GeoDiv: Measuring Concept Diversity of Images Across Geographical Regions

Abhipsa Basu* Mohana Singh* R. Venkatesh Babu
Vision and AI Lab
Indian Institute of Science, Bangalore, India
abhipsabasu@iisc.ac.in

Abstract

Image datasets—real and synthetic—often lack geographical diversity in how concepts are portrayed across regions. Existing metrics rely on curated datasets or visual dissimilarities, limiting interpretability. We propose GeoDiv, a metric that leverages large language models to identify region-specific attribute variations for a concept, uses a VQA model to measure their prevalence in images, and computes entropy over the resulting distributions. Applied to real and synthetic datasets (including Stable Diffusion and Flux.1-Schnell) across three concepts (house, car, bag) and six countries, GeoDiv reveals higher diversity in real-world images, with the UK and Japan being least diverse and Colombia the most. Our results underscore the need for geographical nuance in generative models and we believe that GeoDiv as a step toward measuring and mitigating regional biases.

1. Introduction

Recent studies reveal a lack of geographical diversity in both real and synthetic image datasets, where concepts are often portrayed through regional stereotypes [1, 9, 10]. For instance, prompting Stable Diffusion [15] with "`photo of a car in Africa`" typically yields dusty cars in desert settings, overlooking the continent's rich diversity. Existing geo-diversity metrics either rely on curated datasets [7, 9], which are limited in global coverage [13], or measure visual dissimilarity ignoring the nuances of geographical variations. While metrics like Vendi Score [7] have no dependence on reference datasets, they lack interpretability regarding how a concept's appearance varies (in terms of concrete attributes) across regions (e.g., a car in Africa may appear in desert, coastal, savannah, or urban environments).

In this paper, we propose *GeoDiv*, a metric that leverages the world knowledge of large language models (LLMs) to identify region-specific variations in the attributes of a given concept. Starting with a concept (e.g., house), the LLM is prompted to generate questions about its attributes (e.g., What material is the house made of?). For each question, the LLM then produces region-specific answer candidates (e.g., for Nigeria), capturing intra-regional diversity. After filtering out redundant responses, *GeoDiv* uses a visual question answering (VQA) model to estimate the frequency distribution of these attribute values within a set of images sharing the same concept and region. The final *GeoDiv* score is computed as the average entropy across the answer distributions for all questions.

We use the proposed metric to evaluate the geographical diversity of images for three concepts frequently studied in prior work [2, 10]—house, car, and bag—across six countries spanning multiple continents: the United Kingdom, Nigeria, Japan, Turkey, Colombia, and Indonesia. The images are drawn from both a real-world dataset (GeoDE [13]) and generated using state-of-the-art text-to-image models: Stable Diffusion 2.1, Stable Diffusion 3, and Flux.1-Schnell. Real-world images exhibit greater diversity across all concepts, with SD v3 showing the least. Country-wise, the UK and Japan appear least diverse. Attribute-level variations across countries are especially noticeable in aspects like background (e.g., urban vs. rural), material of construction, and overall condition of the concept in question. Unlike existing metrics, *GeoDiv* enables interpretable, attribute-specific analysis and helps surface potential regional stereotypes embedded in datasets and generative models.

2. Related Work

Metrics Measuring Image Diversity: Image diversity metrics are typically categorized into two types. The first compares a given image set to a reference set—e.g., FID [11], which compares feature distributions using a pre-trained Inception network [16]. We exclude such metrics due to the absence of large-scale geo-diverse reference datasets [8, 13]. The second type assesses variation within the given set. Pairwise Distance Metrics [3, 6] compute av-

*Equal contribution

erage distances between image embeddings (e.g., Inception or CLIP [12]), while Vendi-Score [7] measures entropy over the eigenvalues of the feature kernel matrix. However, these approaches capture only visual variation, not geographical diversity across concept attributes.

Leveraging the World Knowledge of Large-Scale Models: Trained on internet-scale data, LLMs and VLMs encode rich knowledge about global cultures and demographics, which many recent works have utilized. OASIS [5] quantifies stereotypes in text-to-image generation by comparing LLM-predicted attribute distributions for nationalities with those inferred from generated images via a VQA model. GRADE [14] adopts a similar approach to assess visual diversity in everyday objects, while DSG [4] evaluates image-text consistency. To our knowledge, we are the first to apply this framework to analyze region-wise diversity in images of a given concept.

3. GeoDiv

The proposed metric *GeoDiv* quantifies region-wise diversity of a given concept by leveraging the world-knowledge of the LLMs and visual recognition capabilities of the VLMs. Computing this metric entails three stages primarily. The prompts used for each stage are provided in the Appendix.

Question and Answer Generation. Inspired by GRADE [14], we first prompt the LLM to generate nine socio-economically framed questions for a given concept c (e.g., What type of road or terrain is the car on?). For each question, the LLM then generates region-specific answers, avoiding a universal answer set since attribute values often vary across regions. For example, responses to Does the surrounding environment of the car suggest a specific climate? depend heavily on the region's geography. Generated answers are filtered by the LLM to remove redundancy, irrelevance, and those that are visually hard to detect.

Generating Answer Frequency We query a state-of-the-art VQA model with each question, providing the pre-generated answers as options while allowing free-form responses if the correct answer is missing. For each concept c and region r, we compute the answer distribution per question and follow GRADE [14] to calculate its *normalized* entropy — higher entropy indicates greater diversity. *GeoDiv* is defined as the average entropy across all questions for a given (c, r) pair.

$$\text{GeoDiv}(r, c) = \frac{1}{|\mathcal{Q}_c|} \sum_{q \in \mathcal{Q}_c} \hat{H}(q, \mathcal{A}_c^r) \tag{1}$$

where, $\mathcal{Q}_c$ refers to the set of questions pertaining to concept c, $\hat{H}(q, \mathcal{A}_c^r)$ is the entropy for a question $q \in \mathcal{Q}_c$ and the corresponding list of answers for the region r and concept c, denoted by $\mathcal{A}_c^r$. An overview of our proposed pipeline is outlined in Figure 1.

