	10
bn:00029497n	Easter	6
bn:00078872n	Ukraine	2
bn:00068196n	Romania	1
bn:00538675n	Folklore of Romania	1
bn:02889635n	Etymology of Ukraine	1

Table 7. Frequency ranked BabelNet IDs

gloss matching: In this method, we utilize the gloss associated with each BabelNet ID. The gloss is a short definition or description of the entity or concept. This method follows the same approach as lemma matching, where we rank IDs according to gloss-image similarity. In this particular case, this method yields a ranking identical to 'lemma matching'.

frequency matching: This method ranks BabelNet IDs based on their frequency of occurrence within the ID lists associated with the retrieved images. The results of this approach are presented in Table 7, which reports the frequencies computed over a set of 20 retrieved images.

B. Further CAIRE design considerations

B.1. Alternative Cultural relevance scoring techniques

Log-likelihoods: In this approach, we opt to directly use the LM's token log-likelihoods corresponding to each candidate culture label. Given an input prompt consisting of the retrieved textual information about an entity in the image, we prompt the LM with a completion in the form of: "*This text is relevant to [culture label]*". We estimate the relevance of each culture label directly using the LM head log-likelihood of the completion. In the case of VLMs, we also provide the input image as additional conditioning context in calculating the completions' log likelihoods.

We use the notation $\mathcal{L}_m(y, x)$ to denote the negative log likelihood of tokens y from model m conditioned on context X, or

$$\mathcal{L}_m(y, X) = -\log P_m(y \mid X) \tag{1}$$

To compute the log-likelihood of a culture label c, we construct context X containing *retrieved* documents D, the aforementioned "*This text is relevant to*" prompt p, and input image I to compute

$$\text{CAIRE}(I, c) = \mathcal{L}_m(y, (D_i, I, p)) \tag{2}$$

One problem with this approach is that models m predict different base rates $\mathcal{L}_m(c_i, (\emptyset, \emptyset, p))$ for different culture symbols when conditioned on the prompt alone. This introduces a bias where, for example, models will systematically prefer "*relevant to Australia*" over "*relevant to Suriname*" hindering consistent comparison of attribution scores between cultures.

To mitigate this, we apply an affine debiasing method using base rates and hyperparameters λ and T [56]. We adjust the likelihood scores by subtracting a scaled correction term while capping its contribution at a threshold to prevent excessive influence from low-likelihood completions, ensuring a balanced adjustment across all outputs:

$$\hat{\mathcal{L}}(c_i, (D, I, p)) = \mathcal{L}(c_i, (D, I, p)) - \lambda \cdot \max(\mathcal{L}(c_i, (\emptyset, p)), T) \tag{3}$$

In practice, we find that ordinal scoring works better than converting raw log-likelihoods to a relevance rating. 1-5 scoring is also practically feasible to correlate with human judgment. The final CAIRE formulation uses ordinal scoring to grade images across cultures in the label set.

Model	Img	Wiki	Top-20	F1	Δ_{CAIRE}
CAIRE					
Llama-3.2-11B-Vis.-Ins.	✓	✓		47.4	(+6.6)
	✓		✓	47.9	(+7.1)
Qwen2.5-7B-Ins.		✓		**68.9**	(+27.9)
			✓	52.1	(+11.1)
Qwen2.5-VL-7B-Ins.	✓	✓		65.5	(+24.5)
	✓		✓	54.8	(+13.8)
Pangea-7B-hf	✓	✓		42.9	(+22.6)
	✓		✓	29.0	(+8.7)
CAIRE (Log-probabilities)					
Llama-3.2-11B-Vis-Ins.	✓	✓		**61.8**	(+21.0)
Qwen2.5-7B-Ins.		✓		52.5	(+11.5)
Qwen2.5-VL-7B-Ins.	✓	✓		43.6	(+2.6)
Pangea-7B-hf	✓	✓		56.0	(+35.7)

Table 8. F1-scores on the `specific` set. Δ_{CAIRE} represents the improvement each CAIRE implementation achieves over the strongest baseline.

