Culture Affordance Atlas: Reconciling Object Diversity Through Functional Mapping

Joan Nwatu [1], **Longju Bai** [1], **Oana Ignat** [2], **Rada Mihalcea** [1]
[1]University of Michigan-Ann Arbor
[2]Santa Clara University
{jnwatu, longju, mihalcea} @umich.edu, oignat@scu.edu

Abstract

Culture shapes the objects people use and for what purposes, yet mainstream Vision-Language (VL) datasets frequently exhibit cultural biases, disproportionately favoring higher-income, Western contexts. This imbalance reduces model generalizability and perpetuates performance disparities, especially impacting lower-income and non-Western communities. To address these disparities, we propose a novel function-centric framework that categorizes objects by the functions they fulfill, across diverse cultural and economic contexts. We implement this framework by creating the Culture Affordance Atlas, a re-annotated and culturally grounded restructuring of the Dollar Street dataset spanning 46 functions and 288 objects. Through extensive empirical analyses using the CLIP model, we demonstrate that function-centric labels substantially reduce socioeconomic performance gaps between high- and low-income groups by a median of 6 pp (statistically significant), improving model effectiveness for lower-income contexts. Furthermore, our analyses reveals numerous culturally essential objects that are frequently overlooked in prominent VL datasets. Our contributions offer a scalable pathway toward building inclusive VL datasets and equitable AI systems.

Supplemental — https://github.com/MichiganNLP/Culture-Afford-Analysis

Datasets — https://lit.eecs.umich.edu/Culture-Affordance-Atlas/index.html

1 Introduction

Culture, as reflected in data, represents people's way of life (Rosling, Rosling, and Rönnlund 2018; Griswold 2012; Tylor 1871). The geographic and socioeconomic dimensions of culture profoundly influence the objects people use daily – objects that vision-language (VL) models should accurately recognize and interpret. However, popular VL datasets suffer from persistent representational imbalances across these dimensions, despite increasing awareness of these issues (Longpre et al. 2024; Nwatu, Ignat, and Mihalcea 2023; Paullada et al. 2021). Most image datasets (Kuznetsova et al. 2020; Deng et al. 2009; Schuhmann et al. 2022) are object-centric and skewed toward objects common in higher-income, Western contexts. As a result, culturally distinct objects or object functions from lower-income or non-Western regions are often missing or misclassified, leading to poor model generalization.

Figure 1: Various objects are identified to perform the same function across different cultures and income groups. *Best viewed in color.*

Efforts to broaden coverage such as GeoDE (Ramaswamy et al. 2023), Dollar Street (Gaviria Rojas et al. 2022), and Segment Anything (Kirillov et al. 2023), have so far made important strides, but substantial gaps remain. Large-scale crawls still filter out underrepresented data (Fang et al. 2023; Nwatu, Ignat, and Mihalcea 2025), and existing datasets often collapse culturally distinct artifacts under a single Western label (e.g., grouping 'clay pot', 'cooler box', and 'refrigerator' under the label "refrigerator" in Dollar Street).

We argue that object-centric methods alone are insufficient to address these biases. We, instead, propose a function-centric lens. Across all human societies, a core set of universal activities: sleeping, cooking, cleaning, and storing are fulfilled by culturally specific objects (Brown 2004). As illustrated in Figure 1, the function of 'sleeping' may be served

by a Western bed, a Nepali charpai, or a Malawian straw mat; 'cleaning teeth' may involve a plastic toothbrush, a chewing stick, or even fingers. By re-annotating existing image collections based on such cultural affordances, we align diverse artifacts under a shared semantic framework and reveal long-tail objects overlooked in current VL taxonomies.

This paper makes the following contributions.

- We introduce a novel **function-centric framework** for constructing culturally diverse VL datasets, organizing objects by their universal human functions alongside their contextual usage.
- We apply this framework to create the **Culture Affordance Atlas**, a public knowledge base derived by re-annotating the Dollar Street dataset into culturally grounded function categories, with 367 object-function pairs, each backed by at least one ethnographic citation.
- We empirically demonstrate that **function-centric labeling significantly reduces disparities in VL model performance** across socioeconomic groups, highlighting practical benefits for low-income data representation.
- Our function-centric approach **uncovers numerous culturally relevant objects** frequently absent from major VL datasets underscoring its potential to bridge representational gaps.

Across these contributions, we demonstrate a scalable, culturally-aware path toward more inclusive VL systems. By asking *' forwhat is this object used for?* rather than *'what is this object?'*, we leverage a new perspective to annotate object diversity and highlight poorly represented objects in VL research.

2 Related Work

2.1 Representation Bias in Vision-Language Datasets

Multiple studies (Paullada et al. 2021; Shankar et al. 2017; Longpre et al. 2024; Ignat et al. 2024; Liu et al. 2024) show that vision-language datasets such as LAION-5B (Schuhmann et al. 2022), YFCC100M (Thomee et al. 2016), COCO (Lin et al. 2014), and ImageNet (Deng et al. 2009) contain disproportionately large amounts of data from North America and Europe, with far less representation from Africa, South Asia, and parts of Latin America. Although large datasets improve model performance, their construction often depends on institutions with substantial resources. To reduce costs, developers frequently rely on web scraping, which mirrors the distribution and priorities of the internet and tends to elevate Western high-income content (Birhane, Prabhu, and Kahembwe 2021; Paullada et al. 2021; Crawford and Paglen 2021).

