No Filter: Cultural and Socioeconomic Diversity in Contrastive Vision–Language Models

Angéline Pouget[*†] apouget@ethz.ch
Lucas Beyer lbeyer@google.com
Emanuele Bugliarello bugliarello@google.com
Xiao Wang wangxiao@google.com
Andreas Peter Steiner andstein@google.com
Xiaohua Zhai xzhai@google.com
Ibrahim Alabdulmohsin[†] ibomohsin@google.com

Google DeepMind

Abstract

We study cultural and socioeconomic diversity in contrastive vision–language models (VLMs). Using a broad range of benchmark datasets and evaluation metrics, we bring to attention several important findings. First, the common filtering of training data to English image–text pairs disadvantages communities of lower socioeconomic status and negatively impacts cultural understanding. Notably, this performance gap is not captured by—and even at odds with—the currently popular evaluation metrics derived from the Western-centric ImageNet and COCO datasets. Second, pretraining with global, unfiltered data before fine-tuning on English content can improve cultural understanding *without* sacrificing performance on said popular benchmarks. Third, we introduce the task of geo-localization as a novel evaluation metric to assess cultural diversity in VLMs. Our work underscores the value of using diverse data to create more inclusive multimodal systems and lays the groundwork for developing VLMs that better represent global perspectives.

1 Introduction

Contrastive vision–language models (VLMs) have emerged as a powerful and versatile method to bridge the gap between visual and textual information in deep learning systems. They utilize a dual-encoder architecture to map both images and texts into a shared latent space. Representations in this latent space are learned leveraging large datasets of noisy image-text pairs from the web. Work including CLIP [46], ALIGN [31] and SigLIP [75] validates this approach at scale with impressive zero-shot transfer results across a wide range of downstream tasks.

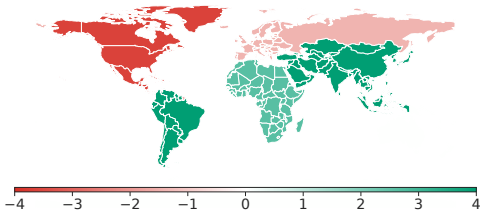
However, due to the growing range of applications for contrastive VLMs, it is imperative to evaluate them not only with respect to standard performance metrics, such as their classification accuracy on ImageNet-ILSRCV2012 [17] or image-text retrieval performance on COCO [38], both of which are Western-oriented [53, 16], but also in terms of "cultural diversity." We illustrate what we mean in Figure 1 (a) and (b). When performing zero-shot classification on images from the Google Landmarks Dataset (GLDv2) [66], SigLIP models trained on only English-language image–text pairs (henceforth

[*]Work performed while interning at Google DeepMind.
[†]Corresponding authors.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

(a) **en**: CN Tower (Canada) **globe**: Milad Tower (Iran)

(b) **en**: Notre-Dame Basilica of Montreal (Canada) **globe**: Catedral da Sé de São Paulo (Brazil)



(c) 0-shot accuracy improvement [%] on the Dollar Street dataset when switching from English-only (**en**) to global data with English translation (**globe-tl**).

Figure 1: Models trained on English image-text pairs exhibit a lack of diversity when evaluated on images from other regions, sometimes confusing landmarks with similar ones located in the West.

denoted **en**) tend to misclassify international landmarks as similar-looking landmarks located in English-speaking countries. Note that this is the currently prevalent pretraining approach in the literature. In contrast, SigLIP models trained on the full, global data (henceforth denoted **globe**) identify the correct landmarks.

A similar observation can also be made in Figure 1 (c), where we compare the zero-shot classification accuracy of a model trained on **en** data to that of a model trained on **globe** data with its text translated to English using the Google Translate API (henceforth denoted **globe-tl**). As can be seen, the **en** model seems to be biased towards data from Western regions. Switching to **globe-tl** data lowers performance for images from North America and Europe but significantly improves performance for other regions such as South America and Africa that are traditionally underrepresented in AI.

It is worth clarifying that cultural diversity in our context is different from "fairness" and the mitigation of societal stereotypes. While recent work has shown that CLIP models perpetuate and amplify social biases and stereotypes present in the training data [27, 7, 2, 43], our emphasis is different. We focus on improving the ability for a VLM to recognize and accurately interpret visual and textual data from a wide range of geographical, socioeconomic and cultural contexts, such as physical surroundings, traditions, customs and everyday goods. Evaluating cultural diversity is critical because it ensures that VLMs perform well across diverse environments and that they recognize and respect the varied perspectives that exist worldwide.