4. Experiments

4.1. Concepts and Regions

To explore images from different regions, we analyze images from six countries, chosen from different parts of the world: the United Kingdom, Nigeria, Colombia, Indonesia, Japan, and Turkey. For the scope of this paper, we analyze the diversity of images from these countries across three concepts: house, car, and bag, selected based on previous works [2, 10].

Datasets: We evaluate geographical diversity using both real and synthetic image datasets. For real data, we use GeoDE [13] due to its global coverage. Synthetic datasets are generated using prompts of the form "A photo of a {c} in {r}" for each concept c and country r, using Stable Diffusion 2.1, 3 [15], and Flux.1-schnell[1]. Image counts are matched to GeoDE's distribution across the six selected countries (Table 10 in Appendix).

4.2. Implementation Details

Recall that computing *GeoDiv* requires question and answer generation using an LLM, after which a VQA model is employed to obtain the distribution of answers (see Section 3). We choose `Gemini-1.5-pro-002` for both tasks as it has been shown to possess sufficient knowledge about the different geographies of the world [17]. Gemini is accessed through the Vertex AI API [2] for all inference tasks.

Question Generation: Our process produces an average of nine questions per concept. To improve quality, we manually preprocess each question by: (1) splitting compound questions into atomic ones (e.g., "Does the chair have upholstery, and if so, what material?" becomes "Does the chair have visible upholstery?" and "What material is the upholstery made from?"); (2) removing indicative examples that may bias model behavior (e.g., "What type of bag is shown (e.g., handbag, backpack)?" becomes "What type of bag is shown?"); (3) adding a background-related question (e.g., indoor vs. outdoor) if missing; (4) reducing vagueness in LLM-generated questions; and (5) tagging each question as having either a fixed (F) or variable (NF) set of plausible answers (e.g., "Is the house single-story or multi-story?" is tagged as F). All questions for each concept can be found in Appendix subsection A.3.

Answer Generation and Filtering: After building the question bank, we pair each question with prompts like "`A photo of a {c} in {r}`" and use an LLM to generate plausible answers in the context of the concept and region (see Appendix 3). For NF-tagged questions, we apply

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

[2] https://cloud.google.com/vertex-ai/docs/reference/rest

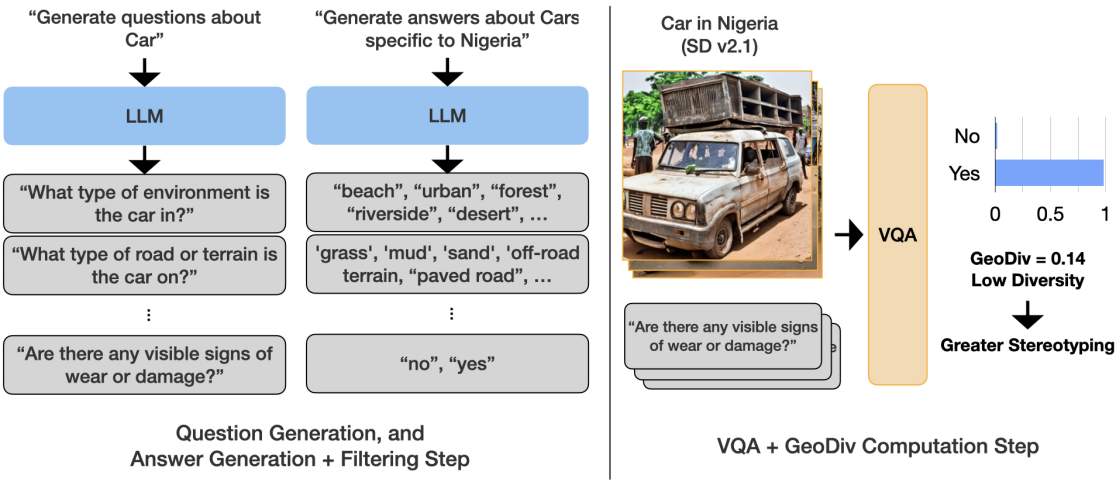

Figure 1. **Pipeline for GeoDiv Computation**. In this figure, we summarize the pipeline for computing the proposed metric. Given a concept c and a country r, first an LLM generates questions about the attributes of c, and non-redundant attribute values relevant to both c and r. In the next phase, a VQA model is employed to find the distribution of these attribute values in a given set of images. Finally, GeoDiv computes the average entropy value of the distributions obtained across all questions.

Table 1. **GeoDiv Scores** for house, car and bag, measured across six countries: the United Kingdom (UK), Nigeria (Nig), Japan (Jap), Turkey (Tur), Colombia (Col), Indonesia (Ind), and four datasets: GeoDE (real), SD v2.1, SD v3, Flux.1-Schnell. Averaged across concepts, the real dataset exhibits most diversity (0.61), followed by SD v2.1 (0.59), Flux.1-Schnell (0.59), SD v3 (0.56). Among nations, the UK images exhibit lowest geographical diversity.