Most VL datasets also use object-centric annotations based on Western taxonomies derived from ImageNet (Mihalcea et al. 2025). This structure supports label consistency but overlooks cultural context, functionality, and alternative uses of objects (Nwatu, Ignat, and Mihalcea 2025, 2023). Recent datasets like WIT (Srinivasan et al. 2021) attempt to enrich grounding through natural language captions, yet caption-based collections still inherit platform-specific biases.

Our approach addresses these limitations by introducing structured, function-based groupings that link objects across cultures and improve representation of underrepresented household artifacts.

2.2 Affordances, Functionality, and the Cultural Framing of Objects

Objects are often classified by perceptual properties such as color, size, shape, texture, and material, features commonly used in deep learning for detection and recognition (Redmon and Farhadi 2018; Russakovsky et al. 2015). Humans, however, frequently understand objects through their *affordances*, the actions an object enables relative to a person's capabilities (Gibson 2014). This perspective highlights meaningful use rather than appearance. Anthropological work, including Brown's Human Universals, identifies recurring domains of activity such as cooking, cleaning, hygiene, and travel (Brown 2004). These domains guide our definition of functional categories that can span visual and cultural variation.

Affordance perception is also shaped by social and cultural context (Costanza-Chock 2020; Maier and Fadel 2009; Ye, Cardwell, and Mark 2009). A metal basin, for instance, may afford bathing in one setting and food storage in another. In Human-Computer Interaction, overlooking such cultural variation often leads to designs that fail to engage users (Norman 1999; Fayard and Weeks 2014).

Although affordance-based models in vision and robotics have improved generalization, especially for object manipulation (Hassanin, Khan, and Tahtali 2021; Mur-Labadia, Guerrero, and Martinez-Cantin 2023; Castellini et al. 2011; Do, Nguyen, and Reid 2018; Bohg et al. 2013), they rarely consider cultural differences in object use. Our work extends affordance theory by centering cultural and functional dimensions rather than solely sensory attributes. We propose a function-based categorization framework that connects visually diverse objects performing similar daily functions across socioeconomic and geographic contexts. This approach strengthens cultural representation and reveals objects often overlooked in existing datasets.

2.3 Improving Cultural Representation in Vision Language Models

Efforts to improve representation in vision-language models have centered on expanding dataset diversity, adapting training methods, and refining prompts to mitigate imbalances across language, region, and socioeconomic background.

A major approach involves building datasets that better reflect underrepresented contexts. Notable examples include Segment Anything (Kirillov et al. 2023), which offers large-scale object masks for generalization, GeoDE (Ramaswamy et al. 2023), which evaluates model performance across geographic regions, and Dollar Street (Gaviria Rojas et al. 2022), which documents household items from a wide range of income levels. Although these datasets broaden coverage, they remain limited by structural issues. Many use data collection or annotation schemes shaped by Western-centric taxonomies, contain mislabeling, or lack sufficient metadata to reveal cultural distinctions. As a result, culturally important objects are often collapsed into generic categories.

Complementary work investigates prompt engineering and language adaptation to improve performance in low-resource contexts. Prior studies show that adding contextual information to prompts can increase retrieval accuracy for underrepresented data (Nwatu, Ignat, and Mihalcea 2025; Buettner et al. 2024; Nguyen et al. 2023). Building on this direction, we re-annotate Dollar Street using a function-centric framework. We categorize images based on the function the depicted object fulfills across cultures, assign accurate names to culturally specific items, correct prior mislabeling, and surface marginalized artifacts. These function-based groupings are then integrated into dataset labels and prompt design to address poor performance on low-income and non-Western images.

3 Methodology

To capture the rich diversity of artifacts serving similar functions across cultures, we propose a function-to-object labeling framework. We start with an existing dataset of cultural objects (DollarStreet) and leverage a state-of-the-art VL model (CLIP) to construct the **Culture Affordance Atlas**, a publicly accessible knowledge base documenting function-to-object mappings across diverse cultural contexts.

3.1 Functions

We define *functions* as the affordances of an object, that is, the specific actions or activities it enables. We explore the culturally and socially meaningful roles that objects serve across different communities, transcending their mere physical appearances. Drawing inspiration from Gibson's theory of affordances (Gibson 2014), and Norman's research in human-computer interaction (Norman 1999), we argue that incorporating functional descriptions into object labels can improve cultural representation in AI datasets and models.

3.2 State-of-the-art Vision-Language Model

We use CLIP (Radford et al. 2021) to conduct our experiments to evaluate the effectiveness of the new labels and captions on lower-income images. CLIP is a VL model trained on a large corpus of image–text pairs, enabling it to jointly embed visual and textual inputs in a shared semantic space. CLIP's open availability, popularity, and relevance to zero-shot image-text retrieval and other downstream tasks, as demonstrated Hessel et al. (2021) make it a suitable tool for assessing the quality and generalizability of image annotations across diverse socioeconomic contexts.