In this work, we present a comprehensive study of cultural and socioeconomic diversity, focusing on the impact of training data source, composition and processing (including translation), using the recently introduced SigLIP models [75] as a case study. We cover a wide range of benchmark datasets and evaluation metrics. Amongst our findings, we show that while SigLIP models trained on English-only image–text pairs (**en**) achieve state of the art results on popular benchmarks (ImageNet, COCO), this filtering disproportionately hurts model performance for low-income households and regions, and negatively impacts cultural diversity. Crucially, these **en** models achieve demonstrably lower performance on cultural diversity benchmarks *even after fine-tuning on more diverse and global data*. Conversely, models pretrained on the full, global data (**globe**, **globe-tl**) followed by brief English-only fine-tuning can match and even outperform the English-only baselines on the popular Western-oriented benchmarks, while also performing demonstrably better in cultural diversity benchmarks. Hence, pretraining on global data yields better foundation models.

In addition, we introduce the task of geo-localization—based on datasets such as XM3600 [58], Dollar Street [49], GeoDE [47] and GLDv2 [66]—as a novel evaluation metric for cultural diversity. We show that, unlike, for example, XM3600 retrieval that evaluates multilinguality, geo-localization is strongly dependent on the global composition of the dataset used during pretraining.

In line with Alabdulmohsin et al. [2], we focus on contrastive learning for several reasons. These models have a wide range of applications, e.g. zero-shot classification and cross-modal retrieval, and are being increasingly adopted in critical domains like healthcare [76, 52], and as a backbone for other models [73, 72, 41]. We study SigLIP [75], the currently best-performing and a widely used CLIP-

style model. SigLIP and CLIP operate on the same principle of aligning representations/embeddings for texts and images and the difference is only in the choice of the loss function.

2 Preliminaries

SigLIP Overview. Given a mini-batch $\mathcal{B} = \{(I_1, T_1), (I_2, T_2), \dots\}$ of image–text pairs, an image encoder $f(\cdot)$ and a text encoder $g(\cdot)$, SigLIP aims to align the image embeddings $\mathbf{x}_i = f(I_i)/\|f(I_i)\|_2$ in the given batch with their corresponding text embeddings $\mathbf{y}_i = g(T_i)/\|g(T_i)\|_2$. The sigmoid-based loss $L(\mathcal{B})$ processes every image–text pair independently with a positive label $z_{ii} = 1$ for the matching pairs (I_i, T_i) and a negative label $z_{ij} = -1$ for all other pairs $(I_i, T_{i\neq j})$:

$$L(\mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \sum_{j=1}^{|\mathcal{B}|} \log\left(1 + \exp\left(z_{ij}(-t\mathbf{x}_i \cdot \mathbf{y}_j + b)\right)\right), \tag{1}$$

where $t > 0$ and b are learnable temperature and bias parameters.We follow Zhai et al. [75] in most aspects: we use a Vision Transformer (ViT) [18] for images and a Transformer [60] for text, both of size Base (B), with an embedding dimension of 768. Images are resized to 256×256 resolution and we use 16×16 patch size. Text input is tokenized using a model trained on mC4 [69], a Common Crawl-based dataset covering 101 languages, with a vocabulary size of 250 k. We keep a maximum of 64 tokens. We use the modified Adafactor optimizer [54, 74] for all our experiments with an initial learning rate of 10^{-3}, reverse square root decay, weight decay of 10^{-4} and 50 k warmup and cooldown steps. Training batch size is 16 k. Models are developed in the Big Vision codebase [6] using Tensor Processing Units (TPUs) [32]. Each model is trained on 10 B image–text pairs (roughly 610 k steps) for about 40 k TPUv2 core-hours, so they are compared on a *compute-matched* regime.

Pretraining Data. We base our analysis on a range of models trained on different subsets of the WebLI dataset, a high-volume image–text dataset collected from the public web [15]. Each example in our filtered subsets contains a caption in the original language as well as an English translation if the caption is not in English. Hence, we can distinguish between three different dataset variants: (1) **globe**, (2) **en**, and (3) **globe-tl**. Here, **globe** denotes the raw, *multilingual* data with minimal filtering applied (e.g., removing sensitive and personally identifiable information [15]). We denote its subset that contains only English captions by **en**. This mirrors the filtering that is currently applied in several influential papers, including CLIP, ALIGN and SigLIP, as well as the common way [22, 21, 67, 56, 57] of using LAION [51] and DataComp [22]. The third and last variant is **globe-tl**, which consists of the same images as **globe** but with an added pre-processing step in which any non-English text is *machine-translated* to English. We use this variant to differentiate between the impact of multilinguality, such as low-resource languages being severely underrepresented [8], and cultural diversity of the trained models. Since it is plausible that VLMs may not leverage their full multilingual potential when prompted in non-English languages, as has been observed for LLMs [20], we report results for both **globe** and **globe-tl** in all experiments and highlight any notable differences. To determine statistical significance of our results, we train 3 models each for **en**, **globe** and **globe-tl** with different random seeds. We perform two-sample t-tests and report the 95% confidence intervals.