Concept	Dataset	UK	Nig	Jap	Tur	Col	Ind	Avg
house	GeoDE	0.468	0.676	0.551	0.693	0.62	0.565	**0.596**
	SD2.1	0.529	0.495	0.531	0.693	0.578	0.527	0.559
	SD3	0.407	0.465	0.504	0.49	0.559	0.488	**0.486**
	Flux1	0.46	0.569	0.554	0.53	0.62	0.503	0.539
	Avg	**0.466**	0.551	0.53	**0.60**	0.59	0.52	-
car	GeoDE	0.523	0.594	0.619	0.688	0.574	0.592	0.598
	SD2.1	0.414	0.390	0.499	0.469	0.373	0.398	**0.424**
	SD3	0.548	0.554	0.491	0.719	0.602	0.504	0.570
	Flux1	0.655	0.557	0.602	0.644	0.636	0.544	**0.606**
	Avg	0.535	0.524	0.553	**0.63**	0.546	**0.509**	-
bag	GeoDE	0.607	0.679	0.7	0.599	0.649	0.625	0.643
	SD2.1	0.705	0.736	0.735	0.736	0.767	0.765	**0.741**
	SD3	0.427	0.652	0.492	0.603	0.703	0.711	**0.598**
	Flux1	0.548	0.629	0.537	0.558	0.735	0.711	0.617
	Avg	**0.57**	0.67	0.62	0.62	**0.71**	0.70	-

an additional filtering step (Appendix 4) to remove redundant or out-of-scope answers. We also generate reasoning for each filtering decision to ensure consistency and transparency.

VQA Stage: We first perform a visibility check by querying the VQA model to determine if the attribute in a question is detectable in the image. If the answer is "No," all follow-up questions about that attribute are discarded, following prior work [4], to reduce hallucinations from model bias. Attributes detectable in fewer than 50% of images are also excluded from further analysis. The remaining questions—fixed (F) and variable (NF)—are then paired with images that passed the check and fed into the VQA model using the prompt in Appendix 5, along with the prefiltered answer list. In addition, multiple option selection is permitted for more precision.

GeoDiv Computation: The output from the VQA step provides the frequency of the generated attribute values within the dataset. Similar to the GRADE method, we also include a "None of the above" option in the list of possible answers. However, unlike GRADE, we find that only a small percentage of images (about 1-2%) fall into this category due to the prior visibility check. As a result, these image-question pairs are not taken into account in the frequency calculation.

Across all stages, the LLM is configured with a temperature of 0.0, top-p value of 0.01, and top-k value of 1 to

enforce deterministic generation. The maximum number of output tokens is set to 2000.

4.3. GeoDiv Scores

Evaluating the geographical diversities of the real and synthetic datasets leads us to multiple interesting questions and insights, which we summarize below.

4.3.1. Dataset-wise Comparison

For every dataset, we observe the GeoDiv scores for all the concepts and countries to be generally lower than < 0.7. Averaged across the concepts, the real dataset exhibits more diversity than those for SD v2.1 (0.59), SD v3 (0.56) and Flux.1-Schnell (0.59). This indicates that the generative models tend to amplify stereotypes of concepts based on countries, highlighting the need for more geographically balanced training sets for these models.

Comparing between the two versions of Stable Diffusion, we find that for house and bag, the GeoDiv scores for SD v3 are lower than that of SD v2.1. However, we observe that the question set for cars leans more heavily into stereotypical representations, and with SD2.1 often generating depictions of old or worn-out cars, leads to overall diversity reduction (Appendix Figures 2 and 3). On the other hand, newer models are biased toward more visually aesthetic imagery (Appendix Fig. 4). Across all three concepts, Flux.1-Schnell images are found to be more geographically diverse than those of SD v3, though they have similar scores with SD v2.

The GeoDiv scores for the individual datasets, concepts and countries are summarized in Table 1. Details on the questions with the least entropy for each concept are shown in Appendix subsection A.4.

4.3.2. Country-wise Analysis

House: The UK shows the lowest geographic diversity across all datasets (GeoDiv$(r, c) = 0.47$), followed by Indonesia (0.52), with houses of the former nation commonly being multistoreyed, brick-built, gabled, and suburban—indicating limited architectural variation. In contrast, Turkey and Colombia are the most diverse (GeoDiv scores of 0.60 and 0.59 respectively), showing broader variation in materials, roofs, storeys, and backgrounds. Notably, 97% of Nigerian house images from Stable Diffusion v3 depict rural settings, unlike GeoDE's balanced distribution. Over 80% of UK images across SD 3, Flux.1, and GeoDE are suburban, while SD 2.1 offers more variety. These trends suggest country-specific stereotyping in generated images.

Car: Averaged across datasets, we find the UK, Japan, Turkey and Indonesia to have similar geographical diversity. Such cars are mostly seen in city streets and urban backgrounds, on paved roads. Surprisingly, Nigeria has the highest diversity, as it has a more balanced distribution of urban and rural backgrounds, and is found on paved roads, sand, off-road terrain, in contrast to other nations, where a more urban background is common. However, the cars from Nigeria mostly appear to be old, whereas those from other countries have higher proportion of new cars. Interestingly, Flux.1-Schnell shows higher diversity for Colombia, with balanced urban, rural and suburban representation, environment of background (e.g., coastal, mountainous, forest, etc). The distribution of car colors is also relatively uniform compared to other datasets.

Bag: Similar to houses, UK has the lowest diversity among all studied countries, especially for SD v3 and Flux.1-Schnell, followed by Japan and Turkey. For all three countries, this happens due to the lack of representations of different materials that bags are generally made of (dominated by leather), as well as in types of bags (dominated by tote and handbags). Nigeria, Colombia and Indonesia have the highest scores across models, which can be attributed to a number of factors - diversity in materials, color, size, etc. As with other concepts, Nigerian backgrounds are overwhelmingly rural ($> 80\%$ on average) for the SD generations, whereas the real images from GeoDe and those generated from Flux.1-Schnell show a more urbanized background. Such observations show that the VQA model itself is not biased towards the country, and can indeed be a useful tool for measuring diversity.

4.3.3. Correlation with Vendi Score

We further compute the correlation of the proposed *GeoDiv* with the popular existing metric Vendi-Score [7], used to measure diversity of images. The average Pearson's Correlation Coefficient ρ across all concepts, countries and datasets turns out to be very weak (0.13), indicating that the proposed metric indeed captures something beyond visual similarities (which is not enough to capture the nuances and complexities of geographical variations across the world), further highlighting the uniqueness of *GeoDiv*. Further details can be found in Appendix subsection A.2.

