From Local Concepts to Universals: Evaluating the Multicultural Understanding of Vision-Language Models

Mehar Bhatia   Sahithya Ravi[*] Aditya Chinchure[*] Eunjeong Hwang   Vered Shwartz
University of British Columbia
Vector Institute for AI
{meharb23, vshwartz}@cs.ubc.ca

Abstract

Despite recent advancements in vision-language models, their performance remains suboptimal on images from non-western cultures, due to underrepresentation in training datasets. Various benchmarks have been proposed to test models' cultural inclusivity, but they have limited coverage of cultures and do not adequately assess cultural diversity across universal as well as culture-specific local concepts. To address these limitations, we introduce the GLOBALRG benchmark, comprising two challenging tasks: *retrieval across universals* and *cultural visual grounding*. The former task entails retrieving culturally-diverse images for universal concepts from 50 countries, while the latter aims at grounding culture-specific concepts within images from 15 countries. Our evaluation across a wide range of models reveals that the performance varies significantly across cultures – underscoring the necessity for enhancing multicultural understanding in vision-language models. Our data and code can be found at https://globalrg.github.io/

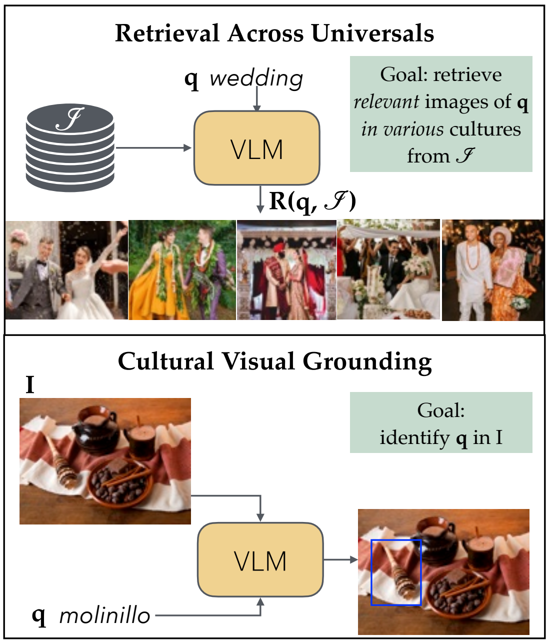


Figure 1: An example instance from each task in GLOBALRG: i) *Retrieval Across Universals* measures the ability of VLMs to retrieve culturally diverse images for a query q. ii) *Cultural Visual Grounding* aims to evaluate the ability of VLMs to identify a cultural concept q.

1 Introduction

Vision-Language Models (VLMs) have shown emergent capabilities through large-scale training that have made them gain popularity in recent years. VLMs show promising results across various vision and language tasks, from image captioning to visual question answering and cross-modal retrieval and grounding. A key component contributing to their strong performance across the board is the scale of their pre-training datasets. However, these large-scale datasets tend to predominantly contain images from Western cultures (Shankar et al., 2017; Ananthram et al., 2024). The underrepresentation of certain cultures in the data translates into performance disparities across cultures. (De Vries et al., 2019; Gustafson et al., 2023).

Several benchmarks and datasets have been proposed to test the cultural inclusivity of VLMs. These include testing the models' performance on questions pertaining to images from certain cultures (Liu et al., 2021a; Yin et al., 2021), on their ability to adapt images from one culture to another (Khanuja et al., 2024), or on stereotypical depiction of various cultures (Jha et al., 2024). Nonetheless, existing benchmarks address a limited set of cultures (5-7), leaving a substantial representational gap. Moreover, current benchmarks leave out a crucial aspect: assessing the cultural diversity in the representation of universal concepts.

To address this gap, we present the GLOBALRG benchmark. GLOBALRG consists of two tasks which are based on popular vision and language tasks: image-text retrieval, which is used, for exam-

[*] Equal Contribution

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 6763–6782
November 12-16, 2024 ©2024 Association for Computational Linguistics

ple, in search engines; and visual grounding, which can test models' vision-language alignment and is crucial for instruction-following applications such as in robotics. Figure 1 presents the multicultural versions of these two tasks.

The first task, **retrieval across universals**, covers images from 50 countries across 10 regions. It assesses the ability of VLMs to retrieve culturally-diverse images pertaining to textual prompts of universal concepts such as "breakfast" and "wedding". In addition to the standard precision@k metric, which verifies that the retrieved images correctly depict the target concept, we also propose a new metric, diversity@k, that measures the cultural-diversity among the retrieved images, allowing us to identify models' bias towards specific countries or regions.

In the second task, **cultural visual grounding**, we cover 15 countries across 8 regions and evaluate models' ability to ground culture-specific concepts (e.g., "molinillo", Mexican whisk) within an image.

Extensive evaluation on 7 models for the retrieval task and 5 models for the grounding task reveals discrepancies across cultures, reassessing findings by prior work (e.g., Liu et al., 2021a; Yin et al., 2021). We further analyze whether VLMs exhibit biases towards certain cultures. In the grounding task, the performance on North America and Europe is substantially higher than on East Asia and South East Asia. This preference is inconsistent across universals in the retrieval task, e.g., a model may retrieve European images of funerals but African images of farming. A closer look reveals that when models retrieve seemingly diverse images, they often share Western elements, such as eggs for breakfast, white dresses at weddings.

GLOBALRG highlights the lack of cultural awareness in VLMs. By identifying and addressing these gaps, we can work towards developing models that perform equally well on inputs pertaining to concepts and images from diverse cultures.

2 Related Work

The Geo-Diversity Problem. Existing large-scale vision and language datasets are imbalanced in their representation of different regions, over-representing the West (Shankar et al., 2017). As a result, models trained on these datasets may exhibit discrepancies in performance when introduced with inputs concerning various demographic and geographic factors (e.g. Gustafson et al., 2023; De Vries et al., 2019). For instance, image generation models—when asked to generate images of universal concepts such as "house", tend to depict the concept as it appears in the US or India, cultures that are more prominently featured in the training data (Basu et al., 2023).

To serve users from diverse cultures fairly, it is imperative to collect large-scale datasets from diverse data sources (Kim et al., 2021; Goyal et al., 2022). Two recent geo-diverse image datasets that are popular for training geo-diverse VLMs, Dollar Street (Rojas et al., 2022) and GeoDE (Ramaswamy et al., 2024), focus on common household items, lacking coverage of more abstract and culture-specific concepts. Finally, to make cross-cultural data collection more feasible, researchers proposed to apply domain adaptation (Kalluri et al., 2023) and active learning (Ignat et al., 2024) based on visual similarity.

Geo-Diverse Benchmarks. With the understanding that language has a social function, there has been growing interest in the NLP community in making models more culturally inclusive (e.g., Hershcovich et al., 2022; Nguyen et al., 2023; Bhatia and Shwartz, 2023). Several benchmarks have been developed to test language models' cultural awareness with respect to values and social norms (Durmus et al., 2023), culinary norms (Palta and Rudinger, 2023), figurative language (Kabra et al., 2023), and more.

In the multimodal domain, benchmarks have been developed to test VLMs on visual question answering and reasoning (Liu et al., 2021a; Yin et al., 2021; Zhou et al., 2022; Nayak et al., 2024), image-text retrieval and visual grounding (Zhou et al., 2022), image captioning (Ye et al., 2023; Burda-Lassen et al., 2024; Karamolegkou et al., 2024), and cultural adaptation (Khanuja et al., 2024; Cao et al., 2024). Despite these efforts, current benchmarks typically cover an incredibly small number of cultures (5-7). To bridge this gap, we introduce a benchmark with two tasks covering 50 and 15 cultures respectively. Moreover, our benchmark tests models both on their familiarity with *culture-specific* concepts and on the diversity of their representation of *universal concepts*.

3 Task 1: Retrieval across Universals

Image-text retrieval is a fundamental task for evaluating VLMs, where the objective is to retrieve relevant images based on textual queries. Existing

Region	Countries
East Asia	China, South Korea, Japan
South East Asia	Vietnam, Thailand, Philippines, Indonesia, Singapore
South Asia	India, Pakistan, Sri Lanka
Middle East Asia	Saudi Arabia, Iran, Turkey, Lebanon, Egypt
Europe	Italy, Greece, France, Germany, Netherlands, Portugal, Spain, United Kingdom, Poland, Sweden, Hungary, Bulgaria, Russia
Africa	Tanzania, Kenya, Uganda, Ghana, Nigeria, Ethiopia, South Africa, Morocco, Somalia, Tunisia
Latin America	Brazil, Peru, Chile, Argentina, Mexico
Caribbean	Jamaica
Oceania	Australia, New Zealand, Fiji
North America	USA, Canada

Table 1: List of cultures covered in the retrieval task.

breakfast	clothing	dance
dessert	dinner	drinks
eating habits	farming	festival
funeral	greetings	head coverings
instrument	lunch	marriage
music	religion	ritual
sports	transport	

Table 2: Human universals used as textual queries in our retrieval dataset.

retrieval benchmarks such as COCO (Lin et al., 2014), Flicker30K (Plummer et al., 2015), ImageCoDe (Krojer et al., 2022), and CIRR (Liu et al., 2021b) contain images predominantly from North America and Europe. To develop globally effective retrieval systems, it is crucial to evaluate models on culturally heterogeneous datasets. In this work, we present a dataset containing images from 50 cultures (Table 1). We introduce the novel task of **Retrieval across Universals**, aimed at retrieving culturally-diverse images for universal concepts such as "wedding". We describe the dataset collection in Sec 3.1.

Image-text retrieval is typically evaluated using precision. Beyond measuring the correctness of the retrieved images, this metric overlooks a significant aspect of retrieval systems: *cultural diversity*. We thus propose an additional evaluation metric to measure the cultural diversity of the retrieved images (Sec 3.2). We evaluate an extensive number of VLMs on the retrieval task (Sec 3.3) and report the results in Sec 3.4.

3.1 Dataset Collection

Textual Queries. The queries in our dataset are human universals—concepts common across cultures worldwide, such as "clothing" and "dance". Table 2 presents the list of 20 human universals used as textual queries in our dataset. The list was adapted from an extensive list of 369 human universals by Brown (2004) and Pinker (2004). We manually selected human universals that can be depicted in images. For example, universals like "clothing" are associated with tangible objects, and "dance" is a ritual that can be visually depicted. In both cases, these universal concepts are expected to be visually represented differently across diverse cultures.[1]

Images. To obtain culturally diverse images corresponding to the textual queries, we first used CANDLE (Nguyen et al., 2023), a comprehensive corpus of cultural knowledge, to extract 3 sentences corresponding to each universal concept and each culture. For example, for "wedding" and "India", CANDLE contains the sentence "*The mehendi ceremony holds significance in Indian tradition*". These sentences provide context and cultural specificity for each universal. We use these sentences to scrape images from Google Images. To ensure the quality of the images, one of the authors manually verified each image in the dataset, filtering out low-resolution images, images with text, and images depicting multiple scenes (i.e., grid images). The final dataset includes a total of 3,000 visually-diverse images (50 cultures $\times$ 20 universals $\times$ 3 images).

3.2 Task Definition and Evaluation Setup

We introduce the novel task of **Retrieval across Universals**, aimed at retrieving culturally diverse images for a given universal concept. Formally, let $\mathcal{Q} = \{q_1, q_2, \ldots, q_n\}$ be a set of textual queries representing universal concepts, and $\mathcal{I} = \{I_1, I_2, \ldots, I_m\}$ the set of images from different cultures. Given a query $q \in \mathcal{Q}$, the goal is to retrieve a ranked list of images $\mathcal{R}(q, \mathcal{I}) = \{I_{r_1}, I_{r_2}, \ldots, I_{r_k}\} \subset \mathcal{I}$ that maximizes both relevance and cultural diversity.

- **Relevance**: $\mathrm{Rel}(q, I)$ refers to how well the image I matches the query q.
- **Diversity**: $\mathrm{Div}(\mathcal{R}(q, \mathcal{I}))$ measures the cultural diversity of the retrieved images.