Evaluation Data. To evaluate cultural diversity, we use five datasets: Dollar Street [49], GeoDE [47], GLDv2 [66], XM3600 [58] and MaRVL [39]. These datasets satisfy the criteria of being of sufficiently high quality and collected with geographical diversity in mind, see Figure 2. In addition, they can readily be used for evaluating contrastive VLMs without a decoder, by supporting either zero- or few-shot classification or cross-modal retrieval.

The Dollar Street (DS) [49] dataset encompasses 38 479 images depicting 289 household items commonly found in everyday settings across 63 countries. Each image is tagged with object descriptors and demographic data such as region, country, and monthly income. Following Rojas et al. [49], we map 96 topics to ImageNet classes, resulting in a subset of 21 536 images. This subset is split into training and testing sets (17 228 and 4 307 images, respectively). Notably, object tags (including the fuzzy matching to ImageNet labels) are not mutually exclusive. We use this dataset for zero-shot object classification and geo-localization. DS has been released under the CC BY-SA 4.0 license.

We also evaluate our models on a geographically diverse subset of the Google Landmarks Dataset v2 (GLDv2) [66], featuring 1 542 images representing 884 landmarks from 84 countries [35]. These

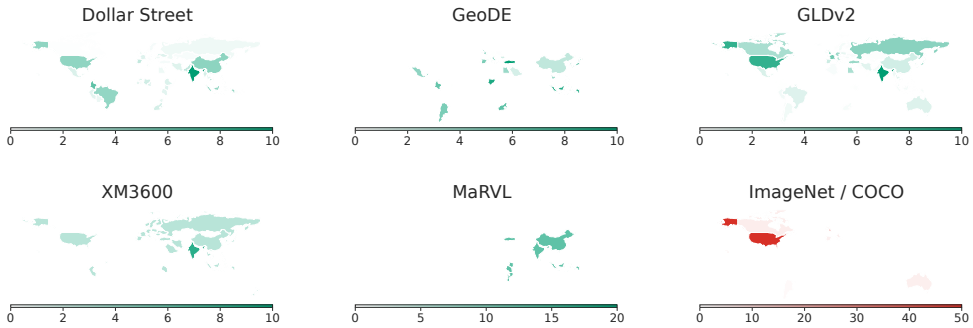


Figure 2: Data distribution [%] for each of the evaluation datasets, only approximate in MaRVL [39] based on the 5 languages collected in the dataset. Dollar Street [49], GeoDE [47], GLDv2 [66] and XM3600 [20] are geographically diverse. MaRVL is included because it focuses on underrepresented regions, such as Asia and East Africa. By comparison, ImageNet examples are mostly from a few Western countries (see for instance [53]). COCO has a nearly identical distribution to ImageNet [16].

images serve as the public and private test datasets for the 2021 Google Landmark Retrieval Challenge, providing a benchmark for evaluating landmark recognition algorithms [34]. We use it for zero-shot landmark classification using the English names of all 884 landmarks as class labels. GLDv2 images have CC-0 or Public Domain licenses. Annotations are licensed by Google LLC under CC BY 4.0.

GeoDE [47] is a geographically diverse dataset comprising 61 940 manually annotated images categorized into 40 classes. The dataset emphasizes object classes across six world regions: Europe, Americas, West Asia, East Asia, Africa, and Southeast Asia. Images were collected through crowd-sourcing, with rigorous manual verification procedures implemented to ensure data integrity. We use GeoDE for zero-shot object classification as well as for geo-localization. Since GeoDE does not have train and test splits, we shuffle the data and take the first 20 000 images for training and the rest for testing in the few-shot geo-localization evaluations. The dataset is licensed under CC BY 4.0.