3.3 Dollar Street Dataset

We use the Dollar Street dataset for our re-annotation and experiments because it is geographically diverse and links household objects to reported income levels. The dataset contains 38,479 images from 63 countries across Africa, the Americas, Asia, and Europe, covering everyday objects and activities such as toothbrushes, toilet paper, and cooking. These are organized into 291 topics. Following (De Vries et al. 2019), we remove 21 subjective topics, yielding 270 objective topics for analysis.

Dollar Street also spans a wide socioeconomic range, with household incomes from $26.9 to 19{,}671.0$ per month, adjusted for purchasing power parity. Following (Gaviria Rojas et al. 2022), we group incomes into geometric ranges and quartiles to support balanced comparisons. Image contributions vary across countries, from 45 images in Canada to 4,704 in India, with a median of 407 per country.

3.4 Objects to Functions Re-annotation

The Dollar Street dataset includes images as well as a topic label and an income value per each image. We re-annotate the Dollar Street dataset to include *functional descriptions* of the topic labels (e.g., 'bed': 'object used for sleeping') and an *identified object label* which is the standard name of the object found in a given image (e.g., an image of a straw mat with the topic label 'bed', will now include a functional description 'object used for sleeping' and an identified object label 'straw mat') as depicted in Figure 2.

Functional Descriptions. We generate functional descriptions for all 270 unique topic labels in Dollar Street using GPT-4o. We use the following prompt; *"A bed is an object that provides a place to sleep. Following the above example, In one short sentence, state what function the object 'topic label' does."*. We then remove the topic label mentioned in the description to get *'object that provides a place to sleep'*.

Identified Objects. To identify which object present in the image is being referred to by the topic label, we tested various pipelines and continued with the pipeline that produced better outcomes. We use LLaVa to caption the images *(prompt- "USER: [image] What does this image show? ASSISTANT:")* and then use GPT-4 to extract the name(s) of the objects present in the image caption that fulfills the associated functional description for the image with the prompt: *'You are given two sentences. One: function Two: description Mention the name of the object or objects referred to in both sentences'* where sentence one is the function and sentence two is the image caption.

Quality Control. We randomly check for incorrect entries to rewrite and manually reannotated 1458 identified object names including those that were previously listed as *NaN*, *None* or *Not identifiable*. For the functional descriptions, we run a quality assessment check with 21 participants across 7 countries (China, India, Ethiopia, Nigeria, United States, Romania, and Russia) selected due to their representation of the 4 continents present in Dollar Street and availability of annotators. For each of these countries, 30 questions are formulated using randomly selected images[1] and their respective functional descriptions. For each country, three native participants evaluate whether the provided functional description accurately describes the image, responding with a binary (yes/no) judgment.

Using responses from the assessment, we calculate the percentage validation for each country and the overall average. We report an overall score of 90% and show in that all scores

[1]All image selections for the user study were drawn by shuffling the dataset using pandas.DataFrame.sample with a fixed random_state (set to 42) to ensure reproducibility.

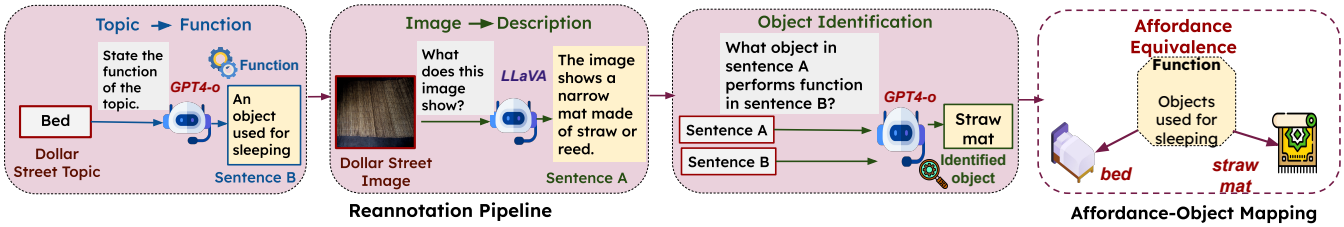

Figure 2: Re-annotation pipeline for generating functional description and identified object from original Dollar Street Image and Topic. Our Pipeline surfaces cross-cultural affordance equivalences by mapping both image-derived objects and topic labels to a shared functional space. *Best viewed in color.*

across the seven countries are above 85% (Table 1). This indicates a high correspondence between the Dollar Street images and the generated functional descriptions. We also include computed multiple inter-annotator agreement metrics in the supplemental materials.