Specifically, relevance is captured by the standard precision@k, the ratio of the top k retrieved images that correctly answer the query. For diversity, we propose the diversity@k metric, which uses entropy to measure the cultural diversity among the top k retrieved images:

$$\textit{diversity}\ @k = \frac{1}{\log\left(\frac{1}{m}\right)} \sum_{i=1}^{m} p_i \log(p_i) \quad (1)$$

[1]The complete list of human universals can be found here: https://condor.depaul.edu/~mfiddler/hyphen/humunivers.htm

Model	Training Data	Data Size	Relevance		Diversity (Country)		Diversity (Region)	
			prec@5	prec@10	div@5	div@10	div@5	div@10
Dual-Encoder:								
CLIP (Radford et al., 2021)	web-scraped	400M	72.5	70.0	93.96	94.16	66.71	64.64
OpenCLIP (Cherti et al., 2023)	LAION-2B	2B	69.5	75.0	95.69	95.14	73.39	**66.93**
Encoder-Decoder:								
CoCA (Yu et al., 2022)	JFT-3B	3B	**81.0**	**79.5**	**98.27**	**95.37**	68.18	64.88
Dual Encoder + Multimodal Fusion Encoder:								
TCL (Yang et al., 2022)	CC-3M, SBU, COCO, VG	4M	76.0	74.5	92.78	91.22	74.04	66.54
ALBEF (Li et al., 2021)	CC-12M, SBU, COCO, VG	14M	68.0	70.0	92.24	91.11	65.75	64.63
BLIP2 (Li et al., 2023)	CC-3/12M, SBU, COCO, VG, LAION-115M	129M	74.0	74.5	**98.27**	92.96	**74.25**	63.26
FLAVA (Singh et al., 2022)	CC-3/12M, SBU, COCO, VG WIT, Red Caps, YFCC	70M	60.0	62.0	96.54	94.95	72.32	66.84

Table 3: Average performance of various VLMs on the the retrieval across universals task, in terms of **Relevance** and **Diversity**.

where p_i is the proportion of images from the i-th culture in the top k retrieved images $\mathcal{R}(q)$, and m is the total number of cultures in the top k. A high normalized entropy value (~ 1) indicates high diversity, meaning the retrieved images are well-distributed across different cultures. Conversely, a low entropy value (~ 0) indicates low diversity, suggesting that the retrieved images are biased towards specific cultures. We report diversity with respect to both the country and the region.

Our balanced focus on relevance and diversity ensures that models are evaluated not only on their ability to understand and represent concepts accurately but also on their capacity to do so across cultures.

3.3 Models

We evaluate the performance of several state-of-the-art VLMs on the retrieval task. The models are categorized based on their architectural design and training methodologies in Table 3. We cover a diverse set of models, including dual encoder and encoder-decoder, as well as dual encoders with multimodal fusion encoder. These models facilitate cross-modal alignment via a multitude of pre-training objectives, including contrastive loss on uni-modal encoders, image-text matching, masked language modelling, and more.[2]

3.4 Results and Analysis

RQ$_1$: Are VLMs able to retrieve relevant and culturally diverse images for universal concept words? Table 3 presents the relevance and diversity scores for each model (see Appendix A.1.1 for a complete breakdown by universal). With respect to relevance, models achieve moderate to high precision scores, with CoCA leading by 5 points.

We note that country-level diversity scores are high for all models, indicating that VLMs can retrieve images from a variety of geographical contexts. Among them, CoCA performs exceptionally well, likely attributed to its extensive training on 3 billion images from Google's proprietary JFT dataset (Zhai et al., 2022).

Similarly, in dual-encoder models, OpenCLIP demonstrates superior cultural diversity, benefiting from its large training dataset of 2 billion images. CLIP, which uses the same dual-encoder architecture and contrastive loss objectives as OpenCLIP but is trained on a dataset five times smaller, exhibits lower performance across all metrics. Naturally, pre-training on a larger-scale dataset increases the chances that the model was exposed to more culturally diverse images. In contrast, regional diversity scores are notably lower across the board. At the same time, for country diversity@5, BLIP-2 stands out as having the highest cultural diversity, leveraging frozen pre-trained encoders (ViT-G (Fang et al., 2023) as the vision encoder and instruction-tuned FlanT5 (Chung et al., 2024) as the language model) and a QFormer architecture.

A particularly surprising finding is the robust performance of TCL with respect to both relevance and diversity – despite being trained on a the smallest dataset among all models (4M images). TCL incorporates a unique uni-modal objective to make the model invariant to data modifications, which likely benefits the cross-modal alignment and joint multi-modal embedding learning. This may suggest that well-designed training objectives can sometimes compensate for smaller datasets, highlighting the significance of pre-training objectives alongside data scale.

[2]We could not evaluate advanced closed-source models like GPT-4v or Gemini on our retrieval task since these models do not support searching through our large collection of images.

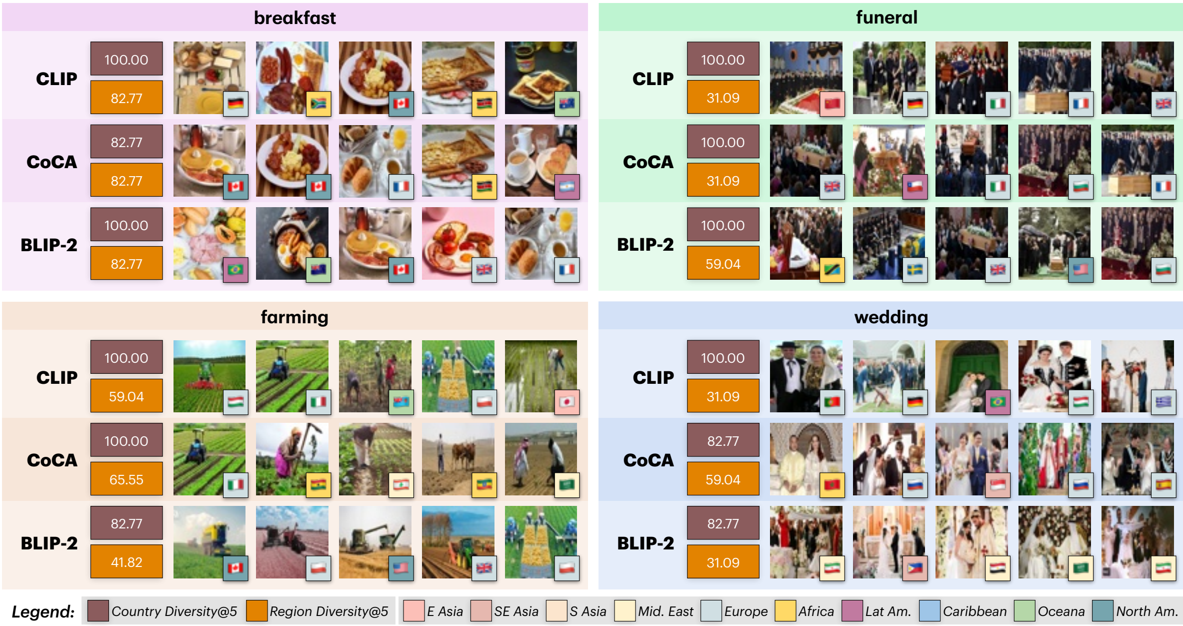


Figure 2: Top 5 images retrieved for a sample of the universals by models CLIP, CoCA and BLIP-2. Each image is annotated with a flag representing the country, and the background colour of the flag represents the region.

RQ$_2$: Do VLMs exhibit biases towards images from specific cultures? From the full results in Appendix A.1.2 and A.1.3 we can observe that there are no countries or regions that are consistently retrieved by models. A closer look reveals that the bias towards specific countries or regions is universal-specific. To demonstrate this point, we plot the top 5 retrieved images for 4 universal concepts, “breakfast”, “funeral”, “farming”, and “wedding”, in Figure 2.

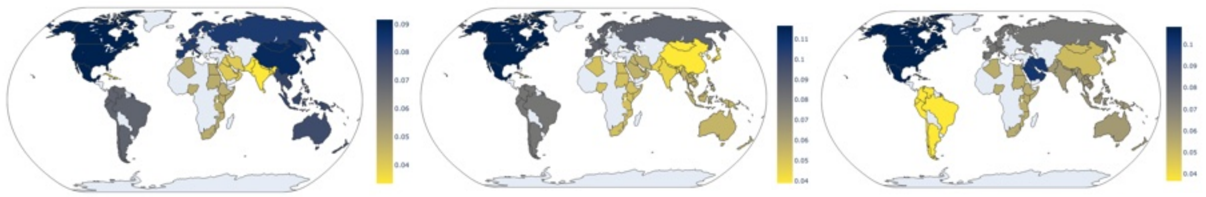
Despite exhibiting high country-level diversity and moderate region-level diversity, Figure 2 shows that the retrieved images for breakfast predominantly contain Western breakfast items such as eggs, sausages and toast. Similarly, the images for “funeral” mostly feature black dresses, and are overwhelmingly from Europe.

With respect to “farming”, CLIP and BLIP-2 mostly retrieve images from Western countries depicting technologically advanced farming tools and large green fields, whereas CoCA retrieves images from Africa and the Middle East of people working in the fields. Finally, the images for “wedding” are diverse across models, although CLIP focuses on Western images whereas BLIP-2 prefers the Middle East (yet still retrieving images of white dresses).

Despite being trained on large datasets, models like CLIP still exhibit notable biases towards Western cultures. While CoCA generally exhibits better diversity compared to CLIP and BLIP-2, all models display certain biases and preferences for Western-style elements, such as black dresses at funerals, white dresses at weddings, and eggs for breakfast.

We also present a visualization in Figure 3, illustrating the geographical biases across three VLMs: CLIP, CoCA, and BLIP-2. The heatmap shows the ratio of how frequently a region appears in the top 10 retrieved images, normalized by how often that region is represented in the retrieval dataset. This is calculated across all 20 universal concepts. Consistently, across all three models, there is a clear overrepresentation of images from North America and Europe in the top retrievals. In contrast, images from regions like Africa, Southeast Asia, and East Asia are retrieved far less frequently, despite being candidates for retrieval. This disparity highlights significant biases in the models’ retrieval mechanisms, emphasizing the need for better training strategies to ensure fairer and more balanced global representation in VLMs.

RQ$_3$: What are the challenges faced by VLMs in achieving high cultural diversity? A low diversity score may be attributed to various factors. First, the scarcity of images from non-Western cultures means that pre-training datasets are predominantly Western-centred (Shankar et al., 2017). Second, many large-scale pre-training datasets are predominantly sourced from Western-centric platforms, leading to the overrepresentation of Western cul-



(A) CLIP (B) CoCA (C) BLIP-2

Figure 3: Performance disparity across regions on the retrieval task, depicted across three VLMs, (a) CLIP, (b) CoCA, (c) BLIP-2. The plot shows the ratio of how frequently a region appears in the top 10 retrieved images, normalized by how often that region is represented in the retrieval dataset.

tures. Finally, typical pre-training objectives are designed to maximize general image-text alignment and do not specifically target cultural diversity, leading models to associate for example breakfast with eggs and weddings with white dresses.

4 Task 2: Cultural Visual Grounding

Visual grounding is essential for human-AI interactions, enabling users to reference regions using spatial cues and models to respond with precise visual answers, such as bounding boxes. Existing grounding datasets such as RefCOCO and its variants (Kazemzadeh et al., 2014; Yu et al., 2016), Flickr Entities (Plummer et al., 2015), Visual Genome (Krishna et al., 2017), and GRIT (Gupta et al., 2022) tend to focus on generic concepts and their images lack cultural contexts.

To address this limitation, we propose the task of **Cultural Visual Grounding**, to evaluate the ability of VLMs to identify culture-specific concepts. We describe our dataset collection (Sec 4.1), the task and evaluation metric (Sec 4.2). We evaluate various models on our task (Sec 4.3), and report the performance in Sec 4.4.

4.1 Dataset Collection

Cultural Keywords. In this task, we focus on 15 countries across 8 regions, detailed in Table 4.[3] We extract 50 cultural keywords for each culture from CANDLE, covering topics such as food, rituals, clothing, etc. The list of keywords is detailed in Appendix A.2. We also provide the distribution of concepts across five facets, i.e., clothing, food, drinks, rituals and traditions, in Appendix A.3.