The Multicultural Reasoning over Vision and Language (MaRVL) dataset [39] presents an ImageNet-style concept hierarchy designed to encompass a broader linguistic and cultural spectrum. The dataset incorporates five languages: Indonesian, Mandarin Chinese, Swahili, Tamil, and Turkish. Both the conceptual categories and associated images are curated exclusively by native speakers of each respective language. While the primary task on MaRVL involves validating statements concerning image pairs, the dataset holds potential for broader applications, including single-image evaluation metrics. The MaRVL texts and features are distributed under the CC BY 4.0 license. Image access is provided only for (non-commercial) research purposes.

Finally, Crossmodal-3600 (XM3600) [58] is a multilingual evaluation dataset that comprises 3 600 images accompanied by 261 375 human-generated reference captions spanning 36 languages. The dataset is sourced from the Open Images Dataset [36], with 100 images per language. Quality assurance measures, including a post-annotation verification process, attest to the overall high quality of the captions. We report both image-caption retrieval and geo-localization results. In the few-shot geo-localization evaluation, since XM3600 does not have train and test splits, we randomly shuffle the data and use the first 1 800 images for training and the second 1 800 images for testing. The annotations are licensed under the CC BY 4.0 license.

Summary of Findings. Before digging into the detailed results, we summarize our key findings:

1. The currently predominant paradigm of directly or indirectly filtering the training data to English image–text pairs negatively impacts cultural diversity and disproportionately hurts communities of lower socioeconomic status, exacerbating existing disparities. Its impact is demonstrably captured by zero-shot classification accuracy on Dollar Street, GLDv2, GeoDE, and MaRVL (Section 3.1, Table 1, Figure 3). This has been known for vision datasets, such as ImageNet and OpenImages [53], but is relatively less explored in image-text pretraining data scraped from the Web.
2. The image features learned by models trained on such filtered data are less culturally diverse. We demonstrate and quantify this by introducing the few-shot geo-localization task. A

Table 1: Filtering training data to English image–text pairs negatively impacts cultural diversity but improves performance on standard benchmarks. Asterisk (*) denotes statistical significance at the 95% confidence level. No statistically significant differences are observed for XM3600 retrieval.

	en	**globe**	**globe-tl**	**en** vs. **globe-tl**
Culturally diverse zero-shot evaluations				
Dollar Street	48.52 ±0.53%	48.82 ±0.34%	49.96 ±0.71%	+1.44%*
GLDv2	43.84 ±0.52%	46.18 ±1.30%	49.46 ±1.17%	+5.62%*
GeoDE	91.82 ±0.39%	92.00 ±0.10%	92.84 ±0.05%	+1.02%*
MaRVL Concepts	68.30 ±0.50%	69.09 ±0.28%	69.96 ±0.35%	+1.66%*
Crossmodal-3600 (XM3600) retrieval top-1 recall				
English captions				
Image → Text	50.60 ±1.54%	49.10 ±0.28%	49.74 ±1.03%	−0.86%
Text → Image	47.74 ±2.23%	45.01 ±0.21%	44.48 ±1.83%	−3.26%
Native captions translated to English				
Image → Text	62.73 ±1.28%	60.70 ±1.38%	62.02 ±1.41%	−0.71%
Text → Image	56.49 ±1.28%	52.60 ±1.21%	54.13 ±1.28%	−2.36%
Prevalent Western-oriented benchmarks				
0-shot ImageNet	70.36 ±0.28%	66.81 ±0.18%	68.23 ±0.19%	−2.13%*
COCO I→T R@1	59.28 ±0.90%	55.81 ±0.59%	54.00 ±1.39%	−5.28%*
COCO T→I R@1	42.91 ±0.56%	38.09 ±0.58%	37.78 ±0.23%	−5.13%*

linear probe on the image encoder shows **en** trained image encoders to be significantly less "world-knowledgeable" than **globe** or **globe-tl** trained image encoders (Section 3.2, Table 2).

3. These performance disparities are not reflected by, and even at odds with, the currently most popular (and often sole reported) benchmarks based on ImageNet [17] and COCO [38]. In addition, we present evidence that benchmarks used to evaluate multilinguality such as XM3600 [58], are *insufficient* to evaluate models' cultural diversity (Section 3.3, Table 1).
4. As a potential way out of this conundrum, we find that pretraining on unfiltered (global) data followed by fine-tuning on English-only data improves cultural diversity *without* sacrificing performance on the popular Western-centric benchmarks. This allows practitioners and researchers to strike a balance between these otherwise competing metrics. These performance improvements across benchmarks are further enhanced by translating training data to English (Section 3.4, Figure 4, Figure 5).