4 Culture Affordance Atlas.

Among the various functions humans perform (e.g., writing, cooking, creating art), we prioritize those universally documented by Brown (2004) in *Human Universals* which is an anthropological theory that examines shared human characteristics and behaviors across cultures. We initially select seven human universal categories: hygienic care, aesthetics, cooking, diurnality, environmental adjustments, healing the sick, and visiting (details in the supplemental materials). We map functional descriptions and associated object names identified from the Dollar Street re-annotation pipeline (Figure 2) to functions within these overarching categories. For each primary category (e.g., cooking), we assign relevant functions and document the corresponding objects identified within the Dollar Street dataset. To ensure rigorous grounding, each entry is supported by at least one published, verifiable source. Approximately 98% of references originate from ethnographic publications accessed through the eHRAF World Cultures database, and Google scholar for the remaining 2%. The initial edition of the Culture Affordance Atlas comprises 367 entries, representative images, and metadata from Dollar Street, all publicly accessible (see supplemental).

General Statistics. Culture Affordance Atlas currently includes 367 entries spanning 7 categories, 46 functions, and 288 unique objects. Each function-object pair (e.g., teeth cleaning-toothbrush) includes a frequency count derived from contextual image annotations (Table 2). Summary statistics and representative examples across the categories of the Atlas are provided in Table 2.

Country	% validated
China	87.78
India	90.00
Ethiopia	91.11
Nigeria	86.67
United States	87.78
Romania	96.67
Russia	90.00
Average	**90.00**

Table 1: Quality control for generated functional descriptions

Culture Affordance Category	#Func	#Obj	Function example	Object example
Practicing hygiene	19	96	cleaning teeth	toothbrush
Beautifying	3	40	adorning the body	necklace
Cooking	7	60	providing heat for cooking	charcoal stove
Maintaining a diurnal cycle	3	16	sleeping	mat
Adjusting to the environment	11	85	providing light	oil lamp
Healing	2	16	providing medication	inhaler
Traveling	1	10	transporting	bicycle

Table 2: Statistics for the Culture Affordance Atlas.

Label Misalignment and Object Diversity within Functions. Through our re-annotation pipeline, we identify discrepancies between original Dollar Street topics and actual object use. A notable percentage of images display mismatches between original topic labels and identified objects, averaging **38.25%** misalignment across topics.[2] Functions such as "bathing" and "clothes washing" demonstrate significant object diversity, characterized by visually disparate artifacts.

Object Use Across Functions. Many objects fulfill multiple functions. For example, "bowls" appear in contexts including "clothes washing", "hand washing", "food serving", and "food preparation". Object-function pairs demonstrating alternative uses, differing from an object's primary function, highlight the versatility and improvisational use of everyday items. "Charcoal" exemplifies this phenomenon: primarily used as "fuel", it also serves "dental hygiene" functions (Figure 3). Such objects, termed *contextually niche*, have dominant uses alongside lesser-known but culturally significant secondary functions. The Culture Affordance Atlas enhances visibility for these secondary functions by documenting relevant contexts and ethnological references.

Long-Tail Objects. Vision-language research often focuses on common object categories, leaving niche or cul-

[2]We prompt GPT4o to identify mismatches by comparing Dollar Street Topics with the identified objects from our pipeline.

Dollar Street Topic: Toothpaste
Generated Function: Cleaning teeth
Identified Object: Charcoal

Image Context (Country): Malawi
Income: $84/month

eHRAF Reference: (Hockings 1980)

Dollar Street Topic: Stove/hob
Generated Function: Providing heat for cooking
Identified Object: Charcoal

Image Context (Country): Togo
Income: $321/month

eHRAF Reference: (Pierce 1964)

Figure 3: Charcoal use across functions. *Best viewed in color.*

turally specific items underrepresented and contributing to a long-tail distribution. Accounting for these overlooked objects is important for building more inclusive and robust models.

We identify long-tail objects in the Culture Affordance Atlas that appear infrequently in our re-annotation. Using the 100 least frequent items (e.g., calabash, earthen pot), we examine their presence across seven widely used VL datasets selected for their broad object coverage and prominence: ImageNet (Deng et al. 2009), COCO (Lin et al. 2014), Open Images (Kuznetsova et al. 2020), LVIS (Gupta, Dollar, and Girshick 2019), GeoDE (Ramaswamy et al. 2023), YFCC100M Entity (Artifacts) (Li et al. 2017), and Dollar Street (Gaviria Rojas et al. 2022). To ensure consistent comparison, we normalize labels using lowercasing, lemmatization, and fuzzy matching with a threshold of 0.8.

Our results in Table 3 show that only OpenImages, with its large label set (about 20k classes), covers more than half of these rare objects. We also identify 34 objects that are absent from all datasets, and 33 that appear in only one. These gaps illustrate the limited representation of culturally diverse and niche artifacts in current VL resources and point to the need for more comprehensive documentation. A full breakdown of the 100 long-tail objects and their coverage across datasets is provided in the supplemental materials.

Dataset	Total classes	Long-tail covered	Missing	Coverage (%)
ImageNet	1000	8	92	8%
COCO	80	2	98	2%
OpenImages	19868	61	39	61%
LVIS	1230	28	72	28%
GeoDE	40	0	100	0%
YFCC100M Entity-Artifacts	325	1	99	1%
Dollar Street	291	5	95	5%

Table 3: Long-tail Objects. Percentage of our 100 long-tail objects are present in each popular vision-language dataset's official class list.