Images. To obtain images corresponding to the keywords, we recruit annotators from the respective cultures through the CloudConnect Platform by Cloud Research.[4] We instructed annotators to find an image depicting the target cultural concept using Google Images. We emphasized that the images should be of high quality and do not solely depict the target concept but also include other visuals, to make sure the grounding task is not trivial. For instance, an image for the Korean sauce "gochujang" may contain gochujang along with other dishes.

Bounding Boxes. After selecting the images, annotators used a bounding box tool to draw a single bounding box (bbox) around the target concept. Each annotator was compensated $50 USD for retrieving and annotating images for 50 concepts in their culture.

Verification. We perform an additional analysis step to verify that the cultural concept is not the main focus of the image. We do so by ensuring that the bbox-to-image ratio is less than 0.3. We also used an off-the-shelf object detection model, YOLOv5, to assess the number of objects in the image, filtering out images with fewer than 3 objects.[5] Additionally, annotators were asked whether the concept was prevalent in their culture, and 1.3% of the concepts were marked as not prevalent. This process resulted in the collection of 591 images. More detailed statistics of the collected data are provided in Table 4.

Finally, we conduct a human evaluation to ensure quality by recruiting annotators from CloudConnect. Each annotator was asked to draw bounding boxes for the given cultural concept word. Annotator agreement was measured by calculating the Intersection over Union (IoU) score between the bounding boxes drawn by two different annotators. The IoU is calculated as: $IoU = \frac{|R_{\text{anno1}} \cap R_{\text{anno2}}|}{|R_{\text{anno1}} \cup R_{\text{anno2}}|}$. Each annotator was compensated $0.1 USD of each annotation. More detailed statistics of the collected

[3] Given that we collected annotations for this task, we restricted the task to countries based on the availability of annotators.

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

[5] https://pytorch.org/hub/ultralytics_yolov5/

Region	Country	Number of Concepts	Average bbox/image Ratio	Average Yolov5 Score	Human Eval (IoU)
Latin America	Argentina	43	0.146	4.442	0.92
	Brazil	32	0.153	3.906	0.87
	Mexico	43	0.163	5.744	0.91
North America	Canada	26	0.118	5.500	0.92
East Asia	China	39	0.163	4.106	0.94
	South Korea	41	0.151	5.317	0.87
South Asia	India	53	0.112	5.698	0.88
	Pakistan	38	0.137	4.162	0.86
Middle-East Asia	Israel	48	0.119	5.255	0.91
South East Asia	Philippines	41	0.138	4.390	0.85
	Vietnam	40	0.129	5.275	0.80
Africa	Nigeria	36	0.137	3.611	0.92
	South Africa	34	0.146	4.118	0.88
Europe	Poland	40	0.216	3.150	0.95
	Russia	37	0.134	4.405	0.92

Table 4: Detailed statistics of annotated images across different cultural groups and regions for Cultural Visual Grounding task.

Model	Training Data	Data Size	Vision Encoder	LM
Specialist Models				
Grounding DINO (Liu et al., 2023)	O365, GoldG, Cap4M	-	Swin-T (DINO)	BERT
Generalist Models				
KOSMOS-2 (Peng et al., 2023)	LAION-2B, COYO, GRIT-91M	2.8B	CLIP-ViT-L	Magneto
MiniGPT-v2 (Chen et al., 2023)	LAION, CC3M, SBU, GRIT-20M, VG, RefCOCO, VQA datasets	-	ViT	LLaMA-2-Chat-7B
QwenVL (Bai et al., 2023)	LAION-en/zh, DataComp, COYO, CC, SBU, COCO	1.4B	ViT-bigG	Qwen-7B
LLaVA-1.5 (Liu et al., 2024)	OKVQA, A-OKVQA, OCRVQA, TextCaps, VG, RefCOCO, GQA, ShareGPT	1.2B	CLIP-ViT-L	Vicuna-13B

Table 5: Overview of models benchmarked for the Cultural Visual Grounding task. **Note: Grounding DINO (Liu et al., 2023) and MiniGPT-v2 (Chen et al., 2023) authors do not provide total training data size in the papers, so we leave that blank to avoid inaccurate numbers.

data and human agreement scores (IoU) are provided in Table 4.

4.2 Task Definition and Evaluation Setup

Given an image I and a query q describing a cultural keyword, the goal is to predict a bounding box R around the region in I that corresponds to q. We evaluate models based on the overlap between the gold standard and predicted regions of interest, using Intersection over Union (IoU) as the metric: $IoU = \frac{|R \cap R_{\text{gold}}|}{|R \cup R_{\text{gold}}|}$. We consider a predicted bounding box correct if its IoU with the ground-truth bounding box is greater than 0.5, and report overall accuracy. It is crucial that models perform consistently well across different cultures.

4.3 Models

We benchmark a series of models on our grounding task, considering both *specialist* models, designed explicitly for visual grounding tasks, and *generalist* models, which can handle a wide range of vision-language tasks, such as captioning, question answering, and grounding. These models are listed in Table 5, along with their training data, vision and language backbones, and training methodology.

The specialist model we include is Grounding DINO (Liu et al., 2023), a zero-shot object detection model that combines a Transformer-based detector (DINO; Zhang et al., 2022) with phrase grounding pre-training (GLIP; Li et al., 2022). The generalist models are multimodal large language models (MLLMs). MLLMs encode visual patches as tokens that a language model can understand. They perform visual grounding by generating bounding boxes in textual format, typically in the format of $\langle X_{\text{left}} \rangle \langle Y_{\text{top}} \rangle \langle X_{\text{right}} \rangle \langle Y_{\text{bottom}} \rangle$, denoting the coordinates of the top-left and bottom-right corners of the generated bounding box.

4.4 Results and Analysis

RQ$_1$: Are VLMs able to identify culture-specific concepts? Figure 5 presents the country-level accuracy of each model on the cultural visual grounding task. The overall performance across models is rather poor. Among all models, the specialist model Grounding DINO shows a relatively higher average performance (47.99%) compared to the generalist models.

Analyzing country-specific performance, we observe that KOSMOS-2 and QwenVL-7B exhibit strong accuracy in grounding elements for Canada and Mexico. Grounding DINO, on the other hand, performs well for Poland and the Philippines. All generalist models perform poorly on images from

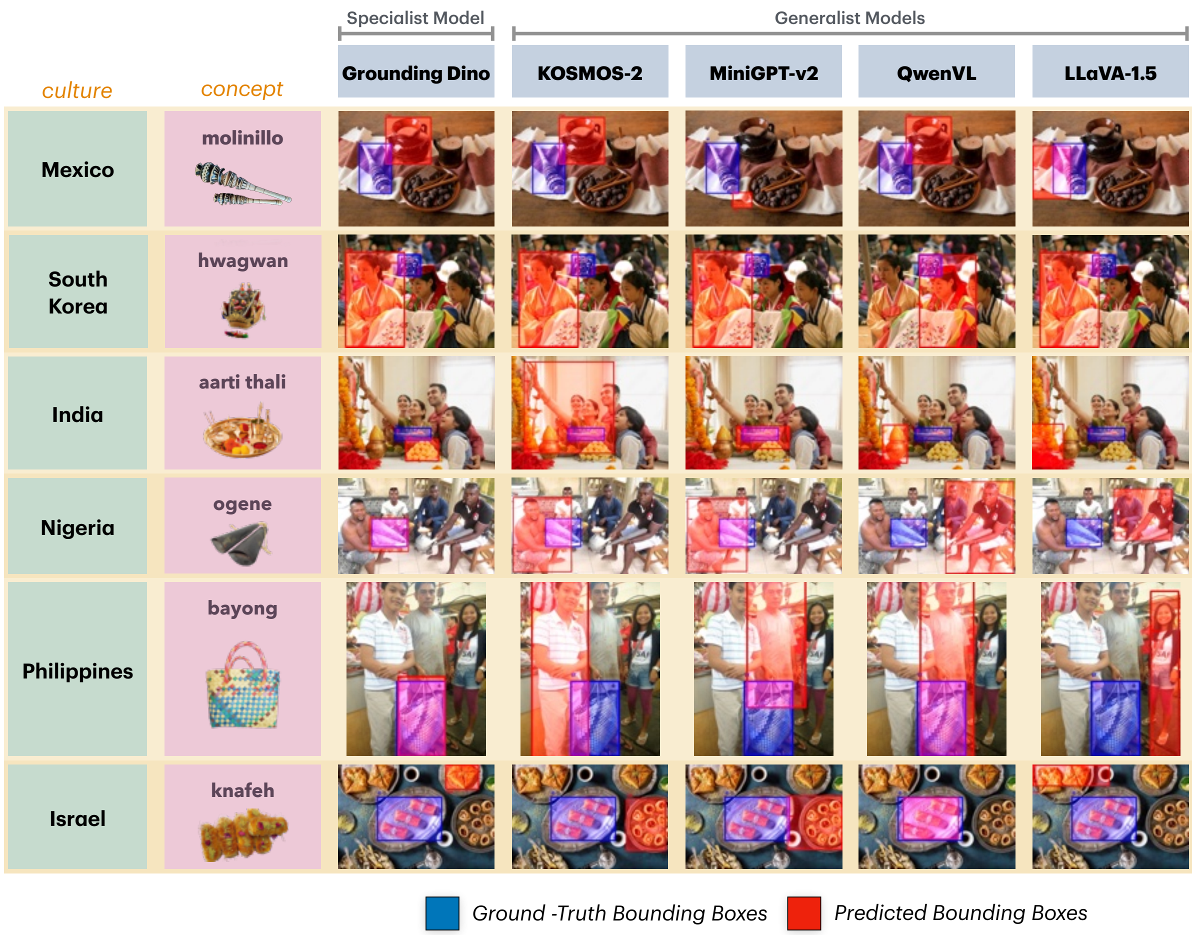


Figure 4: Qualitative Examples showing the performance of specialist and generalist models on Cultural Visual Grounding task.

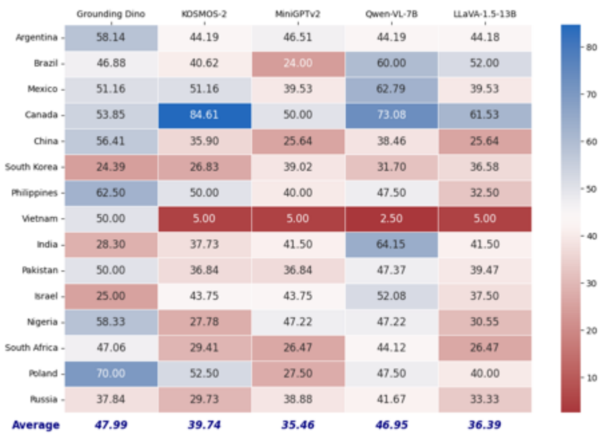


Figure 5: Country-level Accuracy of each model on the Cultural Visual Grounding task.

Vietnam. We hypothesize that this discrepancy stems from a number of factors. First, there may be insufficient coverage of Vietnamese concepts in the pre-training data. Second, the Vietnamese concepts have distinct English spellings and diacritical marks. For example, *wonton* is spelled as “hoành thánh”, which can impact the model’s familiarity with the concept and consequently its accuracy. This highlights the need for future research to focus on developing multilingual, geo-diverse benchmarks to assess performance across a broader range of cultural contexts.

RQ$_2$: Do VLMs exhibit biases towards images from certain cultures? To investigate whether VLMs show biases towards specific cultures, we plot the region-level performance for each model in Figure 6. We observe that almost all models achieve the highest performance on images from North America, with an average accuracy of 64.61%, followed by a considerable drop in performance for images from Latin America (46.99%) and Europe (44.49%). This significant performance disparity may suggest that the VLMs were predominantly trained on images from North America.