In summary, we call for a stop of pretraining on directly or indirectly English-filtered data.

3 Detailed Results

3.1 No filter for improved cultural diversity

To assess models' cultural diversity, we report zero-shot classification accuracy on Dollar Street, GLDv2, GeoDE and MaRVL. This evaluation includes tasks such as recognizing common household items across different countries and income brackets (Dollar Street and GeoDE), identifying significant landmarks and places of worship (GLDv2) and categorizing image concepts (MaRVL). Detailed results are provided in Table 1. Pretraining on globally diverse data yields substantial enhancements across all zero-shot classification metrics related to cultural diversity. However, these improvements stand in contrast to the popular benchmarks of ImageNet zero-shot accuracy and COCO retrieval scores. Given their prominence, it is not surprising that filtering training data to English image–text pairs, directly or indirectly, has quickly established itself as the preferred choice [46, 31, 75, 22, 21, 67, 56, 57]. When considering the cultural and geographical diversity of the resulting models however, a very different picture presents itself with models trained on **globe-tl** and **globe** outperforming those trained on **en** across all four benchmarks by a significant margin.

Indeed, a more fine-grained analysis of zero-shot classification on Dollar Street confirms that filtering to English-only training data disproportionately hurts low-income and non-Western communities;

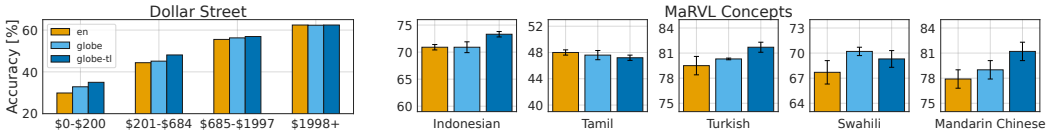


Figure 3: Filtering to English-only data further exacerbates existing performance disparities across socioeconomic subgroups. LEFT: Zero-shot classification results for Dollar Street, disaggregated by income level (x-axis). The performance difference between **en** and **globe-tl** is larger for lower-income households. Also, the performance disparity between the lowest and highest income groups is 32.5% in **en** (from 29.9% in $0-200 income group to 62.4% in $1998+ income group), but this gap is reduced (improved) to 27.4% in **globe-tl**. RIGHT: MaRVL Concepts classification accuracy disaggregated by each of the five languages/regions: Pretraining on **globe-tl** improves performance for Indonesian, Turkish and Mandarin Chinese and yields a similar performance to **en** for Tamil and Swahili.

see Figure 1 (right) and Figure 3 (left). In MaRVL, performance for all regions (except TA) tend to benefit from using globally diverse training data, as shown in Figure 3 (right). The difference in classification accuracy for different income groups and geographic regions is a worrying signal of the inherent biases of filtering to English-only data. Removing the English language filter unsurprisingly leads to a drop in performance for images from Western countries, as exemplified by ImageNet and COCO benchmarks, but it significantly improves performance for the rest of the world.

3.2 Few-shot geo-localization

The disparities between models' cultural diversity become even more pronounced when considering the image encoder's few-shot geo-localization performance, which we introduce in this work. This task involves learning to predict the geographical origin of an image, be it at the country or regional level, with only a limited number of training samples per location. We do this using a linear classification probe on the image representation, employing squared loss and L2 regularization – a problem that admits a closed-form solution. Our training dataset comprises a constrained number of images per location, with results reported for sample sizes of 5, 10, or 25 instances. Should the available examples for a given location in the training set fall below this threshold, we utilize all accessible samples. Our results (Table 2) suggest that few-shot geo-localization holds promise as a novel metric to assess cultural diversity in VLMs.

Independent of the prediction target being more fine-grained (country) or more coarse-grained (region), both **globe** and **globe-tl** have a significant edge over **en**, as shown in Table 2. These results suggest that models that are not trained on sufficiently diverse and global data fail to learn features that capture country- or region-specific information. Another point that is noteworthy, especially when comparing Table 1 to Table 2, is that the difference in performance between the **globe** and

Table 2: Global data improves few-shot geo-localization performance significantly. Performance differences are statistically significant for all reported results at 95% confidence level (*).