C. Experimental Details

C.1. Culturally Similar Matches

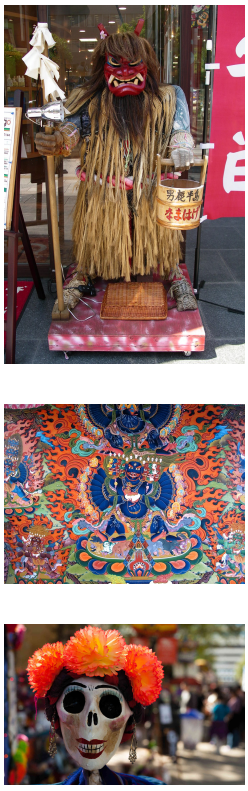
There are instances where the exact entity is not matched. A few of these examples are presented in Figure 7.

Case 1: In this case, the retrieved entity represents a specific aspect of a broader cultural concept. For example, as illustrated in the first case, the retrieved entity is an Oni, a type of Japanese troll, whereas the gold-standard label, Namahage, refers to a broader category of Japanese demon-like creatures. Despite this distinction, the retrieved entity remains culturally relevant and semantically aligned with the gold label.

Case 2: The second case exemplifies instances where the retrieved entity directly represents the associated culture. The image depicts a Thangka painting, a form of Tibetan art. Our retrieval system identifies the broader concept of Tibetan art, which remains an accurate cultural attribution given the overarching goal of cultural relevance assessment.

Case 3: The final case mirrors the first, where the retrieved entity corresponds directly to the deity depicted in the image. This deity is highly relevant and closely associated with the gold-standard label, which represents the broader cultural event, the Day of the Dead.

In such cases, while the retrieval is technically classified as incorrect, the overall cultural scoring remains reliable. This is because the retrieved entities are often culturally analogous to the target entities, and their corresponding Wikipedia pages contain thematically similar content. Consequently, the models can still effectively extract the relevant cultural information.



Gold: Namahage
CAIRE: Oni

Gold: Thangka Painting
CAIRE: Tibetan Art

Gold: Day of the Dead
CAIRE: Mictēcacihuātl

Figure 7. Specific set examples

C.2. GlobalRG Concept List

Table 9 presents the 20 human universals selected by [3].

C.3. Cultural Relevance Scoring Criteria

Table 12 presents descriptions for each score in our 1-5 scoring formulation.

C.4. Example Prompt for 1-5 Scoring

A prompt template used in the 1-5 Scoring method for VLMs is provided in Figure 8.

Prompt Template:
We want to assess how relevant an image is to a given culture.
We have identified this concept to be closely associated with the image: {entity}.
Here is some detailed information about this concept from Wikipedia: {wiki}.

Using the above context, assign a score from 1 to 5 based on how culturally relevant the image is to {target}:
Think step by step, specifically considering cultural symbols, styles, traditions, or any features that align with the culture of {target}.

The final score should be a number between 1 to 5, where the meaning of each score is defined as follows:

- 1 – Not Relevant: The content does not connect with or reflect the target culture at all.
- 2 – Minimally Relevant: The content shows slight or superficial connections to the culture but lacks depth. May include vague references or isolated cultural elements that feel out of place or underdeveloped.
- 3 – Somewhat Relevant: The content contains identifiable cultural references, but they may feel generic, inconsistent, or limited in scope. The connection to the culture is present but could be stronger or more meaningful.
- 4 – Relevant: The content reflects a reasonable understanding of the culture, including accurate and appropriate references. It integrates cultural aspects well, though there may still be areas where more depth could be added.
- 5 – Highly Relevant: The content is deeply connected to the target culture, showing an immersive, accurate, and respectful understanding. Cultural references feel natural, meaningful, and central to the content.

The output should be a single number ONLY.