Different models vary in their performances in the other regions. The generalist models show the

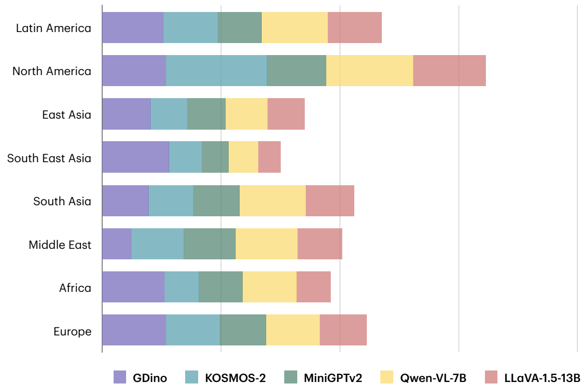


Figure 6: Culture group-level Accuracy for Cultural Visual Grounding.

most difficulty with images from South East Asia (accuracy between 18.75-27.5%) and East Asia (31.11-35.08%) while Grounding DINO performs worst on Middle Eastern images (25%).

RQ$_3$: What challenges do VLMs face in grounding culture-specific concepts? Figure 4 presents some failure cases of the VLMs in the grounding task. We can categorize the errors into two primary types. In the first type, models draw a bounding box around an unrelated object. For example, in the image depicting a "bayong", a type of bag from the Philippines, the models frequently misidentify people as the "bayong". This suggests the model is unfamiliar with the term "bayong" and its visual representation. The other error type occurs when models draw the bounding box around another object with a shape similar to the target object. For instance, for "ogene", a double-bell instrument from Nigeria, some models incorrectly identified a person's arm as the "ogene", which may be due to shape similarity. This may suggest limited familiarity with the concept and its visual form.

5 Conclusion

In this work, we introduced GLOBALRG, a challenging benchmark designed to assess the multicultural understanding of VLMs. GLOBALRG includes two tasks: retrieving culturally diverse images representing universal concepts and visually grounding culture-specific concepts.

Our extensive experiments across a variety of VLMs uncover specific cultural biases. For instance, in the retrieval task, models consistently show varying diversity scores, with a noticeable preference for Western-style elements. Likewise, VLMs demonstrate higher accuracy in grounding images from Western cultures, which underscores the imbalances in their training data. These results reveal an urgent need for more culturally diverse, large-scale datasets and for developing training objectives that explicitly consider a wide range of cultural contexts. Addressing these challenges will pave the way for more inclusive and equitable models that better capture global cultural diversity and mitigate biases in downstream applications.

6 Limitations

While our benchmark, GLOBALRG, provides a comprehensive evaluation of the multicultural understanding of VLMs, it is essential to acknowledge certain limitations as follows,

Cultural Coverage. Although our retrieval task encompasses 50 diverse cultures, the grounding task is restricted to only 15 cultures. This constraint arises from the availability of annotators on the crowdsourcing platform we used, Cloud Research. In future work, we aim to expand the grounding task to include a broader range of cultures.

Restricted cultural concepts. Our study focuses on a selected set of cultural concepts or keywords from the CANDLE dataset. There might be more prominent cultural concepts that we could not cover. This limitation might restrict the comprehensiveness of our evaluation and overlook culturally significant aspects not captured.

Metric for diversity. We currently employ a diversity metric based on entropy to evaluate the cultural diversity of retrieved images. While this metric provides insights into the distribution of images across different cultures, it may not fully capture the nuanced variations in cultural representation. Our approach to regional diversity assessment may lack granularity, potentially overlooking finer distinctions in cultural diversity within regions.

7 Ethical Consideration

Mapping from countries to regions. For the purpose of our tasks, we mapped countries to broad regional categories as specified in Table 1. We acknowledge that cultures do not follow geographic boundaries and that this variation occurs at an individual level, shaped by one's own life experiences. Despite this, we used our mapping as a practical starting point. This approach is a preliminary step, with the ultimate goal of developing systems that can learn from individual user interactions and adapt to diverse and evolving cultures.

Annotator selection and compensation Annotators hired from Cloud Research were predominately based in USA, Canada, Australia, New Zealand, United Kingdom and Ireland. Participation was strictly limited to those who met specific criteria to maintain the relevance of the annotation process. Annotators were required to belong to a chosen ethnicity and to have lived in the designated countries for at least 5 of the past 15 years. This criterion ensured that participants had sufficient cultural context and lived experience relevant to the annotation tasks. We employed a second round of annotators for the human evaluation phase, ensuring none were repeated from the first round.

Inadvertent stereotypes in collect images. We recognize that some images used to capture cultural concepts might inadvertently perpetuate stereotypes. While our goal was to gather authentic cultural representations, we are aware of the ethical implications of including such content. We approached this task with the intention of collecting meaningful cultural data while being mindful of the potential for reinforcing harmful stereotypes.

8 Acknowledgements

This work was funded, in part, by the Vector Institute for AI, Canada CIFAR AI Chairs program, Accelerate Foundation Models Research Program Award from Microsoft, an NSERC discovery grant, and a research gift from AI2.

References

Amith Ananthram, Elias Stengel-Eskin, Carl Vondrick, Mohit Bansal, and Kathleen McKeown. 2024. See it from my perspective: Diagnosing the western cultural bias of large vision-language models in image understanding. *arXiv preprint arXiv:2406.11665*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond.

Abhipsa Basu, R Venkatesh Babu, and Danish Pruthi. 2023. Inspecting the geographical representativeness of images from text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5136–5147.

Mehar Bhatia and Vered Shwartz. 2023. GD-COMET: A geo-diverse commonsense inference model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7993–8001, Singapore. Association for Computational Linguistics.

Donald E Brown. 2004. Human universals, human nature & human culture. *Daedalus*, 133(4):47–54.

Olena Burda-Lassen, Aman Chadha, Shashank Goswami, and Vinija Jain. 2024. How culturally aware are vision-language models? *arXiv preprint arXiv:2405.17475*.

Yong Cao, Yova Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. 2024. Cultural adaptation of recipes. *Transactions of the Association for Computational Linguistics*, 12:80–99.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Terrance De Vries, Ishan Misra, Changhan Wang, and Laurens Van der Maaten. 2019. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 52–59.

Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369.

Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Ishan Misra, Levent Sagun, Armand Joulin, and Piotr Bojanowski. 2022. Vision models are more robust and fair when pretrained on uncurated images without supervision. *arXiv preprint arXiv:2202.08360*.

Tanmay Gupta, Ryan Marten, Aniruddha Kembhavi, and Derek Hoiem. 2022. Grit: General robust image task benchmark.

Laura Gustafson, Megan Richards, Melissa Hall, Caner Hazirbas, Diane Bouchacourt, and Mark Ibrahim. 2023. Pinpointing why object recognition performance degrades across income levels and geographies. *arXiv preprint arXiv:2304.05391*.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Oana Ignat, Longju Bai, Joan C Nwatu, and Rada Mihalcea. 2024. Annotations on a budget: Leveraging geo-data similarity to balance model performance and annotation cost. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1239–1259.

Akshita Jha, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Rida Qadri, Chandan K Reddy, and Sunipa Dev. 2024. Beyond the surface: A global-scale analysis of visual stereotypes in text-to-image generation. *arXiv preprint arXiv:2401.06310*.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8269–8284, Toronto, Canada. Association for Computational Linguistics.

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

Antonia Karamolegkou, Phillip Rust, Yong Cao, Ruixiang Cui, Anders Søgaard, and Daniel Hershcovich. 2024. Vision-language models under cultural and inclusive considerations. *arXiv preprint arXiv:2407.06177*.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar. Association for Computational Linguistics.

Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. 2024. An image speaks a thousand words, but can everyone listen? on translating images for cultural relevance. *arXiv preprint arXiv:2404.01247*.

Zu Kim, André Araujo, Bingyi Cao, Cam Askew, Jack Sim, Mike Green, N Yilla, and Tobias Weyand. 2021. Towards a fairer landmark recognition dataset. *arXiv preprint arXiv:2108.08874*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Benno Krojer, Vaibhav Adlakha, Vibhav Vineet, Yash Goyal, Edoardo Ponti, and Siva Reddy. 2022. Image retrieval from contextual descriptions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3426–3440, Dublin, Ireland. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021a. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. 2021b. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2125–2134.

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

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023*, pages 1907–1917.

Shramay Palta and Rachel Rudinger. 2023. FORK: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9952–9962, Toronto, Canada. Association for Computational Linguistics.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Steven Pinker. 2004. The blank slate: The modern denial of human nature. *New York, NY, Viking. Popper, K.(1974). Unended Quest. Fontana, London*.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

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

Vikram V Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. 2024. Geode: a geographically diverse evaluation dataset for object recognition. *Advances in Neural Information Processing Systems*, 36.

William A Gaviria Rojas, Sudnya Diamos, Keertan Ranjan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. 2022. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. 2017. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650.

Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. 2022. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680.

Andre Ye, Sebastin Santy, Jena D Hwang, Amy X Zhang, and Ranjay Krishna. 2023. Cultural and linguistic diversity improves visual representations. *arXiv preprint arXiv:2310.14356*.

Da Yin, Liunian Harold Li, Ziniu Hu, Nanyun Peng, and Kai-Wei Chang. 2021. Broaden the vision: Geo-diverse visual commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2115–2129, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. 2022. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. 2022. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.

Wangchunshu Zhou, Yan Zeng, Shizhe Diao, and Xinsong Zhang. 2022. Vlue: A multi-task multi-dimension benchmark for evaluating vision-language pre-training. In *International Conference on Machine Learning*, pages 27395–27411. PMLR.

Country	Clothing	Food	Drinks	Rituals	Traditions
Argentina	4	30	3	2	4
Brazil	3	21	3	2	3
Canada	2	17	2	2	3
China	2	13	2	5	17
India	12	21	2	6	12
Israel	5	29	2	7	5
Mexico	5	24	3	4	7
Nigeria	7	18	2	2	7
Pakistan	13	18	1	2	4
Philippines	4	14	3	3	17
Poland	0	37	1	0	2
Russia	5	11	1	2	18
South Africa	3	16	3	2	10
South Korea	6	20	2	2	11
Vietnam	4	27	3	2	4
Total	**75**	**316**	**33**	**43**	**124**
Proportion	**12.69%**	**53.46%**	**5.58%**	**7.27%**	**20.99%**

Table 6: Distribution of Cultural Keywords in Cultural Visual Grounding Dataset.

A Appendix

A.1 Complete Set of Results for Retrieval across Universals task

A.1.1 Results Across All Metrics

Table 7 and 8 details results across all models. We show results for each universal and each metric.

A.1.2 Results Across All Countries

Table 9 and 10 details the first 10 retrieved countries for each model and each universal.

A.1.3 Results Across All Regions

Table 11 and 12 details the first 10 retrieved regions for each model and each universal.

A.2 List of Cultural Keywords in Cultural Visual Grounding dataset

Table 13 lists the cultural concepts for each country in the Cultural Visual Grounding Dataset.

A.3 Distribution of Cultural Keywords in Cultural Visual Grounding Dataset

Table 6 lists the statistics and distribution among the cultural keywords for each country in the Cultural Visual Grounding Dataset to illustrate the diversity and focus of the dataset.