5. Conclusion and Limitations

In this paper, we proposed GeoDiv, an interpretable metric that leverages the world knowledge of LLMs and the visual recognition capabilities of VLMs to measure the geographical diversity of images for any concept across countries. Evaluations across three concepts and six countries show that real-world images are more diverse than those generated by popular text-to-image models, with Stable Diffusion v3 being the least diverse. Country-level analysis reveals limited diversity in the UK and Japan, while generated images of Nigeria often depict rural settings with visible wear, reflecting harmful stereotypes. While GeoDiv may inherit biases from the underlying LLM and VLM, it offers a reference data-free approach for quantifying geographic diversity and uncovering regional biases in image datasets.

References

[1] Reyhane Askari Hemmat, Melissa Hall, Alicia Sun, Candace Ross, Michal Drozdzal, and Adriana Romero-Soriano. Improving geo-diversity of generated images with contextualized vendi score guidance. In *European Conference on Computer Vision*, pages 213–229. Springer, 2024. 1

[2] Abhipsa Basu, R. Venkatesh Babu, and Danish Pruthi. Inspecting the geographical representativeness of images from text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5136–5147, 2023. 1, 2

[3] Victor Boutin, Thomas Fel, Lakshya Singhal, Rishav Mukherji, Akash Nagaraj, Julien Colin, and Thomas Serre. Diffusion models as artists: Are we closing the gap between humans and machines? *arXiv preprint arXiv:2301.11722*, 2023. 1

[4] Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. *arXiv preprint arXiv:2310.18235*, 2023. 2, 3

[5] Sepehr Dehdashtian, Gautam Sreekumar, and Vishnu Naresh Boddeti. Oasis uncovers: High-quality t2i models, same old stereotypes. *arXiv preprint arXiv:2501.00962*, 2025. 2

[6] Lijie Fan, Kaifeng Chen, Dilip Krishnan, Dina Katabi, Phillip Isola, and Yonglong Tian. Scaling laws of synthetic images for model training... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7382–7392, 2024. 1

[7] Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023. 1, 2, 4

[8] William Gaviria Rojas, Sudnya Diamos, Keertan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. *Advances in Neural Information Processing Systems*, 35:12979–12990, 2022. 1

[9] Melissa Hall, Candace Ross, Adina Williams, Nicolas Carion, Michal Drozdzal, and Adriana Romero Soriano. Dig in: Evaluating disparities in image generations with indicators for geographic diversity. *arXiv preprint arXiv:2308.06198*, 2023. 1

[10] Melissa Hall, Samuel J Bell, Candace Ross, Adina Williams, Michal Drozdzal, and Adriana Romero Soriano. Towards geographic inclusion in the evaluation of text-to-image models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 585–601, 2024. 1, 2

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 1

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[13] Vikram V Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. Geode: a geographically diverse evaluation dataset for object recognition. *Advances in Neural Information Processing Systems*, 36:66127–66137, 2023. 1, 2

[14] Royi Rassin, Aviv Slobodkin, Shauli Ravfogel, Yanai Elazar, and Yoav Goldberg. Grade: Quantifying sample diversity in text-to-image models. *arXiv preprint arXiv:2410.22592*, 2024. 2

[15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2

[16] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, 2017. 1

[17] Zhiqiang Wang, Dejia Xu, Rana Muhammad Shahroz Khan, Yanbin Lin, Zhiwen Fan, and Xingquan Zhu. Llmgeo: Benchmarking large language models on image geolocation in-the-wild. *arXiv preprint arXiv:2405.20363*, 2024. 2

A. Appendix

A.1. Prompts used while computing *GeoDiv*

The prompt used for question generation, answer generation and answer filtering can be found in Tables 2, 3 and 4 respectively. For the VQA step, the prompt is shown in Table 5.

A.2. Vendi Score

The Vendi Scores for each concept, country and dataset can be found in Table 6. In addition to the Vendi-scores, we also provide their Pearson's Correlation Coefficient (ρ) with respect to the proposed *GeoDiv* values.

A.3. Generated Questions by Concept

A.3.1. Concept: house

1. Are any cultural or religious symbols visible on or around the house?
2. Does the house appear to be located in a rural, suburban, or urban environment?
3. Does the house appear to have access to electricity?
4. How is the area immediately surrounding the house landscaped?
5. Is the house single-story or multi-story?
6. What is the general condition of the house?
7. What is the primary material used for the exterior walls of the house?
8. What kind of climate does the surrounding environment suggest?

Prompt for Concept-wise Question Generation

```
You are a helpful assistant.
Help me ask questions about images that depict certain concepts.
I will provide you a concept.  Your task is to analyze the concept's typical
visual attributes and generate **clear and simple questions** about the concept.
The questions should examine for stereotypical biases, socioeconomic indicators
(e.g., presence of luxury brands, construction materials, overall condition such as
well-maintained, dilapidated, decent), and geographic and cultural diversity of the
concept.
Your questions should involve concrete attributes and be answerable purely by visually
inspecting the image.
Include at least one question that evaluates the geographical context (e.g., rural,
urban, high-tech, desert, mountainous, etc.).
Do NOT ask follow-up or compound questions within the same question.
Do NOT ask questions that cannot be answered by visually inspecting the image or
require inference or external context beyond what is shown.
Do NOT ask more than 10 questions.
Here's an example:
**concept**:  a house
**assumed attributes**:  houses can have different architectural, geographic,
socioeconomic, and cultural representations.
**questions**:
1.  What is the type of the house?
2.  What primary construction material is used for the house walls?
3.  What type of roof does the house have?
4.  Is the house single-storey or multi-storey?
5.  Are there any religious or cultural symbols visible on the house?
6.  Do the surroundings suggest an urban setting with modern infrastructure or a rural
one?
7.  What is the general condition of the house?
8.  What kind of ground cover is visible in front of or around the house?
9.  What kind of climate or geography does the house appear to be situated in?
```

Table 2. System prompt used to generate questions for each concept. Note that the model is specifically instructed to generate questions about stereotypical biases, socio-economic indicators, overall condition and background in geographical context. Such instructions help the model construct questions relevant for investigating geographical diversity. The prompt also provides with an example for the concept *house*.