Example Conversation Format:

System: *You are an expert in evaluating the cultural relevance of images.*

User: [Image] [Text: Prompt Template]
Model:
Final Score: [1-5]

Figure 8. Prompt Template for CAIRE methods

Breakfast	Clothing	Dance	Drinks
Dessert	Dinner	Farming	Festival
Eating Habits	Funeral	Greetings	Head Coverings
Instrument	Lunch	Marriage	Music
Religion	Ritual	Sports	Transport

Table 9. Cultural Categories Table

C.5. Upper Bound Analysis

We obtain the results in Table 10 by replacing the retrieved Wikipedia pages with the ground-truth pages corresponding to the entities in our `specific set`. By comparing the F1-scores of our retrieved pages to those of the gold-standard pages, we quantify the retrieval-induced error in our pipeline. Specifically, we compute the ratio of the F1-scores from our VEL system to those obtained using the gold-standard pages. As shown in Table 11, our method consistently achieves scores comparable to the gold standard, highlighting the effectiveness of our retrieval system

Model	Image	Wikipedia	F1 Score
CAIRE (1-5 Scores)			
Llama-3.2-11B-Vision-Instruct	✓	✓	48.8
Qwen2.5-7B-Instruct		✓	77.7
Qwen2.5-VL-7B-Instruct	✓	✓	**77.7**
Pangea-7B-hf	✓	✓	46.6
CAIRE (Log-probabilities)			
Llama-3.2-11B-Vision	✓	✓	**65.7**
Qwen2.5-7B		✓	58.6
Qwen2.5-VL-7B-Instruct	✓	✓	45.1
Pangea-7B-hf	✓	✓	62.6

Table 10. F1-scores with Gold context

Model	Image	Ratio
CAIRE (1-5 Scores)		
Llama-3.2-11B-Vision-Instruct	✓	97.1
Qwen2.5-7B-Instruct		88.7
Qwen2.5-VL-7B-Instruct	✓	84.3
Pangea-7B-hf	✓	92.1
CAIRE (Log-probabilities)		
Llama-3.2-11B-Vision	✓	94.1
Qwen2.5-7B		89.6
Qwen2.5-VL-7B-Instruct	✓	96.7
Pangea-7B-hf	✓	89.5

Table 11. Ratio between CAIRE and Gold context

Score	Relevance Level	Description
1	Not Relevant	The content does not connect with or reflect the target culture at all.
2	Minimally Relevant	The content shows slight or superficial connections to the culture but lacks depth. May include vague references or isolated cultural elements that feel out of place or underdeveloped.
3	Somewhat Relevant	The content contains identifiable cultural references, but they may feel generic, inconsistent, or limited in scope. The connection to the culture is present but could be stronger or more meaningful.
4	Relevant	The content reflects a reasonable understanding of the culture, including accurate and appropriate references. It integrates cultural aspects well, though there may still be areas where more depth could be added.
5	Highly Relevant	The content is deeply connected to the target culture, showing an immersive, accurate, and respectful understanding. Cultural references feel natural, meaningful, and central to the content.

Table 12. Cultural Relevance Scoring Criteria

in identifying culturally relevant entities.

C.6. Specific Set Label Set

Table 13 shows the full set of candidate culture labels for the images in the `Specific Set`:

Category	Labels
Geographical Entities	Countries of the world Countries of Africa States of Mexico States of the US States of India Major cities of India Cities of Indonesia Cities of Nepal
Ethnic & Cultural Groups	Ethnicities of Africa Ethnicities of India Ethnicities of Australia Ethnicities of the US Ethnicities of Indonesia
Festivals & Traditions	Festivals of India
Philosophy & Religion	Japanese philosophy World religions
Historical Civilizations	Bronze Age civilizations

Table 13. Cultural labels for the Specific Set

D. Human Annotation Analysis

D.1. Concept-Wise Correlation

We compute the Pearson correlation scores for each of the 20 GlobalRG categories to analyze trends across different concepts. The image-wise correlations are averaged for each concept, and the final results are presented in Table 15. This analysis helps us highlight which concepts exhibit strong alignment and which ones show weaker correlations. This can provide insights into potential biases and inconsistencies in concept representation. Notably, the correlation scores for the language model without image input are consistently the lowest across all categories for both Natural and Generated images. Additionally, we observe that the correlation values for Generated images are slightly lower than those for Natural images. The correlation scores for the concept of *"Dinner"* are notably low in the case of Qwen2.5-7B. The concept of *"Greetings"* shows the highest correlation values, likely because greetings often have clear, culturally distinctive forms that are easier to visually identify compared to other concepts. Food-related concepts such as *"Breakfast"*, *"Eating Habits"*, and *"Lunch"* tend to exhibit lower correlation scores, particularly for Generated images, likely because the models may not always produce specific, culturally distinctive dishes.