Task	Shots	en	globe	globe-tl	en vs. globe-tl
Dollar Street (country)	5	11.33 ±0.22%	14.16 ±0.23%	12.81 ±0.60%	+1.48%*
	10	17.45 ±0.33%	21.51 ±0.56%	20.42 ±0.61%	+2.97%*
	25	24.40 ±0.38%	30.11 ±0.50%	29.17 ±0.27%	+4.77%*
XM3600 (country)	5	14.35 ±0.36%	19.04 ±0.52%	19.24 ±0.46%	+4.89%*
	10	18.56 ±0.17%	25.83 ±0.50%	25.98 ±0.94%	+7.42%*
	25	25.96 ±0.53%	34.76 ±0.90%	33.85 ±0.28%	+7.89%*
GeoDE (country)	5	12.66 ±0.37%	19.59 ±0.63%	19.91 ±1.37%	+7.25%*
	10	16.41 ±0.56%	26.29 ±0.38%	26.23 ±0.81%	+9.82%*
	25	23.24 ±0.26%	37.13 ±0.51%	36.54 ±0.27%	+13.30%*
GeoDE (region)	5	28.18 ±1.22%	33.32 ±0.37%	34.51 ±1.90%	+6.33%*
	10	32.03 ±1.01%	40.48 ±0.34%	41.09 ±1.53%	+9.06%*
	25	38.86 ±0.75%	49.61 ±0.73%	50.38 ±1.36%	+11.53%*

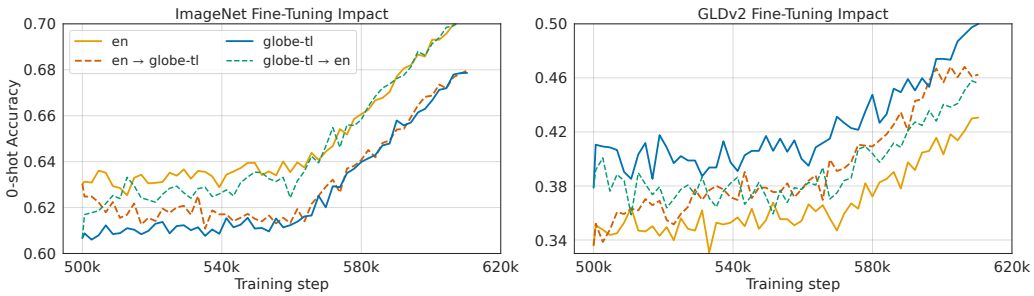


Figure 4: Fine-tuning **globe-tl** on **en** quickly catches up with **en** for ImageNet zero-shot evaluation while also performing better on GLDv2. Conversely, fine-tuning **en** on **globe-tl** does not suffice to close the gap in performance on culturally diverse benchmarks.

globe-tl models is notably smaller in the few-shot setting. This is not surprising since the standard zero-shot classification prompt templates and object or landmark names are all in English and, hence, this more closely resembles the **globe-tl** training data version and does not leverage the multilingual capabilities of the **globe** model. In the few-shot setting, by contrast, we only use the image encoder's representations and hence the impact of the text tower on the evaluation results is reduced.

3.3 Decoupling multilinguality and cultural diversity

As shown in Table 1, models trained on English-only data (**en**) perform best on Western-oriented benchmarks. New benchmarks such as XM3600 have recently been introduced to evaluate multilinguality in VLMs. To recall, XM3600 contains 100 images each from 36 different linguistic regions, captioned by native speakers in all 36 languages. At first, it might seem that performing image–text retrieval based on the English captions or the English translation of the native captions for all 3 600 images could serve as a viable signal for cultural diversity.

However, when comparing our models' performance on these two tasks, we do not find any statistically significant differences between the three models. Unsurprisingly, the **globe** model performs best when performing retrieval on *non-English* captions since the other two models have only seen English texts during training. However, when evaluating all models on the English-language captions or the English translations of captions in other languages, there are no statistically significant differences between the three model variants. We hypothesize that this is because XM3600 is derived from the Open Images dataset, which contains primarily Western images [53] or, images quite similar to what is already available in English domains. Closer inspection of the XM3600 images confirms that most images from non-Western countries are likely taken by tourists. For example, among the 100 images from the Arab world, 12% are images of cars. These images do not adequately reflect cultural differences. Moreover, since the original caption language is often English, similar images are likely to have been included in the **en** training data, explaining why there are no statistically significant differences between the three models. When comparing retrieval scores between English captions and English translations of captions in other languages across all three models, we observe a significant difference. Upon further investigation, we found that non-English captions tend to provide more detailed descriptions of the target image, leading to higher retrieval accuracy. This discrepancy might not be associated with cultural diversity or multilinguality but rather reflects variance in annotators.

Based on these findings, we argue that datasets originally created for evaluating multilinguality, such as XM3600, might not be sufficient for evaluating cultural diversity in multimodal systems.