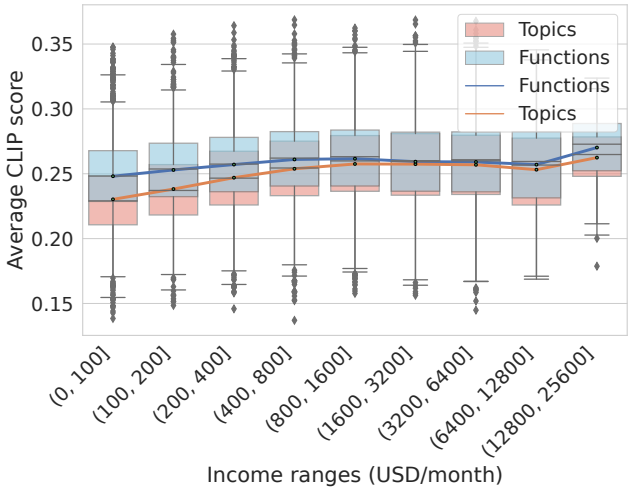

Figure 4: Comparison of CLIP alignment scores between Topic-Image (red) and Function-Image (blue) pairs across Dollar Street images from varying income levels. Trend lines indicate mean scores per income bin. The slope of each line reflects the extent of the digital divide. *Best viewed in color.*

5 Experiments

5.1 The Effects of Function Captions on VL models.

We investigate the impact of function-based captions on VL model performance through two distinct experiments comparing original Dollar Street topic labels against corresponding function descriptions generated via our re-annotation pipeline (Figure 2). In the first experiment, termed the *CLIP association test*, we compute cosine similarities between Dollar Street image embeddings and text embeddings of both topics (baseline) and functions. Images are grouped according to geometric income bins as described in (Gaviria Rojas et al. 2022). In the second experiment, we evaluate image retrieval performance. Specifically, we calculate CLIP scores between all images and each Dollar Street topic to derive topic-image associations, repeating the process using corresponding function captions to obtain function-image associations. We then measure retrieval performance using Recall: we retrieve the top N images based on CLIP scores—where N equals the number of ground-truth images—and compare these retrieved images against the ground-truth (i.e., number of true predictions / N). Below, we present key findings from these experiments.

Effects of Functions on Income Performance Gaps. Figure 4 compares CLIP alignment scores for images grouped by income, using functions (blue) and topics (red). Trend lines depict the average CLIP scores per income bin. Consistent with findings from (Nwatu, Ignat, and Mihalcea 2023), CLIP scores generally increase with income, indicating a performance gap favoring higher-income contexts. However, the function-based scores produce a flatter trajectory compared to topics. Linear regression confirms this visual observation, with slopes of 0.002 for functions versus 0.004 for topics,

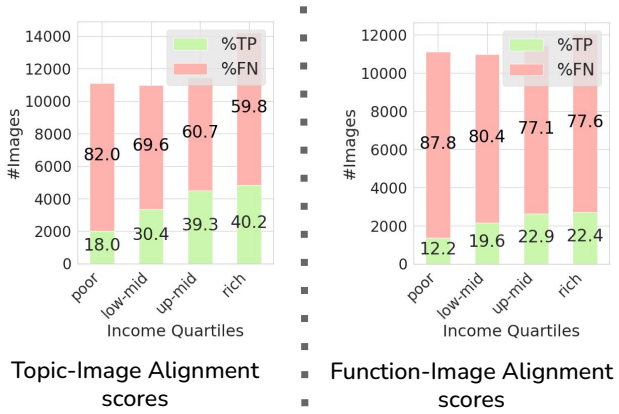

Figure 5: CLIP Recall across all images using Topic-Image (left) and Function-Image (right) alignment scores. We report the percentage of true positives and false negatives for each income quartile. Function-Image Recall shows less variation across income levels compared to Topic-Image Recall, indicating greater robustness to income-based distribution shifts. *Best viewed in color.*

indicating that function-based captions significantly reduce performance disparities.

For each of the 270 topics, we computed the difference in recall between high (rich and up-mid) and low-income (poor and low-mid) image sets under topic-only and function-based prompts, then took the per-topic gap reduction (Δgap) [3]. A Wilcoxon signed-rank test (Woolson 2005) shows that (Δgap) is significantly greater than zero (median = 0.06, p = 1.62e-17), confirming that function-based prompting reliably narrows socioeconomic performance gaps.

Figure 5 illustrates Recall results across income quartiles (poor, low-mid, upper-mid, and rich). Echoing the trends observed previously, functions yield less pronounced Recall disparities across income groups. Nevertheless, overall Recall scores for functions are lower compared to topics, suggesting trade-offs when employing functional descriptions in retrieval tasks. This phenomenon is explored further in the next subsection.

Tradeoffs and Challenges in Functions for VL tasks. Functions prompt CLIP to broaden retrieval and include culturally diverse, non-standard objects that topic prompts often miss. As shown in Figure 6, various forms of "stoves", "couches", and "water sources" from across cultures are correctly retrieved when framed by their function rather than their topic.