A.4 Model Checkpoints

- **CLIP**: laion/CLIP-ViT-g-14-laion2B-s12B-b42K
- **OpenCLIP**: clip-vit-base-patch32
- **Coca** CoCa-ViT-B-32-laion2B-s13B-b90k
- **llava**: llava-hf/llava-1.5-13b-hf
- **Qwen**: Qwen/Qwen-VL-Chat

Metric	Model	breakfast	clothing	dance	dessert	dinner	drinks	eating habits	farming	festival	funeral
Regional Diversity @ 10	CLIP	65.35	69.9	65.35	65.35	63.88	69.9	73.65	69.9	47.29	65.05
	OpenCLIP	73.65	69.9	73.65	73.65	79.67	79.67	63.88	67.62	40.97	63.88
	CoCA	65.35	81.94	63.88	79.67	50.74	59.03	65.05	45.81	59.33	53.31
	TCL	63.88	55.58	63.88	73.65	79.67	73.65	73.65	61.6	57.86	69.9
	ALBEF	71.37	57.06	71.37	75.92	65.35	79.67	59.03	40.84	63.88	69.9
	BLIP-2	55.58	71.37	65.05	61.6	71.37	81.94	73.65	34.82	73.65	65.05
	FLAVA	69.9	27.75	81.94	67.62	59.33	65.35	73.65	67.62	69.9	59.03
Regional Diversity @5	CLIP	82.77	59.04	82.77	82.77	65.55	82.77	59.04	59.04	59.04	31.09
	OpenCLIP	59.04	82.77	82.77	100	82.77	65.55	59.04	82.77	41.82	65.55
	CoCA	82.77	100	65.55	82.77	0	82.77	82.77	65.55	82.77	31.09
	TCL	82.77	65.55	65.55	82.77	100	82.77	100	65.55	82.77	65.55
	ALBEF	82.77	65.55	82.77	100	82.77	82.77	82.77	31.09	65.55	31.09
	BLIP-2	82.77	65.55	82.77	82.77	100	100	82.77	41.82	100	59.04
	FLAVA	82.77	59.04	100	65.55	65.55	59.04	82.77	65.55	59.04	82.77
Country Diversity @10	CLIP	93.98	100	100	87.96	100	87.96	100	93.98	93.98	93.98
	OpenCLIP	93.98	85.69	100	100	93.98	93.98	93.98	100	93.98	93.98
	CoCA	79.67	100	100	100	100	93.98	93.98	100	93.98	87.96
	TCL	87.96	100	87.96	93.98	93.98	100	87.96	93.98	93.98	87.96
	ALBEF	79.67	93.98	85.69	100	93.98	93.98	100	100	87.96	93.98
	BLIP-2	85.69	100	93.98	100	87.96	100	87.96	87.96	100	93.98
	FLAVA	100	85.69	93.98	100	79.67	93.98	100	93.98	100	93.98
Country Diversity @5	CLIP	100	100	100	100	100	82.77	100	100	82.77	100
	OpenCLIP	100	82.77	100	100	82.77	100	100	100	82.77	82.77
	CoCA	82.77	100	100	100	100	100	100	100	100	100
	TCL	82.77	100	100	100	100	100	100	82.77	82.77	100
	ALBEF	82.77	100	100	100	100	100	100	100	82.77	82.77
	BLIP-2	100	100	100	100	100	100	100	82.77	100	100
	FLAVA	100	100	100	100	82.77	100	100	100	100	100
Relevance@10	CLIP	100	100	100	0	100	100	100	100	0	100
	OpenCLIP	100	100	100	100	0	100	100	100	0	100
	CoCA	100	90	80	100	20	100	100	100	30	100
	TCL	100	30	100	90	30	100	100	100	80	90
	ALBEF	90	30	80	100	20	100	100	100	50	100
	BLIP-2	100	50	100	90	0	100	90	100	90	100
	FLAVA	80	20	70	40	20	90	100	100	30	100
Relevance@5	CLIP	100	80	100	50	50	100	90	100	40	100
	OpenCLIP	100	60	70	60	0	100	100	100	30	100
	CoCA	100	80	100	100	20	100	100	100	40	100
	TCL	100	40	100	100	40	100	100	100	80	80
	ALBEF	80	20	100	100	0	100	100	100	60	100
	BLIP-2	100	40	100	100	0	100	80	100	100	100
	FLAVA	100	0	80	40	20	80	100	100	20	100

Table 7: First half of the results across all metrics and models for *Retrieval Across Universals* task.

Metric	Model	greeting	headcoverings	instrument	lunch	marriage	music	religion	ritual	sports	transport
	CLIP	53.01	55.58	69.9	61.6	47.29	53.01	73.65	73.65	75.92	73.65
	OpenCLIP	63.88	65.35	63.88	61.6	81.94	63.88	65.35	65.35	73.65	47.29
	CoCA	73.65	71.37	53.31	55.58	63.88	59.03	75.92	75.92	71.37	73.65
Regional Diversity@10	TCL	79.67	79.67	69.9	63.88	50.74	73.65	34.82	63.88	65.35	75.92
	ALBEF	69.9	73.65	73.65	67.62	73.65	40.84	73.65	53.01	67.62	44.72
	BLIP-2	73.65	57.06	75.92	73.65	50.74	69.9	50.74	67.62	39	53.01
	FLAVA	65.35	75.92	63.88	81.94	44.72	67.62	73.65	81.94	69.9	69.9
	CLIP	59.04	59.04	59.04	65.55	100	82.77	82.77	82.77	82.77	65.55
	OpenCLIP	41.82	82.77	82.77	65.55	100	82.77	82.77	82.77	65.55	59.04
	CoCA	82.77	59.04	31.09	65.55	59.04	59.04	82.77	65.55	82.77	100
Regional Diversity@5	TCL	82.77	82.77	82.77	59.04	41.82	100	31.09	59.04	65.55	82.77
	ALBEF	41.82	82.77	65.55	65.55	65.55	59.04	65.55	31.09	65.55	65.55
	BLIP-2	100	82.77	82.77	59.04	31.09	82.77	59.04	82.77	41.82	65.55
	FLAVA	65.55	100	65.55	82.77	65.55	65.55	82.77	82.77	82.77	31.09
	CLIP	87.96	93.98	100	100	93.98	100	87.96	85.69	93.98	87.96
	OpenCLIP	100	93.98	100	100	100	85.69	85.69	100	93.98	93.98
	CoCA	87.96	100	100	93.98	93.98	93.98	100	100	100	87.96
Country Diversity@10	TCL	93.98	93.98	93.98	93.98	50.74	100	93.98	93.98	93.98	87.96
	ALBEF	87.96	100	93.98	87.96	93.98	79.67	93.98	93.98	87.96	73.65
	BLIP-2	81.94	87.96	93.98	100	87.96	93.98	93.98	100	87.96	93.98
	FLAVA	100	100	100	93.98	93.98	87.96	93.98	100	93.98	93.98
	CLIP	82.77	82.77	100	100	100	100	82.77	82.77	100	82.77
	OpenCLIP	100	100	100	100	100	100	100	100	82.77	100
	CoCA	100	100	100	100	82.77	100	100	100	100	100
Country Diversity@5	TCL	100	82.77	100	82.77	41.82	100	100	100	100	100
	ALBEF	65.55	100	100	100	82.77	82.77	100	100	100	65.55
	BLIP-2	100	100	100	100	82.77	100	100	100	100	100
	FLAVA	100	100	100	82.77	100	82.77	100	100	100	82.77
	CLIP	0	100	100	0	100	0	100	0	100	100
	OpenCLIP	100	100	100	0	0	0	100	100	100	100
	CoCA	60	90	100	30	90	60	90	50	100	100
Relevance@10	TCL	10	20	90	70	80	70	80	40	100	100
	ALBEF	10	30	100	50	80	90	40	30	100	100
	BLIP-2	40	50	80	0	90	90	90	30	100	100
	FLAVA	50	40	100	20	20	30	90	50	90	100
	CLIP	50	60	90	30	90	0	80	40	100	100
	OpenCLIP	40	100	100	30	60	30	70	40	100	100
	CoCA	40	100	100	0	100	80	100	60	100	100
Relevance@5	TCL	0	20	80	60	80	80	80	60	100	100
	ALBEF	20	0	100	40	80	100	40	20	100	100
	BLIP-2	20	60	80	0	80	100	80	40	100	100
	FLAVA	60	0	100	20	40	20	80	60	80	100

Table 8: Second half of the results across all metrics and models for *Retrieval Across Universals* task.

Model	First@10 Country	breakfast	clothing	dance	dessert	dinner	drinks	eating_habits	farming	festival	funeral
CLIP	country 0	germany	italy	italy	fiji	vietnam	australia	ethiopia	hungary	hungary	china
	country 1	south africa	sri lanka	us	thailand	russia	thailand	netherlands	italy	hungary	germany
	country 2	canada	france	france	uk	tunisia	iran	srilanka	fiji	sweden	italy
	country 3	kenya	greece	australia	phillippines	ethiopia	italy	germany	poland	singapore	france
	country 4	australia	fiji	chile	south africa	portugal	italy	poland	japan	new zealand	uk
	country 5	somalia	morocco	canada	new zealand	us	jamaica	canada	ethiopia	japan	australia
	country 6	italy	hungary	uk	hungary	germany	peru	south korea	lebanon	greece	us
	country 7	argentina	australia	philippines	new zealand	hungary	greece	brazil	spain	bulgaria	peru
	country 8	france	indonesia	brazil	egypt	canada	indonesia	peru	japan	australia	italy
	country 9	canada	mexico	argentina	uk	peru	greece	japan	peru	poland	tanzania
OpenCLIP	country 0	germany	peru	jamaica	new zealand	phillippines	indonesia	bulgaria	spain	portugal	us
	country 1	canada	morocco	uganda	chile	us	iran	peru	pakistan	bulgaria	australia
	country 2	france	france	bulgaria	netherlands	peru	phillippines	uk	ethiopia	tunisia	uk
	country 3	italy	turkey	philippines	egypt	morocco	jamaica	france	lebanon	bulgaria	germany
	country 4	tanzania	peru	tanzania	tanzania	phillippines	egypt	egypt	ghana	kenya	us
	country 5	new zealand	mexico	new zealand	singapore	south korea	france	vietnam	bulgaria	new zealand	chile
	country 6	singapore	south korea	sweden	saudi arabia	sweden	phillippines	netherlands	egypt	france	mexico
	country 7	kenya	japan	australia	south africa	vietnam	tanzania	vietnam	india	nigeria	turkey
	country 8	argentina	peru	greece	lebanon	iran	australia	us	phillippines	uganda	portugal
	country 9	canada	singapore	chile	indonesia	france	brazil	brazil	hungary	morocco	italy
CoCA	country 0	canada	morocco	spain	ethiopia	uk	iran	uk	italy	uk	uk
	country 1	canada	italy	australia	indonesia	poland	australia	peru	ghana	sweden	chile
	country 2	france	indonesia	us	somalia	italy	italy	egypt	lebanon	somalia	italy
	country 3	kenya	argentina	italy	mexico	sweden	indonesia	us	ethiopia	new zealand	bulgaria
	country 4	argentina	egypt	canada	south korea	france	greece	netherlands	saudi arabia	chile	france
	country 5	uk	chile	france	germany	peru	italy	poland	portugal	kenya	bulgaria
	country 6	south africa	us	chile	saudi arabia	kenya	france	france	south africa	australia	australia
	country 7	kenya	iran	hungary	spain	chile	peru	phillippines	nigeria	peru	japan
	country 8	canada	japan	jamaica	us	brazil	singapore	kenya	hungary	new zealand	italy
	country 9	singapore	france	new zealand	italy	australia	bulgaria	poland	spain	tunisia	indonesia
TCL	country 0	canada	ethiopia	tanzania	italy	france	thailand	brazil	kenya	australia	peru
	country 1	south africa	ghana	kenya	tanzania	us	greece	nigeria	italy	uk	fiji
	country 2	uk	hungary	australia	us	china	bulgaria	france	india	argentina	germany
	country 3	singapore	saudi arabia	peru	south korea	india	egypt	us	italy	china	spain
	country 4	canada	spain	brazil	south africa	indonesia	australia	egypt	tunisia	china	mexico
	country 5	uk	turkey	australia	russia	bulgaria	indonesia	poland	germany	peru	mexico
	country 6	hungary	tunisia	lebanon	canada	south korea	peru	chile	saudi arabia	brazil	peru
	country 7	poland	somalia	sri lanka	canada	china	jamaica	us	nigeria	new zealand	indonesia
	country 8	philippines	phillippines	mexico	new zealand	peru	vietnam	brazil	morocco	mexico	uganda
	country 9	australia	germany	mexico	brazil	fiji	turkey	south korea	thailand	germany	lebanon
ALBEF	country 0	canada	ethiopia	tanzania	italy	france	thailand	brazil	kenya	australia	peru
	country 1	south africa	ghana	kenya	tanzania	us	greece	nigeria	italy	uk	fiji
	country 2	uk	hungary	australia	us	china	bulgaria	france	india	argentina	germany
	country 3	singapore	saudi arabia	peru	south korea	india	egypt	us	italy	china	spain
	country 4	canada	spain	brazil	south africa	indonesia	australia	egypt	tunisia	china	mexico
	country 5	uk	turkey	australia	russia	bulgaria	indonesia	poland	germany	peru	mexico
	country 6	hungary	tunisia	lebanon	canada	south korea	peru	chile	saudi arabia	brazil	peru
	country 7	poland	somalia	sri lanka	canada	china	jamaica	us	nigeria	new zealand	indonesia
	country 8	philippines	phillippines	mexico	new zealand	peru	vietnam	brazil	morocco	mexico	uganda
	country 9	australia	germany	mexico	brazil	fiji	turkey	south korea	thailand	germany	lebanon
BLIP-2	country 0	brazil	morocco	italy	italy	thailand	indonesia	saudi arabia	canada	new zealand	tanzania
	country 1	new zealand	ghana	australia	uganda	france	ghana	france	poland	uk	sweden
	country 2	canada	canada	egypt	egypt	peru	australia	egypt	us	us	uk
	country 3	uk	lebanon	us	us	egypt	pakistan	canada	uk	japan	bulgaria
	country 4	france	egypt	portugal	sweden	us	iran	jamaica	poland	lebanon	us
	country 5	canada	chile	greece	australia	sweden	france	singapore	russia	south korea	bulgaria
	country 6	sweden	poland	spain	somalia	iran	mexico	jamaica	russia	italy	australia
	country 7	italy	tunisia	ethiopia	hungary	france	greece	tanzania	nigeria	canada	turkey
	country 8	argentina	turkey	jamaica	bulgaria	morocco	thailand	spain	hungary	jamaica	mexico
	country 9	canada	srilanka	spain	south africa	egypt	ethiopia	france	france	china	spain
FLAVA	country 0	jamaica	south korea	china	saudi arabia	jamaica	kenya	brazil	italy	ghana	germany
	country 1	canada	somalia	thailand	ghana	ethiopia	vietnam	kenya	south africa	new zealand	italy
	country 2	south africa	nigeria	pakistan	egypt	south africa	italy	china	saudi arabia	pakistan	australia
	country 3	poland	srilanka	tanzania	us	jamaica	tunisia	ethiopia	egypt	morocco	indonesia
	country 4	kenya	tunisia	greece	kenya	vietnam	somalia	greece	portugal	somalia	turkey
	country 5	new zealand	tunisia	new zealand	tanzania	jamaica	fiji	italy	nigeria	portugal	uk
	country 6	ghana	tunisia	australia	srilanka	hungary	poland	pakistan	saudi arabia	uk	vietnam
	country 7	turkey	ghana	india	germany	indonesia	jamaica	canada	india	thailand	italy
	country 8	germany	kenya	tanzania	italy	greece	vietnam	netherlands	fiji	tanzania	russia
	country 9	somalia	morocco	jamaica	canada	south africa	france	us	pakistan	jamaica	uganda