9. What type of house is shown?
10. What type of roof does the house have?

A.3.2. Concept: car

1. Are there any people visible in or around the car?
2. Are there any visible signs of damage or wear on the car?
3. Does the background suggest an urban or rural setting?
4. Does the car appear luxurious or economical?
5. Does the car appear to be new or old based on its overall condition and style?
6. Does the surrounding environment suggest a specific climate?
7. What is the color of the car?
8. What type of car is visible?
9. What type of road or terrain is the car on?

A.3.3. Concept: bag

1. Are there any cultural or symbolic motifs on the bag?
2. Are there any logos or brand markings visible on the bag?
3. Does the bag appear commercially made with professional stitching, or does it appear more makeshift and locally handmade?
4. Does the bag appear new, moderately used, or worn?
5. Does the environment surrounding the bag appear to be rural or urban?
6. What color is the bag?
7. What is the context in which the bag is shown?

Prompt for Answer Generation Given a Concept and Region

`I have a question that is asked about an image. I will provide you with the question and a caption of the image. Your job is to first carefully read the question and analyze, then hypothesize plausible answers to the question assuming you could examine the image (instead, you examine the caption).`
`The answers should be in a list, as in the example below.`
`Do not write anything other than the plausible answers.`
`Do not provide extra details to your answers in parentheses (e.g., white and NOT 'white (for decorated cookies)').`
`Do your best to be succinct and not overly-specific.`
`If the question is very open-ended, like 'Is there anything on the table?' or 'Is the cake decorated with any specific theme or design?', the answer should be strictly ['yes', 'no'].`
`Example:`
`Caption: a helmet in a bike shop`
`Question: What type of helmet is depicted in the image?`
`Plausible answers: ["motorcycle helmets", "bicycle helmets", "football helmets", "construction helmets", "military helmets", "firefighter helmets", "rock climbing helmets", "hockey helmets"]`

Table 3. System Prompt used to generate answers for a given question and caption pertaining to a concept c and region r. We provide an example question and plausible answers, as specified in GRADE. In the user prompt, we pass the question, and the caption is set as `a c in r`.

8. What is the size of the bag?
9. What material is the bag made of?
10. What type of bag is shown?

A.4. Question-wise Diversity Scores

We report the three questions with the lowest entropy scores for house, car and bag in Tables 7, 8 and 9 respectively.

Prompt for Filtering Answers Generated

You are provided with a concept, a question about an image of this concept, and a list of possible answers.

Your task is to filter out answers that do not belong in the final list based on the following three filtering criteria:

(1) Out of Scope -- If an answer belongs to a completely different category than the rest, remove it. Example: If all answers describe number of table legs, but one says "wooden surface", remove it.

(2) \None of the Above" -- Do not allow answers that suggest no correct answer exists, such as "none", "no visible toppings", etc. Remove these.

(3) Semantic Redundancy -- If two answers mean the same thing but one is more specific, keep the broader term and remove the more specific one. Example: Keep "chocolate" and remove "chocolate drizzle".

How to Respond: First, carefully read the concept, question and answers. Then, apply each filtering rule and explain which answers are removed and why. Finally, provide the reasoning and the filtered answers list obtained by taking into account the reasoning steps. Provide the response in JSON format with the following structure:

reasoning_steps: ["Step 1", "Step 2", ...],

"filtered_answers": ["answer1", "answer2", "answer3"]

Example 1

Concept: A photo of Popcorn

Question: Are there any visible toppings or additions, such as butter or cheese?

Answers: ["no", "yes", "chocolate", "cinnamon", "butter", "none", "chocolate drizzle", "no visible toppings", "plain", "caramel", "cheese"]

reasoning_steps: [""no" and "yes" -- Out of scope, as they do not describe specific toppings whereas the other answers do (Criterion 1)", ""none" and "no visible toppings" -- Removed (Criterion 2: "None of the above")", ""chocolate drizzle" and "chocolate" -- "chocolate drizzle" is more specific, so remove it (Criterion 3: Redundancy)"]

filtered_answers: ['chocolate', 'cinnamon', 'butter', 'plain', 'caramel', 'cheese']

Example 2

Concept: A photo of a table

Question: How many legs does the table have?

Answers:["no legs", "no", "yes", "one central pedestal", "one leg", "two trestle supports", "a trestle base", "two legs", "six legs", "a pedestal base", "three legs", "multiple legs", "five legs", "four legs"]

reasoning_steps: ["anything with 'trestle' is too specific and out of scope (criterion 1)", ""no leg" Removed as it matches Criterion 2", ""two legs", "three legs", "four legs", "five legs", "six legs" -- Redundant. Keep the broadest term, "multiple legs", and remove the others (Criterion 3)"]

filtered_answers: ['one leg', 'multiple legs']

Table 4. System Prompt used to filter irrelevant and redundant answers for a given question and the corresponding answer list belonging to concept c and region r. For simplicity, we keep the example same as in GRADE. In the user prompt, we pass the question, the answers generated for the same, and the 'concept' attribute in the prompt (different from the concept we define in the paper) is set as `'a c in r'`.

Prompt for Visual Question Answering

```
You are a helpful assistant.
Answer the given question by selecting one or more categories from the provided list.
Select "None of the above" if none of the other options are relevant.
To come up with the correct answer, carefully analyze the image and think step-by-step
before providing the final answer.
Provide the reasoning steps that lead to the final conclusion and the final list of
answers, taking into account the reasoning steps.  Provide the response in JSON format
with the following structure:
reasoning_steps:  ['Step 1', 'Step 2', ...]
answer:  ['category 1', 'category2', ...]
```

Table 5. System prompt used to for the VQA step. Note that we allow an additional option to the pre-generated set of answers: None of the above, to account for missing options.