However, while function-based labels help reduce digital divides and equalize performance across socioeconomic contexts, they also decrease retrieval accuracy. We explore this trade-off qualitatively, by examining false positives produced by function prompts. We find that misinterpretation of functions by the model can lead to the retrieval of contextually incorrect objects, for example, the tendency of CLIP to emphasize nouns independently rather than interpreting

[3]Per topic, gap in Topic prompt - gap in Function Prompt

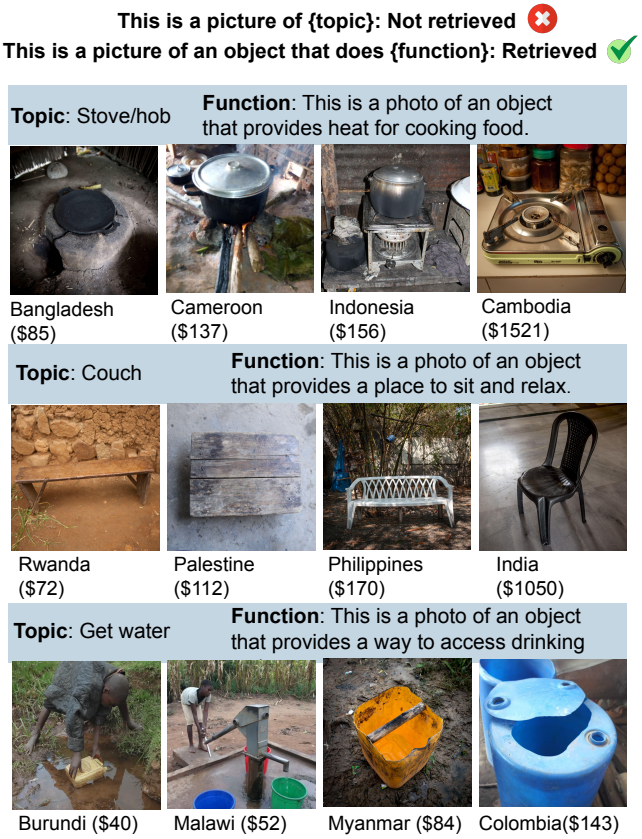

Figure 6: Qualitative analysis of Images 'forgotten' during retrieval with *topic prompt* but successfully retrieved with the *function prompt Best viewed in color.*

full sentence contexts (Castro, Ignat, and Mihalcea 2023). In Figure 7, the function prompts retrieve dishware and cultery instead of "objects that clean dishes", and retrieve images related to general cleaning of the toilets but miss the intended context of personal hygiene in the prompt. Effective use of functions thus requires careful crafting of prompts that are precise yet sufficiently inclusive to capture reasonable diversity (e.g., preferring *object that cleans dishes* over *machine that cleans dishes*).

Integrating Topics and Functions. Motivated by these insights, we explore integrating topics with their respective functions into combined prompts. For example, the topic *dishwasher* and its function *an object that cleans dishes and utensils* merge into the prompt *A dishwasher that cleans dishes and utensils*.

Repeating our experiments with these combined function-topic prompts, we compare performance against standalone topics (see details in the supplemental materials). Similar to function prompts, function-topic prompts eliminate performance disparities across income levels, achieving a near-zero slope (-0.0002) in CLIP association tests, with particularly notable performance gains for lower-income bins. Consequently, recall scores improve markedly for lower income

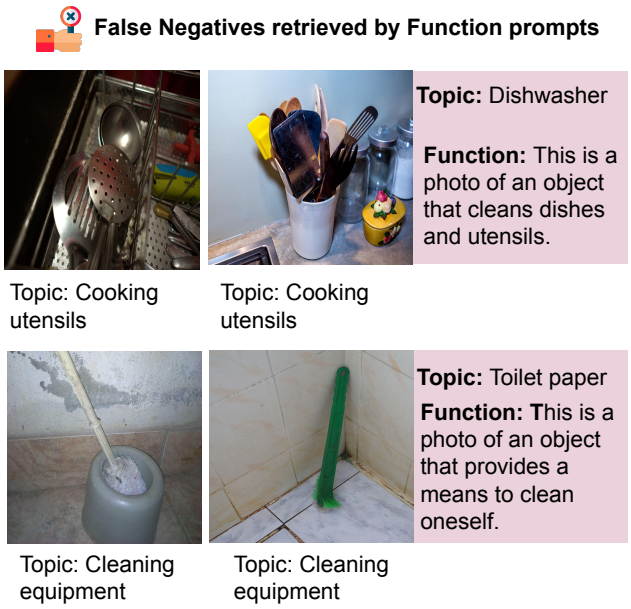

Figure 7: Qualitative analysis of false positives retrieved by *function prompts. Best viewed in color.*

quartiles, although slight reductions occur in higher income quartiles.

6 Generalizability

We confirm our findings on the reduction of performance disparity across income across vision-language models by repeating our experiments on the siglip2-so400m-patch14-384 model (Tschannen et al. 2025). Consistent with our findings on CLIP, we find that function prompts significantly reduce the high–low income performance gap by 11% ($p = 5.9e - 11$, Wilcoxon signed-rank test). Plots for the siglip2-so400m-patch14-384 model alignment scores and recall can be found in the supplemental materials.