D.2. Inter-Annotator Correlation

We compute inter-annotator correlation for annotators from 10 countries by averaging pairwise correlations within each country. Using 1-5 ratings for 200 images, we evaluate Spearman's and Pearson's correlations along with additional metrics. These results are shown in Table 14.

The additional metrics we report are Krippendorff's, and Weighted-Krippendorff's Alpha.

Metric	Brazil		China		Egypt		Germany		India		Indonesia		Mexico		Nigeria		Russia		USA		Avg.	
	(N)	(G)	(N)	(G)	(N)	(G)	(N)	(G)	(N)	(G)	(N)	(G)	(N)	(G)	(N)	(G)	(N)	(G)	(N)	(G)	(N)	(G)
Kripp Alpha	34	30	43	22	35	27	28	53	31	32	43	17	18	29	19	18	49	37	4	24	30	28
Weighted Alpha	35	30	52	23	39	25	28	52	34	31	46	17	18	25	22	17	54	34	5	25	33	27
Avg Spearman	44	43	50	29	45	33	35	62	41	43	54	35	28	38	23	26	55	49	12	34	39	39
Avg Pearson	46	42	57	29	46	32	38	61	45	43	57	34	32	36	25	27	60	47	12	34	42	38

Table 14. Inter-annotator correlation metrics across different countries (scaled to 100). **(N)** denotes natural, and **(G)** denotes generated.

Krippendorff's Alpha (α):

$$\alpha = 1 - \frac{D_o}{D_e} \tag{4}$$

where D_o is the observed disagreement and D_e is the expected disagreement by chance.

Weighted Alpha:

$$\alpha_w = 1 - \frac{D_o^w}{D_e^w} \tag{5}$$

where D_o^w and D_e^w are the weighted observed and expected disagreements, respectively, with the weights reflecting the degree of disagreement between values.

The inter-annotator agreement is notably low for certain regions. In particular, the agreement for Natural images from the USA is quite low, likely due to the country's high cultural diversity, making it harder to converge on a single cultural interpretation. Conversely, the agreement for Generated images in regions like Germany and Russia is relatively high, possibly because the generated content tends to align with more homogeneous or stereotypical cultural representations.

D.3. Averaged Correlation

In this section, we present two types of averaged correlations to analyze global trends and country-level agreement, as shown in Table 16. The values in Table 2 were computed as image-wise correlations and then averaged, reflecting local consistency by measuring how well different methods align with human annotations at the image level.
To assess broader agreement, we compute two additional correlations:

1. **Concept-Averaged Correlation:**
   - We first compute the average rating (on a scale of 1–5) from CAIRE across all images for each concept.
   - Then, we calculate the Pearson correlation between these averaged scores and the corresponding human annotator scores.
   - Finally, we report the overall average correlation across all concepts.

Concept	Llama-3.2-11B		Qwen2.5-7B		Qwen2.5-VL-7B		Pangea-7B	
	(N)	(G)	(N)	(G)	(N)	(G)	(N)	(G)
Breakfast	68	45	26	28	60	51	49	44
Clothing	77	67	63	53	76	66	75	66
Dance	69	69	65	49	70	68	76	68
Dessert	63	61	59	57	68	71	65	58
Dinner	71	33	4	6	69	34	62	56
Drinks	35	31	52	33	61	49	46	38
Eating Habits	68	38	56	46	73	42	69	51
Farming	46	47	42	27	52	55	62	62
Festival	65	58	60	74	71	76	71	70
Funeral	47	12	35	16	50	34	51	31
Greetings	80	72	69	69	80	74	63	67
Head Coverings	31	45	41	41	51	57	40	54
Instrument	59	41	45	40	47	62	57	37
Lunch	73	40	54	15	61	38	64	28
Marriage	50	75	49	51	49	70	55	49
Music	69	53	55	61	70	75	66	47
Religion	65	45	62	51	79	59	71	47
Ritual	62	41	64	45	70	40	68	50
Sports	74	41	55	27	74	49	49	52
Transport	70	39	54	30	70	48	60	48

Table 15. CAIRE : Pearson's correlation with human judgment across 20 concepts (Scaled to 100). N = Natural data, G = Generated data.