Table 6. **Vendi Scores** for house, car and bag, measured across six countries: the United Kingdom (UK), Nigeria (Nig), Japan (Jap), Turkey (Tur), Colombia (Col), Indonesia (Ind), and four datasets: GeoDE (real), SD v2.1, SD v3, Flux.1-Schnell. The average Pearson's Correlation Coefficient between *GeoDiv* and Vendi-Scores is 0.13, indicating the insufficiency of metrics based on visual similarity in capturing the nuances and complexities of geographical variations across the globe.

Concept	Dataset	UK	Nig	Jap	Tur	Col	Ind	Avg	ρ (GeoDiv, Vendi)
house	GeoDE	2.45	2.38	2.39	2.53	2.65	3.19	2.60	-0.10
	SD2.1	3.27	3.01	4.16	3.08	4.25	3.51	3.55	-0.12
	SD3	1.97	2.32	3.49	2.37	3.31	3.18	2.77	0.79
	Flux1	3.28	3.06	3.44	3.34	3.99	3.76	3.48	0.37
car	GeoDE	2.81	2.59	2.93	2.58	2.93	3.51	2.89	-0.55
	SD2.1	3.18	3.30	3.89	3.45	4.08	3.94	3.64	0.28
	SD3	2.89	2.70	3.53	2.88	3.98	3.95	3.32	-0.20
	Flux1	4.67	4.28	4.44	4.13	4.54	5.18	4.54	-0.06
bag	GeoDE	2.53	3.08	2.68	2.63	3.09	3.19	2.87	0.26
	SD2.1	2.67	3.20	3.75	3.24	3.38	3.52	3.29	0.67
	SD3	3.16	2.92	3.91	2.97	3.43	3.95	3.39	0.09
	Flux1	4.42	4.25	4.96	4.00	4.38	4.98	4.50	0.15

Table 7. Question-wise GeoDiv Scores for **house** measured across six countries: the United Kingdom (UK), Nigeria (Nig), Japan (Jap), Turkey (Tur), Colombia (Col), Indonesia (Ind), and four datasets: GeoDE (real), SD v2.1, SD v3, Flux.1-Schnell. Only lowest scoring three questions are reported for brevity.

Question	Dataset	UK	Nig	Jap	Tur	Col	Ind	Avg
Are any cultural or religious symbols visible on or around the house?	GeoDE	0.341	0.238	0.349	0.529	0.381	0.421	0.377
	SD2.1	0.276	0.191	0.830	0.784	0.270	0.421	0.462
	SD3	0.454	0.032	0.999	0.425	0.310	0.391	0.435
	Flux1	0.631	0.492	0.999	0.708	0.650	0.556	0.673
What is the general condition of the house?	GeoDE	0.398	0.639	0.282	0.601	0.333	0.568	0.470
	SD2.1	0.227	0.369	0.00	0.481	0.557	0.350	0.331
	SD3	0.309	0.820	0.027	0.602	0.545	0.142	0.407
	Flux1	0.367	0.130	0.094	0.060	0.033	0.053	0.123
What type of roof does the house have?	GeoDE	0.160	0.538	0.627	0.566	0.474	0.543	0.485
	SD2.1	0.305	0.649	0.257	0.486	0.651	0.591	0.490
	SD3	0.092	0.349	0.159	0.400	0.405	0.526	0.322
	Flux1	0.309	0.483	0.514	0.496	0.523	0.498	0.470

Table 8. Question-wise GeoDiv Scores for **car** measured across six countries: the United Kingdom (UK), Nigeria (Nig), Japan (Jap), Turkey (Tur), Colombia (Col), Indonesia (Ind), and four datasets: GeoDE (real), SD v2.1, SD v3, Flux.1-Schnell. Only lowest scoring three questions are reported for brevity.

Question	Dataset	UK	Nig	Jap	Tur	Col	Ind	Avg
What type of road or terrain is the car on?	GeoDE	0.147	0.234	0.454	0.604	0.000	0.245	0.281
	SD2.1	0.568	0.357	0.042	0.732	0.386	0.221	0.385
	SD3	0.201	0.000	0.042	0.614	0.134	0.285	0.213
	Flux1	0.147	0.000	0.392	0.187	0.098	0.172	0.166
Does the surrounding environment suggest a specific climate?	GeoDE	0.646	0.194	0.503	0.561	0.520	0.000	0.404
	SD2.1	0.212	0.080	0.480	0.487	0.287	0.000	0.258
	SD3	0.217	0.146	0.453	0.424	0.534	0.000	0.295
	Flux1	0.371	0.230	0.594	0.611	0.697	0.128	0.439
Does the background suggest an urban or rural setting?	GeoDE	0.082	0.578	0.602	0.751	0.136	0.436	0.431
	SD2.1	0.568	0.956	0.221	0.978	0.230	0.207	0.527
	SD3	0.540	0.190	0.000	0.581	0.199	0.642	0.359
	Flux1	0.772	0.629	0.422	0.345	0.621	0.567	0.559

Table 9. Question-wise GeoDiv Scores for **bag** measured across six countries: the United Kingdom (UK), Nigeria (Nig), Japan (Jap), Turkey (Tur), Colombia (Col), Indonesia (Ind), and four datasets: GeoDE (real), SD v2.1, SD v3, Flux.1-Schnell. Only lowest scoring three questions are reported for brevity.