7 Lessons Learned

We highlight key insights learned from our findings and present them below.

Cultural and socioeconomic biases persist in mainstream VL datasets. Our analysis reveals that 34% of our curated long-tail objects are entirely absent from seven widely used VL datasets, while 38.25% of Dollar Street topics suffer from label misalignment. These findings expose a critical blind spot in dataset construction: culturally specific artifacts are routinely omitted or mislabeled. Left unaddressed, such gaps risk entrenching cultural erasure, amplifying systemic bias, and further marginalizing underrepresented communities in AI applications.

Function-centric labeling effectively reduces the socioeconomic performance gap. We find that function-centric labeling reduces the recall gap between high and low-income image sets by a median of 6 percentage points in CLIP and 11 percentage points in SigLIP2 ($p < 5.9e - 11$). These results demonstrate that focusing on what objects afford, rather than how they are named, leads to more equitable model performance across income levels. Notably, designing precise, yet inclusive function prompts is crucial, as overly broad prompts reduce retrieval accuracy, while overly specific prompts fail to capture object diversity.

Systematic documentation is necessary for improving object representation in VL datasets. Following the recommendation from Nwatu, Ignat, and Mihalcea (2023) to annotate diversity and subjectivity in datasets, we present the Culture Affordance Atlas as an effort to reconcile object appearance diversity within labels and inclusively categorize them. We call for more interdisciplinary research efforts involving the AI community and domain experts toward representative data annotation.

8 Conclusion

In this paper, we demonstrated that culturally informed, function-based labeling substantially reduces representational disparities in vision-language models. By introducing the Culture Affordance Atlas, we systematically captured culturally diverse object-function relationships, addressing critical gaps in mainstream VL datasets. Our empirical analysis using CLIP showed that integrating object functions with traditional labeling significantly improves performance equity across socioeconomic contexts. This work demonstrates that adopting culturally aware frameworks such as function-centric annotation in dataset construction facilitates the development of inclusive AI systems capable of reliably serving diverse global communities.

9 Limitations

Culture Affordance Atlas Object Documentation: Each object–function pair in the Atlas is supported by at least one verifiable published source confirming its real-world use. Due to resource constraints, we did not exhaustively document all possible citations per culture. Most entries (89%) include one credible reference, the first reliably identified one, while 11% contain multiple sources. We position this Atlas as a foundational layer of cultural grounding that future work can expand through additional eHRAF searches, museum and archival materials, and community contributions. Current references should be viewed as initial guides for deeper inquiry.

Long-Tail Objects: The 34 long-tail objects represent items that did not match any class in ImageNet, COCO, OpenImages, LVIS, GeoDE, YFCC100M Entity (Artifacts), or Dollar Street after normalization and fuzzy matching (threshold 80). Because synonym variation may occur, we provide a 'Notes' column in the supplemental table to flag potential ambiguities.

Validation Scope: Our validation involved 21 participants across seven countries, providing geographically diverse and cross-cultural assessment. Nonetheless, broader participation would further strengthen the robustness of the validation.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback. We are also grateful to the Language and Information Technologies (LIT) lab members at the University of Michigan for their insightful discussions and feedback during the project's early stages. This project was partially funded by an award from OpenAI. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Open AI.

References

Birhane, A.; Prabhu, V. U.; and Kahembwe, E. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*.

Bohg, J.; Morales, A.; Asfour, T.; and Kragic, D. 2013. Data-driven grasp synthesis—a survey. *IEEE Transactions on robotics*, 30(2): 289–309.

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

Buettner, K.; Malakouti, S.; Li, X. L.; and Kovashka, A. 2024. Incorporating Geo-Diverse Knowledge into Prompting for Increased Geographical Robustness in Object Recognition. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13515–13524. IEEE.

Castellini, C.; Tommasi, T.; Noceti, N.; Odone, F.; and Caputo, B. 2011. Using object affordances to improve object recognition. *IEEE transactions on autonomous mental development*, 3(3): 207–215.

Castro, S.; Ignat, O.; and Mihalcea, R. 2023. Scalable Performance Analysis for Vision-Language Models. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (* SEM 2023)*, 284–294.

Costanza-Chock, S. 2020. *Design justice: Community-led practices to build the worlds we need*. The MIT Press.

Crawford, K.; and Paglen, T. 2021. Excavating AI: The politics of images in machine learning training sets. *Ai & Society*, 36(4): 1105–1116.

De Vries, T.; Misra, I.; Wang, C.; and Van der Maaten, L. 2019. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 52–59.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Do, T.-T.; Nguyen, A.; and Reid, I. 2018. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *2018 IEEE international conference on robotics and automation (ICRA)*, 5882–5889. IEEE.

Fang, A.; Jose, A. M.; Jain, A.; Schmidt, L.; Toshev, A. T.; and Shankar, V. 2023. Data Filtering Networks. In *NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models*.

Fayard, A.-L.; and Weeks, J. 2014. Affordances for practice. *Information and Organization*, 24(4): 236–249.