Model	Concept-Avg. Corr. (N)	Concept-Avg. Corr. (G)	Global-Avg. Corr. (N)	Global-Avg. Corr. (G)
Vision-Language Encoders (*Baseline-1*)				
CLIP	1.2	25.1	1.7	33.7
mSigLIP	32.8	-6.4	57.7	-5.9
Vision-Language Models (*Baseline-2*)				
Llama-3.2-11B-Vis.-Ins.	11.8	16.5	16.3	9.2
Qwen2.5-VL-7B-Ins.	-1.1	16.9	-11.8	30.1
Pangea-7B-hf	45.7	39.3	61.2	60.4
CAIRE				
Llama-3.2-11B-Vis.-Ins.	39.4	35.3	56.7	61.7
Qwen2.5-VL-7B-Ins.	58.0	49.7	**85.2**	**76.7**
Pangea-7B-hf	**66.9**	**54.4**	77.7	70.8

Table 16. Averaged correlation values (Scaled to 100) between models and human annotation. The highest correlations are highlighted in bold.

2. **Global-Averaged Correlation:**
   - Instead of averaging per concept, we first compute the mean rating across all 200 images.

- We then compute the correlation between this globally averaged score and human annotator scores.

This approach helps reduce noise in image-wise correlations and captures broader patterns of agreement between our model and human annotations. In particular, averaging scores before computing correlations provides a more stable measure of overall alignment, mitigating the impact of per-image fluctuations.

CAIRE outperform both baseline vision-language encoders and models, with Pangea-7B-hf achieving the highest correlations across all metrics. Qwen2.5-VL-7B-Ins. shows significant improvements under CAIRE, highlighting its enhanced alignment. We also note larger variation between the baselines and our method as compared to the image-wise average correlations.

D.4. Alternate CAIRE-Human correlation

Country	Mean Corr		Max Corr		Min Corr		Median Corr		Std Dev		With Average	
	(N)	(G)	(N)	(G)	(N)	(G)	(N)	(G)	(N)	(G)	(N)	(G)
Brazil	43	37	66	52	11	16	45	37	16	9	59	54
China	47	31	57	47	29	17	49	27	8	12	60	51
Egypt	37	23	51	56	16	-9	39	25	8	21	53	35
Germany	47	45	68	61	8	22	56	48	22	11	71	56
India	45	37	70	55	5	15	43	41	17	11	63	54
Indonesia	45	15	66	24	27	2	46	17	10	7	57	24
Mexico	36	28	60	45	-10	0	42	28	21	12	58	43
Nigeria	23	31	40	47	-10	4	33	32	16	11	43	56
Russia	55	40	63	55	46	30	56	37	6	10	68	52
USA	23	44	53	65	-39	6	36	51	28	19	58	72

Table 17. Comparison of correlation statistics by country for (N) and (G) (scaled by 100, rounded).

Following the approach in subsection D.2, we compute per-country correlations between human ratings and our best-performing method: **CAIRE w/ Qwen2.5-VL-7B-Ins**. The procedure is as follows:

For each country, we collect the 1-5 image-wise scores from all annotators belonging to that country, resulting in 200 ratings per annotator. We then extract the corresponding 1-5 scores from our model, specifically for that country's outputs.

Next, we compute the Pearson correlation between the model's scores and the human ratings. We report the following statistics per country: the mean pairwise correlation, maximum, minimum, standard deviation, median, and the correlation between the model's scores and the averaged human ratings (i.e., scores averaged across all annotators for each image).

As shown in Table 17, the correlation between the model and the averaged human ratings is consistently higher than the mean and median of the pairwise correlations. This suggests that averaging human scores before computing correlations effectively reduces the impact of outliers, resulting in a stronger alignment with the model's predictions. This justifies our choice of evaluation in Table 2.