Question	Dataset	UK	Nig	Jap	Tur	Col	Ind	Avg
What is the context in which the bag is shown?	GeoDE	0.188	0.135	0.597	0.178	0.375	0.103	0.263
	SD2.1	0.482	0.601	0.541	0.362	0.439	0.174	0.433
	SD3	0.236	0.486	0.363	0.287	0.296	0.400	0.345
	Flux1	0.707	0.725	0.545	0.700	0.476	0.502	0.609
Does the bag appear commercially made with professional stitching, or does it appear locally handmade?	GeoDE	0.427	0.330	0.233	0.315	0.265	0.138	0.285
	SD2.1	0.380	0.403	0.794	0.566	0.761	0.950	0.642
	SD3	0.995	0.900	0.077	0.931	0.911	0.000	0.636
	Flux1	0.712	0.532	0.135	0.267	0.312	0.079	0.340
Does the environment surrounding the bag appear to be rural or urban?	GeoDE	0.206	0.378	0.487	0.725	0.000	0.104	0.317
	SD2.1	0.985	0.993	0.187	0.404	0.385	0.397	0.558
	SD3	0.549	0.790	0.000	0.501	0.212	0.000	0.342
	Flux1	0.957	0.799	0.108	0.616	0.123	0.244	0.474

Table 10. **GeoDE distribution**. Object counts by country.

Object	**UK**	**Nig**	**Jap**	**Tur**	**Col**	**Ind**
house	63	307	168	150	108	117
car	92	203	139	161	97	136
bag	103	176	212	178	126	312

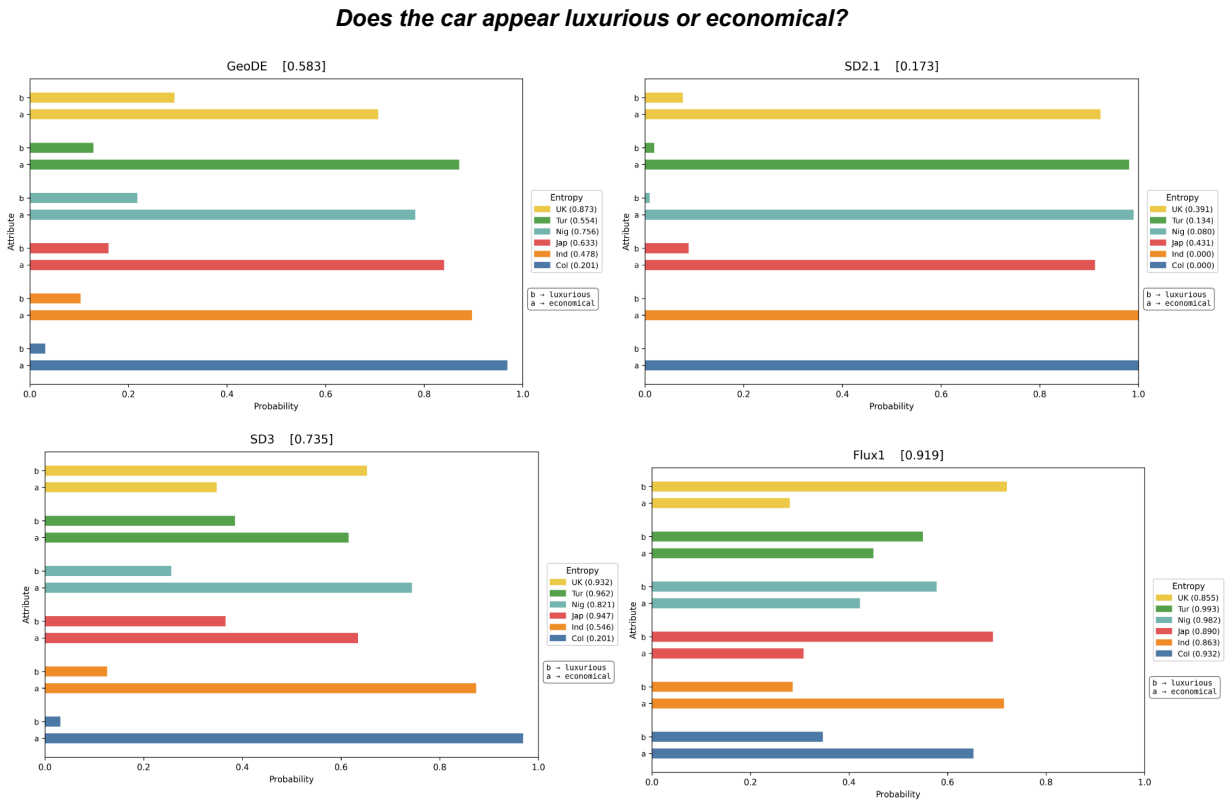

Figure 2. The plots display the probabilities across 6 countries for each value of the attribute (question: "Does the car appear luxurious or economical?"), for each dataset studied in this work. The plots are titled as "`Dataset-name [Average GeoDiv scores over all 6 countries]`". As the figure shows, SD2.1 is highly polarised with most generated images of cars being classified as "economical", leading to much lower GeoDiv scores compared to the other datasets.