Table 9: First half of the results for first 10 retrieved countries for *Retrieval Across Universals* task.

Model	First@10 Country	greetings	headcoverings	instrument	lunch	marriage	music	religion	ritual	sports	transport
CLIP	country 0	vietnam	tunisia	hungary	vietnam	portugal	somalia	thailand	thailand	thailand	ethiopia
	country 1	thailand	morocco	netherlands	sweden	germany	russia	china	srilanka	italy	peru
	country 2	us	greece	sweden	indonesia	brazil	turkey	peru	germany	us	china
	country 3	south korea	china	vietnam	portugal	hungary	peru	jamaica	thailand	hungary	peru
	country 4	thailand	morocco	brazil	peru	greece	netherlands	china	china	fiji	kenya
	country 5	vietnam	egypt	chile	phillippines	australia	hungary	tanzania	ethiopia	japan	thailand
	country 6	indonesia	hungary	iran	germany	canada	uk	kenya	morocco	turkey	india
	country 7	morocco	lebanon	france	hungary	russia	uganda	thailand	japan	hungary	nigeria
	country 8	japan	iran	morocco	south korea	argentina	argentina	morocco	turkey	netherlands	egypt
	country 9	tunisia	somalia	pakistan	tanzania	russia	france	saudi arabia	thailand	ghana	india
OpenCLIP	country 0	phillippines	canada	kenya	peru	srilanka	mexico	lebanon	australia	phillippines	nigeria
	country 1	singapore	australia	sweden	sweden	russia	ghana	somalia	china	india	kenya
	country 2	chile	south africa	iran	france	iran	germany	australia	india	somalia	china
	country 3	mexico	germany	ethiopia	vietnam	singapore	argentina	iran	indonesia	india	egypt
	country 4	indonesia	ghana	canada	indonesia	argentina	australia	pakistan	south korea	ethiopia	ethiopia
	country 5	ghana	iran	us	us	phillippines	ghana	ethiopia	singapore	srilanka	ethiopia
	country 6	somalia	greece	netherlands	hungary	tunisia	tanzania	pakistan	argentina	germany	tunisia
	country 7	bulgaria	lebanon	hungary	south korea	us	saudi arabia	pakistan	brazil	indonesia	peru
	country 8	thailand	italy	germany	spain	canada	poland	tanzania	peru	argentina	brazil
	country 9	srilanka	ghana	china	phillippines	india	ghana	indonesia	phillippines	egypt	uganda
CoCA	country 0	vietnam	canada	sweden	peru	morocco	netherlands	lebanon	mexico	phillippines	singapore
	country 1	peru	greece	portugal	brazil	russia	peru	south africa	brazil	jamaica	pakistan
	country 2	srilanka	italy	chile	jamaica	singapore	spain	jamaica	srilanka	south korea	kenya
	country 3	singapore	tunisia	hungary	germany	russia	kenya	tanzania	india	india	saudi arabia
	country 4	tunisia	germany	netherlands	hungary	spain	hungary	italy	singapore	srilanka	brazil
	country 5	tunisia	argentina	us	canada	argentina	somalia	morocco	south korea	uk	saudi arabia
	country 6	china	egypt	france	sweden	iran	us	new zealand	somalia	egypt	somalia
	country 7	russia	iran	china	us	hungary	portugal	chile	fiji	sweden	kenya
	country 8	singapore	turkey	kenya	france	phillippines	iran	ghana	morocco	bulgaria	poland
	country 9	hungary	singapore	uk	canada	saudi arabia	hungary	pakistan	vietnam	pakistan	russia
TCL	country 0	chile	australia	japan	japan	thailand	turkey	mexico	ghana	phillippines	pakistan
	country 1	germany	egypt	australia	south korea	india	japan	ghana	kenya	vietnam	lebanon
	country 2	peru	pakistan	sweden	japan	india	tunisia	kenya	fiji	hungary	egypt
	country 3	china	phillippines	italy	somalia	india	italy	ethiopia	singapore	tanzania	thailand
	country 4	tanzania	australia	pakistan	spain	thailand	pakistan	nigeria	morocco	ghana	jamaica
	country 5	china	thailand	uk	uganda	nigeria	ethiopia	somalia	iran	indonesia	fiji
	country 6	srilanka	singapore	sweden	canada	india	canada	tanzania	thailand	india	jamaica
	country 7	singapore	brazil	nigeria	hungary	india	poland	peru	kenya	india	uganda
	country 8	jamaica	poland	fiji	china	thailand	india	bulgaria	srilanka	somalia	pakistan
	country 9	brazil	tunisia	saudi arabia	turkey	egypt	hungary	ethiopia	saudi arabia	egypt	australia
ALBEF	country 0	chile	australia	japan	japan	thailand	turkey	mexico	ghana	phillippines	pakistan
	country 1	germany	egypt	australia	south korea	india	japan	ghana	kenya	vietnam	lebanon
	country 2	peru	pakistan	sweden	japan	india	tunisia	kenya	fiji	hungary	egypt
	country 3	china	phillippines	italy	somalia	india	italy	ethiopia	singapore	tanzania	thailand
	country 4	tanzania	australia	pakistan	spain	thailand	pakistan	nigeria	morocco	ghana	jamaica
	country 5	china	thailand	uk	uganda	nigeria	ethiopia	somalia	iran	indonesia	fiji
	country 6	srilanka	singapore	sweden	canada	india	canada	tanzania	thailand	india	jamaica
	country 7	singapore	brazil	nigeria	hungary	india	poland	peru	kenya	india	uganda
	country 8	jamaica	poland	fiji	china	thailand	india	bulgaria	srilanka	somalia	pakistan
	country 9	brazil	tunisia	saudi arabia	turkey	egypt	hungary	ethiopia	saudi arabia	egypt	australia
BLIP-2	country 0	china	ethiopia	pakistan	uk	iran	singapore	ghana	vietnam	singapore	somalia
	country 1	fiji	hungary	iran	iran	phillippines	japan	greece	indonesia	thailand	tunisia
	country 2	portugal	south korea	south africa	spain	egypt	france	ethiopia	sweden	srilanka	lebanon
	country 3	somalia	turkey	nigeria	greece	saudi arabia	new zealand	jamaica	brazil	indonesia	thailand
	country 4	egypt	germany	japan	brazil	iran	indonesia	tunisia	morocco	india	egypt
	country 5	china	nigeria	sweden	new zealand	saudi arabia	tunisia	morocco	srilanka	phillippines	india
	country 6	chile	poland	pakistan	south korea	spain	pakistan	greece	kenya	vietnam	indonesia
	country 7	egypt	nigeria	indonesia	lebanon	sweden	portugal	mexico	germany	vietnam	kenya
	country 8	somalia	lebanon	turkey	fiji	australia	portugal	bulgaria	pakistan	india	kenya
	country 9	saudi arabia	turkey	germany	phillippines	greece	hungary	uganda	netherlands	brazil	ghana
FLAVA	country 0	vietnam	vietnam	italy	peru	iran	mexico	phillippines	srilanka	srilanka	nigeria
	country 1	indonesia	portugal	sweden	jamaica	russia	sweden	tanzania	canada	bulgaria	china
	country 2	uganda	south korea	china	jamaica	tunisia	china	jamaica	thailand	somalia	somalia
	country 3	us	egypt	uganda	ethiopia	tanzania	china	canada	lebanon	morocco	ethiopia
	country 4	canada	nigeria	ghana	sweden	portugal	italy	ghana	vietnam	indonesia	ethiopia
	country 5	somalia	argentina	south africa	vietnam	nigeria	netherlands	phillippines	argentina	singapore	saudi arabia
	country 6	chile	jamaica	mexico	fiji	tanzania	pakistan	peru	china	phillippines	fiji
	country 7	singapore	somalia	pakistan	south africa	ethiopia	saudi arabia	kenya	japan	indonesia	russia
	country 8	tunisia	south africa	argentina	phillippines	lebanon	pakistan	spain	portugal	egypt	lebanon
	country 9	greece	ghana	kenya	pakistan	poland	chile	mexico	india	china	pakistan

Table 10: Second half of the results for first 10 retrieved countries for *Retrieval Across Universals* task.