3.4 Bridging the gap

Fine-tuning. As shown in Table 1, improving diversity generally results in a loss of performance on standard benchmarks. In general, we found that the two objectives typically *compete* with each other: in Figure 4, where we take two models pretrained on either **en** or **globe-tl** data and fine-tune them on the other dataset for a short duration. Clearly, improving culturally diverse metrics is accompanied by a loss in performance on Western-oriented benchmarks and vice versa. This is even more clear when looking at the correlation coefficients across metrics of over 40 models, shown in Figure 5 (b).

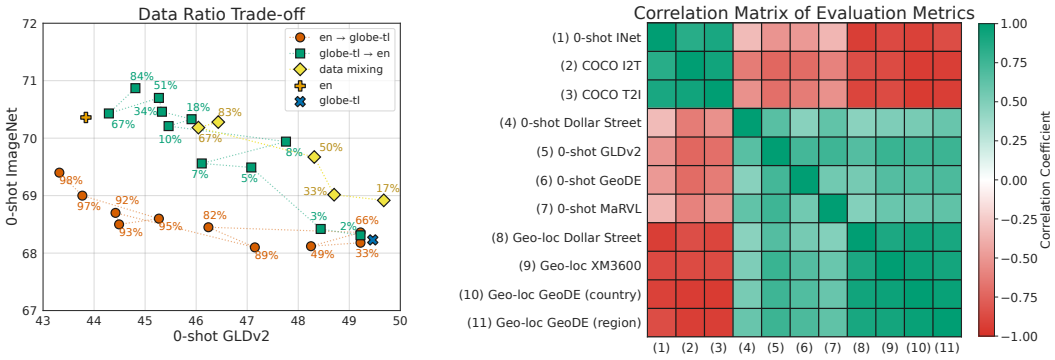


Figure 5: LEFT: Fine-tuning allows for a controlled trade-off between cultural diversity and performance on standard benchmarks. Fine-tuning **globe-tl** on **en** is strictly better than fine-tuning **en** on **globe-tl**, but mixing training data in different proportions achieves a better trade-off overall. Values in percentages [%] correspond to the fraction of time training is restricted to **en**data. RIGHT: Correlation coefficients of the evaluation metrics computed based on over 40 fully trained models.

In Figure 5 (a), this trade-off is further visualized. We first pretrain models on either **en** or **globe-tl** for a set number of steps (varying between $100k$ and $600k$) and we then switch to the other dataset for the remainder of the training duration. Each model is trained for approximately $610k$ steps in total. We plot the resulting performance at the end of training for each of these switch-over points. In line with what we see in Figure 4, we find that very little fine-tuning is sufficient to significantly impact model performance and improve performance either for standard benchmarks (green squares) or for the cultural diversity evaluations (red circles), depending on which dataset was used at the end. Hence, if we fine-tune for more than $100k$ steps, final model performance is mostly determined by the dataset that the model is fine-tuned on.

Notably, it is possible to improve performance on cultural understanding (e.g. improving 0-shot accuracy in GLDv2 from 44% in **en** to over 45.5%) *without* impacting ImageNet 0-shot accuracy by pretraining on **globe-tl** and fine-tuning that model on **en** data. However, the opposite does not hold. English-only pretrained models achieve lower performance on culturally diverse benchmarks even after fine-tuning on global data with the tradeoff curve being clearly suboptimal to the one observed when fine-tuning culturally diverse models on **en** data. Therefore, pretraining on globally diverse data before fine-tuning on English-only image–text pairs allows for a nuanced trade-off between catering to cultural diversity and achieving strong performance on well-established benchmarks.

Data mixing. Besides fine-tuning, we also study the impact of mixing the two versions of data during pretraining. Because **en** is a subset of **globe-tl**, mixing is equivalent to assigning more weight to English data. Figure 5 highlights that data mixing is as good as fine-tuning (if not better) in achieving a balance between Western-oriented and culturally diverse benchmarks. Moreover, it is applicable in settings where we may have more than 2 different data splits. However, data mixing entails training new models *ab initio*, thus incurring higher computational cost compared to fine-tuning. In our setting, we observe that fine-tuning for as few as 50k steps is often sufficient. By contrast, training a new model from scratch is more than 12 times as expensive. Table 3 provides a detailed comparison between data mixing and fine-tuning for similar mixing ratios.