Gaviria Rojas, W.; Diamos, S.; Kini, K.; Kanter, D.; Janapa Reddi, V.; and Coleman, C. 2022. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. *Advances in Neural Information Processing Systems*, 35: 12979–12990.

Gibson, J. J. 2014. The theory of affordances:(1979). In *The people, place, and space reader*, 56–60. Routledge.

Griswold, W. 2012. *Cultures and societies in a changing world*. Sage.

Gupta, A.; Dollar, P.; and Girshick, R. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5356–5364.

Hassanin, M.; Khan, S.; and Tahtali, M. 2021. Visual affordance and function understanding: A survey. *ACM Computing Surveys (CSUR)*, 54(3): 1–35.

Hessel, J.; Holtzman, A.; Forbes, M.; Le Bras, R.; and Choi, Y. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, 7514–7528.

Ignat, O.; Bai, L.; Nwatu, J. C.; and Mihalcea, R. 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)*, 1239–1259.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A.; et al. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7): 1956–1981.

Li, Y.; Yang, J.; Song, Y.; Cao, L.; Luo, J.; and Li, L.-J. 2017. Learning from noisy labels with distillation. In *Proceedings of the IEEE international conference on computer vision*, 1910–1918.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, B.; Wang, L.; Lyu, C.; Zhang, Y.; Su, J.; Shi, S.; and Tu, Z. 2024. On the cultural gap in text-to-image generation. In *ECAI 2024*, 930–937. IOS Press.

Longpre, S.; Singh, N.; Cherep, M.; Tiwary, K.; Materzynska, J.; Brannon, W.; Mahari, R.; Obeng-Marnu, N.; Dey, M.; Hamdy, M.; et al. 2024. Bridging the Data Provenance Gap Across Text, Speech and Video. *arXiv preprint arXiv:2412.17847*.

Maier, J. R.; and Fadel, G. M. 2009. Affordance based design: a relational theory for design. *Research in Engineering Design*, 20: 13–27.

Mihalcea, R.; Ignat, O.; Bai, L.; Borah, A.; Chiruzzo, L.; Jin, Z.; Kwizera, C.; Nwatu, J.; Poria, S.; and Solorio, T. 2025. Why AI Is WEIRD and Shouldn't Be This Way: Towards AI for Everyone, with Everyone, by Everyone. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 28657–28670.

Mur-Labadia, L.; Guerrero, J. J.; and Martinez-Cantin, R. 2023. Multi-label affordance mapping from egocentric vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5238–5249.

Nguyen, T.; Gadre, S. Y.; Ilharco, G.; Oh, S.; and Schmidt, L. 2023. Improving multimodal datasets with image captioning. *Advances in Neural Information Processing Systems*, 36: 22047–22069.

Norman, D. A. 1999. Affordance, conventions, and design. *interactions*, 6(3): 38–43.

Nwatu, J.; Ignat, O.; and Mihalcea, R. 2023. Bridging the Digital Divide: Performance Variation across Socio-Economic Factors in Vision-Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 10686–10702.

Nwatu, J.; Ignat, O.; and Mihalcea, R. 2025. Uplifting Lower-Income Data: Strategies for Socioeconomic Perspective Shifts in Large Multi-modal Models. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2127–2144. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-189-6.

Paullada, A.; Raji, I. D.; Bender, E. M.; Denton, E.; and Hanna, A. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11).

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*.

Ramaswamy, V. V.; Lin, S. Y.; Zhao, D.; Adcock, A.; van der Maaten, L.; Ghadiyaram, D.; and Russakovsky, O. 2023. Geode: a geographically diverse evaluation dataset for object recognition. *Advances in Neural Information Processing Systems*, 36: 66127–66137.

Redmon, J.; and Farhadi, A. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Rosling, H.; Rosling, O.; and Rönnlund, A. R. 2018. *Factfulness: ten reasons we're wrong about the world–and why things are better than you think*. Flatiron books.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35: 25278–25294.

Shankar, S.; Halpern, Y.; Breck, E.; Atwood, J.; Wilson, J.; and Sculley, D. 2017. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*.

Srinivasan, K.; Raman, K.; Chen, J.; Bendersky, M.; and Najork, M. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 2443–2449.

Thomee, B.; Shamma, D. A.; Friedland, G.; Elizalde, B.; Ni, K.; Poland, D.; Borth, D.; and Li, L.-J. 2016. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2): 64–73.

Tschannen, M.; Gritsenko, A.; Wang, X.; Naeem, M. F.; Alabdulmohsin, I.; Parthasarathy, N.; Evans, T.; Beyer, L.; Xia, Y.; Mustafa, B.; Hénaff, O.; Harmsen, J.; Steiner, A.; and Zhai, X. 2025. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. arXiv:2502.14786.

Tylor, E. B. 1871. *Primitive culture: researches into the development of mythology, philosophy, religion, art, and custom*, volume 2. J. Murray.

Woolson, R. F. 2005. Wilcoxon signed-rank test. *Encyclopedia of Biostatistics*, 8.

Ye, L.; Cardwell, W.; and Mark, L. S. 2009. Perceiving multiple affordances for objects. *Ecological Psychology*, 21(3): 185–217.