Model	First@10 Regions	breakfast	clothing	dance	dessert	dinner	drinks	eating_habits	farming	festival	funeral
CLIP	region 0	Europe	Europe	Europe	Oceania	SE Asia	Oceania	Africa	Europe	Europe	East Asia
	region 1	Africa	South Asia	N America	SE Asia	Europe	SE Asia	Europe	Europe	Europe	Europe
	region 2	N America	Europe	Europe	Europe	Africa	ME Asia	South Asia	Oceania	Europe	Europe
	region 3	Africa	Europe	Oceania	SE Asia	Africa	Europe	Europe	Europe	SE Asia	Europe
	region 4	Oceania	Oceania	L America	Africa	Europe	Europe	Europe	East Asia	Oceania	Europe
	region 5	Africa	Africa	N America	Oceania	N America	Caribbean	N America	Africa	East Asia	Oceania
	region 6	Europe	Europe	Europe	Europe	Europe	L America	East Asia	ME Asia	Europe	N America
	region 7	L America	Oceania	SE Asia	Oceania	Europe	Europe	L America	Europe	Europe	L America
	region 8	Europe	SE Asia	L America	ME Asia	N America	SE Asia	L America	East Asia	Oceania	Europe
	region 9	N America	L America	L America	Europe	L America	Europe	East Asia	L America	Europe	Africa
OpenCLIP	region 0	Europe	L America	Caribbean	Oceania	SE Asia	SE Asia	Europe	Europe	Europe	N America
	region 1	N America	Africa	Africa	L America	N America	ME Asia	L America	South Asia	Europe	Oceania
	region 2	Europe	Europe	Europe	Europe	L America	SE Asia	Europe	Africa	Africa	Europe
	region 3	Europe	ME Asia	SE Asia	ME Asia	Africa	Caribbean	Europe	ME Asia	Europe	Europe
	region 4	Africa	L America	Africa	Africa	SE Asia	ME Asia	ME Asia	Africa	Africa	N America
	region 5	Oceania	L America	Oceania	SE Asia	East Asia	Europe	SE Asia	Europe	Oceania	L America
	region 6	SE Asia	East Asia	Europe	ME Asia	Europe	SE Asia	Europe	ME Asia	Europe	L America
	region 7	Africa	East Asia	Oceania	Africa	SE Asia	Africa	SE Asia	South Asia	Africa	ME Asia
	region 8	L America	L America	Europe	ME Asia	ME Asia	Oceania	N America	SE Asia	Africa	Europe
	region 9	N America	SE Asia	L America	SE Asia	Europe	L America	L America	Europe	Africa	Europe
CoCA	region 0	N America	Africa	Europe	Africa	Europe	ME Asia	Europe	Europe	Europe	Europe
	region 1	N America	Europe	Oceania	SE Asia	Europe	Oceania	L America	Africa	Europe	L America
	region 2	Europe	SE Asia	N America	Africa	Europe	Europe	ME Asia	ME Asia	Africa	Europe
	region 3	Africa	L America	Europe	L America	Europe	SE Asia	N America	Africa	Oceania	Europe
	region 4	L America	ME Asia	N America	East Asia	Europe	Europe	Europe	ME Asia	L America	Europe
	region 5	Europe	L America	Europe	Europe	L America	Europe	Europe	Europe	Africa	Europe
	region 6	Africa	N America	L America	ME Asia	Africa	Europe	Europe	Africa	Oceania	Oceania
	region 7	Africa	ME Asia	Europe	Europe	L America	L America	SE Asia	Africa	L America	East Asia
	region 8	N America	East Asia	Caribbean	N America	L America	SE Asia	Africa	Europe	Oceania	Europe
	region 9	SE Asia	Europe	Oceania	Europe	Oceania	Europe	Europe	Europe	Africa	SE Asia
TCL	region 0	N America	Africa	Africa	Europe	Europe	SE Asia	L America	Africa	Oceania	L America
	region 1	Africa	Africa	Africa	Africa	N America	Europe	Africa	Europe	Europe	Oceania
	region 2	Europe	Europe	Oceania	N America	East Asia	Europe	Europe	South Asia	L America	Europe
	region 3	SE Asia	ME Asia	L America	East Asia	South Asia	ME Asia	N America	Europe	East Asia	Europe
	region 4	N America	Europe	L America	Africa	SE Asia	Oceania	ME Asia	Africa	East Asia	L America
	region 5	Europe	ME Asia	Oceania	Europe	Europe	SE Asia	Europe	Europe	L America	L America
	region 6	Europe	Africa	ME Asia	N America	East Asia	L America	L America	ME Asia	L America	L America
	region 7	Europe	Africa	South Asia	N America	East Asia	Caribbean	N America	Africa	Oceania	SE Asia
	region 8	SE Asia	SE Asia	L America	Oceania	L America	SE Asia	L America	Africa	L America	Africa
	region 9	Oceania	Europe	L America	L America	Oceania	ME Asia	East Asia	SE Asia	Europe	ME Asia
ALBEF	region 0	N America	Africa	Africa	Europe	Europe	SE Asia	L America	Africa	Oceania	L America
	region 1	Africa	Africa	Africa	Africa	N America	Europe	Africa	Europe	Europe	Oceania
	region 2	Europe	Europe	Oceania	N America	East Asia	Europe	Europe	South Asia	L America	Europe
	region 3	SE Asia	ME Asia	L America	East Asia	South Asia	ME Asia	N America	Europe	East Asia	Europe
	region 4	N America	Europe	L America	Africa	SE Asia	Oceania	ME Asia	Africa	East Asia	L America
	region 5	Europe	ME Asia	Oceania	Europe	Europe	SE Asia	Europe	Europe	L America	L America
	region 6	Europe	Africa	ME Asia	N America	East Asia	L America	L America	ME Asia	L America	L America
	region 7	Europe	Africa	South Asia	N America	East Asia	Caribbean	N America	Africa	Oceania	SE Asia
	region 8	SE Asia	SE Asia	L America	Oceania	L America	SE Asia	L America	Africa	L America	Africa
	region 9	Oceania	Europe	L America	L America	Oceania	ME Asia	East Asia	SE Asia	Europe	ME Asia
BLIP-2	region 0	L America	Africa	Europe	Europe	SE Asia	SE Asia	ME Asia	N America	Oceania	Africa
	region 1	Oceania	Africa	Oceania	Africa	Europe	Africa	Europe	Europe	Europe	Europe
	region 2	N America	N America	ME Asia	ME Asia	L America	Oceania	ME Asia	N America	N America	Europe
	region 3	Europe	ME Asia	N America	N America	ME Asia	South Asia	N America	Europe	East Asia	Europe
	region 4	Europe	ME Asia	Europe	Europe	N America	ME Asia	Caribbean	Europe	ME Asia	N America
	region 5	N America	L America	Europe	Oceania	Europe	Europe	SE Asia	Europe	East Asia	Europe
	region 6	Europe	Europe	Europe	Africa	ME Asia	L America	Caribbean	Europe	Europe	Oceania
	region 7	Europe	Africa	Africa	Europe	Europe	Europe	Africa	Africa	N America	ME Asia
	region 8	L America	ME Asia	Caribbean	Europe	Africa	SE Asia	Europe	Europe	Caribbean	L America
	region 9	N America	South Asia	Europe	Africa	ME Asia	Africa	Europe	Europe	East Asia	Europe
FLAVA	region 0	Caribbean	East Asia	East Asia	ME Asia	Caribbean	Africa	L America	Europe	Africa	Europe
	region 1	N America	Africa	SE Asia	Africa	Africa	SE Asia	Africa	Africa	Oceania	Europe
	region 2	Africa	Africa	South Asia	ME Asia	Africa	Europe	East Asia	ME Asia	South Asia	Oceania
	region 3	Europe	South Asia	Africa	N America	Caribbean	Africa	Africa	ME Asia	Africa	SE Asia
	region 4	Africa	Africa	Europe	Africa	SE Asia	Africa	Europe	Europe	Africa	ME Asia
	region 5	Oceania	Africa	Oceania	Africa	Caribbean	Oceania	Europe	Africa	Europe	Europe
	region 6	Africa	Africa	Oceania	South Asia	Europe	Europe	South Asia	ME Asia	Europe	SE Asia
	region 7	ME Asia	Africa	South Asia	Europe	SE Asia	Caribbean	N America	South Asia	SE Asia	Europe
	region 8	Europe	Africa	Africa	Europe	Europe	SE Asia	Europe	Oceania	Africa	Europe
	region 9	Africa	Africa	Caribbean	N America	Africa	Europe	N America	South Asia	Caribbean	Africa

Table 11: First half of the results for first 10 retrieved regions *Retrieval Across Universals* task. Note: N America, L America, ME Asia, and SE Asia stand for North America, Latin America, Middle East Asia, and South East Asia, respectively.

Model	First@10 Regions	greeting	headcoverings	instrument	lunch	marriage	music	religion	ritual	sports	transport
CLIP	region 0	SE Asia	Africa	Europe	SE Asia	Europe	Africa	SE Asia	SE Asia	SE Asia	Africa
	region 1	SE Asia	Africa	Europe	Europe	Europe	Europe	East Asia	South Asia	Europe	L America
	region 2	N America	Europe	Europe	SE Asia	L America	ME Asia	L America	Europe	N America	East Asia
	region 3	East Asia	East Asia	SE Asia	Europe	Europe	L America	Caribbean	SE Asia	Europe	L America
	region 4	SE Asia	Africa	L America	L America	Europe	Europe	East Asia	East Asia	Oceania	Africa
	region 5	SE Asia	ME Asia	L America	SE Asia	Oceania	Europe	Africa	Africa	East Asia	SE Asia
	region 6	SE Asia	Europe	ME Asia	Europe	N America	Europe	Africa	Africa	ME Asia	South Asia
	region 7	Africa	ME Asia	Europe	Europe	Europe	Africa	SE Asia	East Asia	Europe	Africa
	region 8	East Asia	ME Asia	Africa	East Asia	L America	L America	Africa	ME Asia	Europe	ME Asia
	region 9	Africa	Africa	South Asia	Africa	Europe	Europe	ME Asia	SE Asia	Africa	South Asia
OpenCLIP	region 0	SE Asia	N America	Africa	L America	South Asia	L America	ME Asia	Oceania	SE Asia	Africa
	region 1	SE Asia	Oceania	Europe	Europe	Europe	Africa	Africa	East Asia	South Asia	Africa
	region 2	L America	Africa	ME Asia	Europe	ME Asia	Europe	Oceania	South Asia	Africa	East Asia
	region 3	L America	Europe	Africa	SE Asia	SE Asia	L America	ME Asia	SE Asia	South Asia	ME Asia
	region 4	SE Asia	Africa	N America	SE Asia	L America	Oceania	South Asia	East Asia	Africa	Africa
	region 5	Africa	ME Asia	N America	N America	SE Asia	Africa	Africa	SE Asia	South Asia	Africa
	region 6	Africa	Europe	Europe	Europe	Africa	Africa	South Asia	L America	Europe	Africa
	region 7	Europe	ME Asia	Europe	East Asia	N America	ME Asia	South Asia	L America	SE Asia	L America
	region 8	SE Asia	Europe	Europe	Europe	N America	Europe	Africa	L America	L America	L America
	region 9	South Asia	Africa	East Asia	SE Asia	South Asia	Africa	SE Asia	SE Asia	ME Asia	Africa
CoCA	region 0	SE Asia	N America	Europe	L America	Africa	Europe	ME Asia	L America	SE Asia	SE Asia
	region 1	L America	Europe	Europe	L America	Europe	L America	Africa	L America	Caribbean	South Asia
	region 2	South Asia	Europe	L America	Caribbean	SE Asia	Europe	Caribbean	South Asia	East Asia	Africa
	region 3	SE Asia	Africa	Europe	Europe	Europe	Africa	Africa	South Asia	South Asia	ME Asia
	region 4	Africa	Europe	Europe	Europe	Europe	Europe	Europe	SE Asia	South Asia	L America
	region 5	Africa	L America	N America	N America	L America	Africa	Africa	East Asia	Europe	ME Asia
	region 6	East Asia	ME Asia	Europe	Europe	ME Asia	N America	Oceania	Africa	ME Asia	Africa
	region 7	Europe	ME Asia	East Asia	N America	Europe	Europe	L America	Oceania	Europe	Africa
	region 8	SE Asia	ME Asia	Africa	Europe	SE Asia	ME Asia	Africa	Africa	Europe	Europe
	region 9	Europe	SE Asia	Europe	N America	ME Asia	Europe	South Asia	SE Asia	South Asia	Europe
TCL	region 0	L America	Oceania	East Asia	East Asia	SE Asia	ME Asia	L America	Africa	SE Asia	South Asia
	region 1	Europe	ME Asia	Oceania	East Asia	South Asia	East Asia	Africa	Africa	SE Asia	ME Asia
	region 2	L America	South Asia	Europe	East Asia	South Asia	Africa	Africa	Oceania	Europe	ME Asia
	region 3	East Asia	SE Asia	Europe	Africa	South Asia	Europe	Africa	SE Asia	Africa	SE Asia
	region 4	Africa	Oceania	South Asia	Europe	SE Asia	South Asia	Africa	Africa	Africa	Caribbean
	region 5	East Asia	SE Asia	Europe	Africa	Africa	Africa	Africa	ME Asia	SE Asia	Oceania
	region 6	South Asia	SE Asia	Europe	N America	South Asia	N America	Africa	SE Asia	South Asia	Caribbean
	region 7	SE Asia	L America	Africa	Europe	South Asia	Europe	L America	Africa	South Asia	Africa
	region 8	Caribbean	Europe	Oceania	East Asia	SE Asia	South Asia	Europe	South Asia	Africa	South Asia
	region 9	L America	Africa	ME Asia	ME Asia	ME Asia	Europe	Africa	ME Asia	ME Asia	Oceania
ALBEF	region 0	L America	Oceania	East Asia	East Asia	SE Asia	ME Asia	L America	Africa	SE Asia	South Asia
	region 1	Europe	ME Asia	Oceania	East Asia	South Asia	East Asia	Africa	Africa	SE Asia	ME Asia
	region 2	L America	South Asia	Europe	East Asia	South Asia	Africa	Africa	Oceania	Europe	ME Asia
	region 3	East Asia	SE Asia	Europe	Africa	South Asia	Europe	Africa	SE Asia	Africa	SE Asia
	region 4	Africa	Oceania	South Asia	Europe	SE Asia	South Asia	Africa	Africa	Africa	Caribbean
	region 5	East Asia	SE Asia	Europe	Africa	Africa	Africa	Africa	ME Asia	SE Asia	Oceania
	region 6	South Asia	SE Asia	Europe	N America	South Asia	N America	Africa	SE Asia	South Asia	Caribbean
	region 7	SE Asia	L America	Africa	Europe	South Asia	Europe	L America	Africa	South Asia	Africa
	region 8	Caribbean	Europe	Oceania	East Asia	SE Asia	South Asia	Europe	South Asia	Africa	South Asia
	region 9	L America	Africa	ME Asia	ME Asia	ME Asia	Europe	Africa	ME Asia	ME Asia	Oceania
BLIP-2	region 0	East Asia	Africa	South Asia	Europe	ME Asia	SE Asia	Africa	SE Asia	SE Asia	Africa
	region 1	Oceania	Europe	ME Asia	ME Asia	SE Asia	East Asia	Europe	SE Asia	SE Asia	Africa
	region 2	Europe	East Asia	Africa	Europe	ME Asia	Europe	Africa	Europe	South Asia	ME Asia
	region 3	Africa	ME Asia	Africa	Europe	ME Asia	Oceania	Caribbean	L America	SE Asia	SE Asia
	region 4	ME Asia	Europe	East Asia	L America	ME Asia	SE Asia	Africa	Africa	South Asia	ME Asia
	region 5	East Asia	Africa	Europe	Oceania	ME Asia	Africa	Africa	South Asia	SE Asia	South Asia
	region 6	L America	Europe	South Asia	East Asia	Europe	South Asia	Europe	Africa	SE Asia	SE Asia
	region 7	ME Asia	Africa	SE Asia	ME Asia	Europe	Europe	L America	Europe	SE Asia	Africa
	region 8	Africa	ME Asia	ME Asia	Oceania	Oceania	Europe	Europe	South Asia	South Asia	Africa
	region 9	ME Asia	ME Asia	Europe	SE Asia	Europe	Europe	Africa	Europe	L America	Africa
FLAVA	region 0	SE Asia	SE Asia	Europe	L America	ME Asia	L America	SE Asia	South Asia	South Asia	Africa
	region 1	SE Asia	Europe	Europe	Caribbean	Europe	Europe	Africa	N America	Europe	East Asia
	region 2	Africa	East Asia	East Asia	Caribbean	Africa	East Asia	Caribbean	SE Asia	Africa	Africa
	region 3	N America	ME Asia	Africa	Africa	Africa	East Asia	N America	ME Asia	Africa	Africa
	region 4	N America	Africa	Africa	Europe	Europe	Europe	Africa	SE Asia	SE Asia	Africa
	region 5	Africa	L America	Africa	SE Asia	Africa	Europe	SE Asia	L America	SE Asia	ME Asia
	region 6	L America	Caribbean	L America	Oceania	Africa	South Asia	L America	East Asia	SE Asia	Oceania
	region 7	SE Asia	Africa	South Asia	Africa	Africa	ME Asia	Africa	East Asia	SE Asia	Europe
	region 8	Africa	Africa	L America	SE Asia	ME Asia	South Asia	Europe	Europe	ME Asia	ME Asia
	region 9	Europe	Africa	Africa	South Asia	Europe	L America	L America	South Asia	East Asia	South Asia