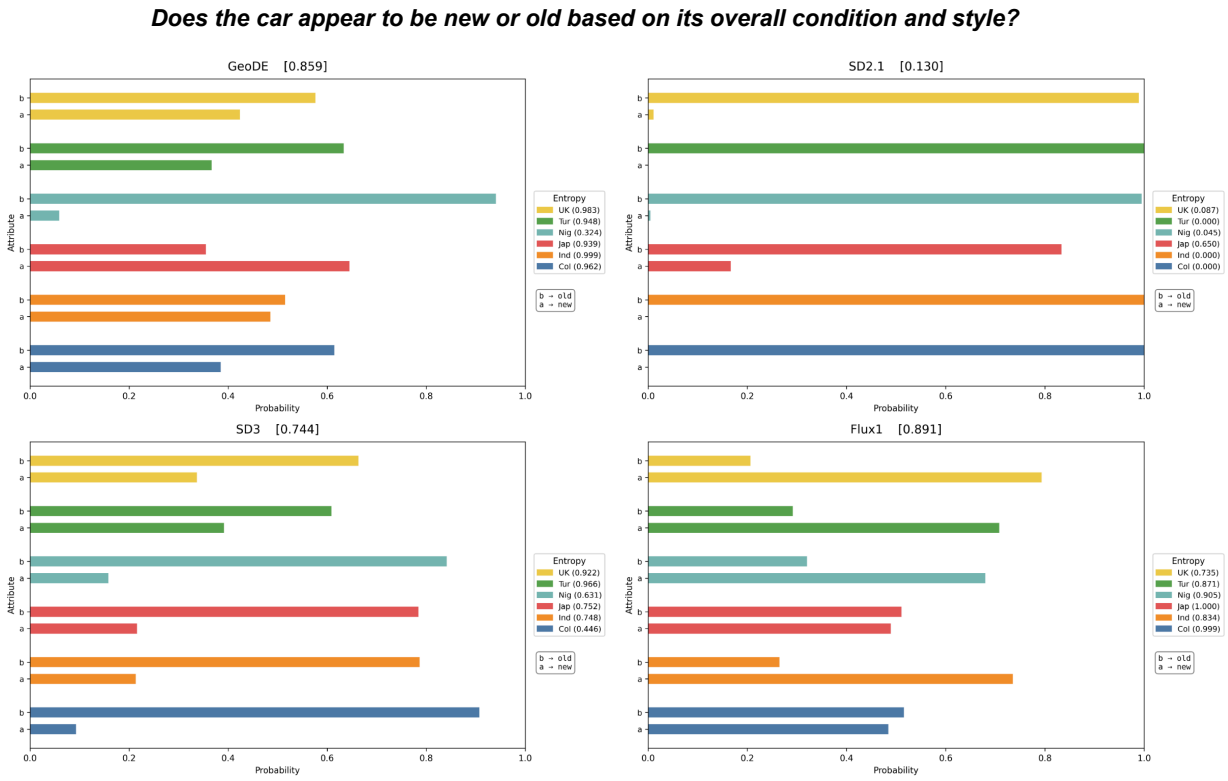

Figure 3. The plots display the probabilities across 6 countries for each value of the attribute (question: "Does the car appear to be new or old based on its overall condition and style?"), for each dataset studied in this work. The plots are titled as "`Dataset-name [Average GeoDiv scores over all 6 countries]`". As the figure shows, SD2.1 overwhelmingly generates images of "old" cars, leading to much lower GeoDiv scores compared to the other datasets.

Are there any visible signs of damage or wear on the car?

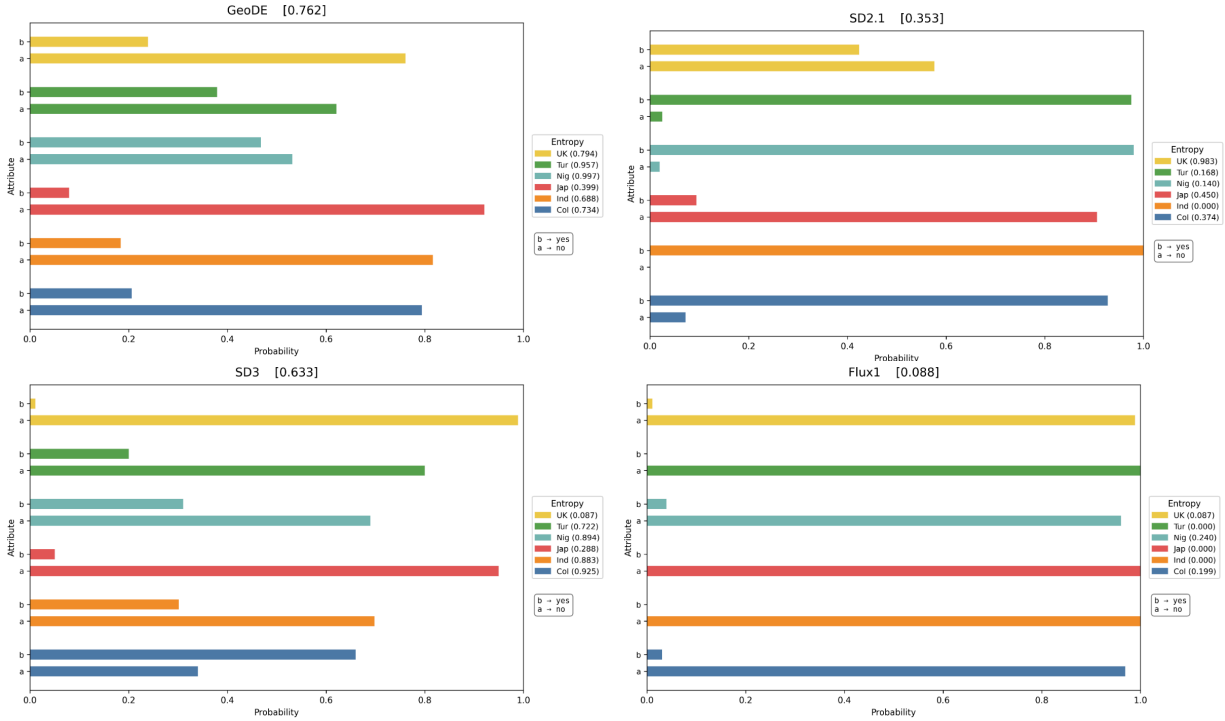

Figure 4. The plots display the probabilities across 6 countries for each value of the attribute (question: "Are there any visible signs of damage or wear on the car?"), for each dataset studied in this work. The plots are titled as "`Dataset-name [Average GeoDiv scores over all 6 countries]`". For all countries except the UK, SD2.1 overwhelmingly generates images of cars with "signs of damage and wear", leading to much lower GeoDiv scores compared to the other datasets. On the other hand, Flux1 is polarized in the opposite direction, hardly generating any car images with this attribute.