Table 12: Second half of the results for first 10 retrieved regions *Retrieval Across Universals* task. Note: N America, L America, ME Asia, and SE Asia stand for North America, Latin America, Middle East Asia, and South East Asia, respectively.

Country	Cultural Concepts
Argentina	alfajor, alpargatas, asado, bandoneon, bifes a la criolla, boina, bolero, bombilla, carbonada, chimichurri, chipa, chocotorta, choripan, churros rellenos, dulce de batata, dulce de leche, dulce de membrillo, empanada, facturas, faina, gaucho knife, humita, locro, lomito sandwich, malbec, matambre, mate, medialuna, milanesa, morcilla, parrilla, pascualina, pastel de papa, pebete, picada, provoleta, rabanito, ravioles, rosca de pascua, sandwich de miga, torta frita, vino patero, yerba
Brazil	acai, acaraje, alfajor, baiao, bombacha, bumba-meu-boi, brigadeiro, cachaca, caipirinha, carimbo, chimarrao, churrasco, cocar, cuica, empada, espetinho, farofa, feijoada, frescobol, moqueca, pacoca, pao de queijo, rapadura, requeijao, rosca, romeu e julieta, samba, sarongue, tapioca, tucupi, vatapa
Canada	bagel, bannock, beavertail pastry, blueberry grunt, butter tart, caribou, cipaille, caesar cocktail, cretons, date squares, donair, flipper pie, garlic fingers, inukshuk, jiggs dinner, maple taffy, nanaimo bar, peameal bacon, pemmican, persian roll, poutine, rappie pie, sugar pie, toboggan, toque, tourtiere
China	baozi, bianlian, bianzhong, biang biang noodles, chinese knot, chinese lantern, chinese seal, cong you bing, doufu, dragon beard candy, erhu, fenghuang crown, gongbi, guzheng, hongbao, hotpot, huanghuali furniture, hulusi, jiaozi, jinghu, laziji, liuli, longjing tea, luo han guo, mala tang, mahjong tiles, mooncake, paper cutting, peking opera mask, pipa, qipao, shengjianbao, suzhou embroidery, wushu sword, xiao long bao, xun, yuanyang hotpot, zongzi
India	aarti thali, achaar, bangles, bhang, bhatura, bharatanatyam, bindi, biryani, chapati, chai, diya, dosa, dhoti, gajra, ganesha, idli, jalebi, jhumka, kathakali, kulfi, kurta, kumkum, laddu, lassi, lehenga, lungi, mangalsutra, mehndi, mojaris, mridangam, murukku, namaste, pani puri, papadum, paratha, payal, rasam, rasgulla, rangoli, raita, salwar kameez, sari, shehnai, sherwani, sitar, tabla, tanpura, tandoor, tikka, turban, veena, vada
Israel	baba ganoush, baklava, bourekas, challah, chamsa, chuppah, eshet chayil candlesticks, fattoush, falafel, galabeya, hamentashen, halva, hatzilim, hamsa, jachnun, kibbeh, kippah, krembo, ketubah, knafeh, kubbeh, kiddush cup, knafeh, labaneh, malabi, matbucha, matzah, menorah, muhammara, matkot, ptitim, rugelach, sabich, sambusak, sefer torah, shakshuka, shofar, skhug, stuffed grape leaves, sufganiyah, tallit, tefillin, tembel hat, tabbouleh, tzatziki, tzitzit, yemenite kudu horn
Mexico	aguas frescas, alebrije, banderita, barbacoa, calavera, cantarito, carnitas, cemitas, ceviche, chalupa, chapulines, chicharrones, churro, cochinita pibil, enchilada, gordita, huarache, huipil, menudo, metate, mole, molinillo, nopal, ofrenda, panucho, papel picado, pinata, pozole, pulque, quesadilla, quexquemitl, rebozo, salbute, sarape, sopes, taco, talavera, tamale, teponaztli, tlayuda, torta, vihuela, zarape
Nigeria	abacha, abeti aja, agbada, agidi, akara, amala, aso oke, asoke, buba, chin chin, danfo, dodo, edikang ikong, egusi soup, ekwe, ewedu, fila, fufu, gbegiri, gele, garri, isi ewu, jollof rice, keke napep, kilishi, kuli kuli, moi moi, ogene, okapi, oha soup, pounded yam, sakara, suya, talking drum, zobo
Pakistan	achaar, ajrak, balochi sajji, banarasi saree, balti, biryani, chapli kabab, chitrali cap, cobalt pottery, dholki, falooda, gulab jamun, gilgit cap, haleem, henna, hunza cap, karahi, kheer, khadi, khussa, kulfi, lacha paratha, lehnga choli, miswak, moti choor ladoo, multani sohan halwa, nan khatai, nihari, paan, pakol, pathani suit, peshawari chappal, saag, samosa, sharbat, sheermal, shalwar kameez, sindhi topi.
Philippines	adobo, anting-anting, arnis sticks, bahay kubo, balangay, balisong, balut, bangus, barong tagalog, bayong, bulul, calamansi, carabao, dinuguan, durian, guling, halo-halo, ifugao hut, jeepney, kalesa, kinilaw, kulintang, lechon, malong, maranao gong, pamaypay, pan de regla, pandesal, palabok, pinya fabric, puto bumbong, salakot, santol, sinigang, singkaban, tarsier, tapsilog, terno, tinikling.
Poland	barszcz, basolia, bigos, bryndza, chrzan, flaki, faworki, golonka, kasza gryczana, kaszanka, kabanos, kartacze, kielbasa, kiszka, knysza, kogel mogel, kompot, kotlet schabowy, kluski slaskie, makowiec, mizeria, oscypek, pasztecik szczecinski, paczek, pierogi, pierniki, placki ziemniaczane, ptasie mleczko, rosol, rogalswietomarcinski, ryba po grecku, sledz, smalec, ser bialy, sekacz, szarlotka, tatar, zrazy, zurek.
Russia	babushka, balalaika, bayan, blini, borshch, budyonovka, caviar, chak-chak, domra, dymkovo toys, fabergï¿½ eggs, garmon, gusli, gzhel, khokhloma, kasha, kokoshnik, kvass, lapti, matryoshka, okroshka, pelmeni, podstakannik, pryanik, pirozhki, russian blue, samovar, sarafan, shchi, soljanka, sushki, syrniki, telnyashka, treshchotka, ushanka, valenki, varenyky.
South Africa	amarula, amagwinya, beadwork, biltong, boeremusiek instruments, boerewors, braai, bunny chow, chakalaka, djembe, dompas, fufu, geelbek, hadeda ibis, kaross, knobkerrie, makarapa, malva pudding, marula fruit, melktert, mopane worms, pap, potjiekos, protea, rooibos, rondavel, shweshwe, sosatie, spaza shop, txalaparta, umqombothi, vetkoek, vuvuzela.
South Korea	bibimbap, bokjumeoni, bossam, bulgogi, buchaechum, dduk, ddukbokki, dongchimi, galbi, gat, gayageum, geomungo, gochujang, gimbap, haeguem, hahoetal, hanbok, hangwa, hanji, hwagwan, jeogori, jeon, jokduri, janggu, jeotgal, kimchi, makgeolli, naengmyeon, norigae, pyeongyeong, samgyeopsal, samulnori, seonji, sikhye, sotdae, sundubu-jjigae, soju, tteok, tteokguk, tteokbokki, yeot.
Vietnam	ao ba ba, ao dai, ao thu than, banh bao, banh canh, banh chung, banh cuon, banh gio, banh giay, banh khuc, banh mi, banh pia, banh xeo, ca phe trung, cao lau, cafe sua da, canh chua, chao long, che, dua mon, gio lua, gio lua, goi cuon, hoanh thanh, kem xoi, keo dua, khanh ran, my quang, non la, nuoc mam, pho, sinh to, thit kho tau, thit heo quay, trong com, trung vit lon, thung chai boat, bun cha, bun bo hue, bun thit nuong, com tam.

Table 13: List of cultures concepts covered in Cultural Visual Grounding dataset