To conclude, both fine-tuning models pretrained on **globe-tl** as well as choosing an appropriate data mix during training can be viable approaches to navigate the trade-off between cultural diversity and optimizing performance on Western-oriented, but well-studied benchmarks, such as ImageNet.

3.5 Quality Filters

We apply quality filtering using an internal model trained on image-text pairs to calculate the image-text similarity score, in order to assess if our empirical findings continue to hold in this context. The model was trained on global data to make sure it accurately assesses quality for global data, of which English data is a substantial fraction, and its threshold was tuned to balance quality and quantity. We filter out about 60% of the data in our experiments. Appendix B shows that our main

findings continue to hold even when quality filters are applied. For instance, quality-filtered **globe-tl** performs better than quality-filtered **en** on 0-shot Dollar Street, GLDv2, and GeoDE but performs worse on Western-oriented benchmarks, such as ImageNet and COCO retrieval. In addition, the improvement in quality-filtered **globe-tl** over quality-filtered **en** is particularly significant for few-shot geo-localization tasks.

4 Related Work

A range of prior work has studied biases in zero-shot classifiers related to a range of sensitive attributes including gender, race and age [1, 25, 27, 12, 24] and found that CLIP models perpetuate biases present in the training data [7, 23, 2]. More recent work extends this analysis to zero-shot performance across groups of different income levels and geographic regions [43, 78]. While several of these papers highlight the central role of training data and the potentially large impact of design choices such as the data source or filtering techniques [22] or translation of English captions for cross-modal multilingual encoders [10, 11, 45], to our knowledge, we are the first to study the impact of image–text pair filtering and text translation on cultural understanding in contrastive VLMs.

Contrastive models are usually evaluated on a range of benchmark datasets including ImageNet [50] (and variations [29, 5]), COCO [38] and Flickr30K [71]. These datasets have been shown to reflect a heavy Western bias [53, 16, 55]. Over the last few years, a range of alternative benchmarks have been proposed, including DollarStreet [49], GeoDE [47], Crossmodal-3600 [58], GLDv2 [66], GeoNet [33], Geo-YFCC [19], xGQA [44], MaXM [13], Ego4D [26] and GD-VCR [70]. While we have used some of these in our experiments, we decided against using others for the following reasons. GeoNet uses only images from North America and Asia and is hence not sufficiently diverse. GeoYFCC contains images from 62 different countries, but Europe-centric and based on images with noisy tags [47]. xGQA mainly evaluates multilinguality: the starting point for its creation is a monolingual English dataset that was then translated. MaXM is an adaptation of XM3600 (which we already use) to multilingual VQA, which contrastive VLMs do not support. Ego4D contains images from only 9 countries with a majority of the images from English-speaking countries (US, UK). GD-VCR is a geo-diverse commonsense benchmark, but is a VQA task unsuited for contrastive VLMs.

The closest work to ours is [48], which argues that progress in global data (Dollar Street and GeoDE) has been much slower than the progress on ImageNet. Unlike their work, however, we study the impact of the training data mixture, study the impact of translation, suggest a setup where both types of metrics can be improved, and propose geo-localization as an evaluation metric. We also consider a broader set of datasets in our study, such as XM3600, GLDv2, and MaRVL.

5 Limitations and Future Work

While our work highlights the importance of incorporating cultural and socioeconomic diversity considerations into contrastive VLMs, several limitations should be acknowledged. Firstly, our experimental results are based on recently popular contrastive, encoder-only SigLIP models. While our analysis offers valuable insights, it should be extended to generative VLMs [40, 37, 77, 3, 14, 59, 61, 63, 4, 68, 65, 64, 62]. Secondly, our work primarily highlights the importance of utilizing all available data when pretraining foundation models, but there is potential for additional measures to further improve cultural diversity, such as via regularization, data balancing, or weight averaging [30].

Table 3: A comparison between data mixing and fine-tuning using identical ratios of **en** vs. **globe-tl** examples; e.g., **en** $5:1$ **globe-tl** for **en** $\rightarrow$ **globe-tl** means that the model was pretrained on **en** data for 508k steps and then fine-tuned on **globe-tl** data for approximately 102k steps.

	0-shot accuracy on GLDv2			0-shot accuracy on ImageNet		
Proportions	**Data mixing**	**en $\rightarrow$ globe-tl**	**globe-tl $\rightarrow$ en**	**Data mixing**	**en $\rightarrow$ globe-tl**	**globe-tl $\rightarrow$ en**
en $5:1$ **globe-tl**	46.43%	46.24%	44.81%	70.28%	68.45%	70.87%
en $2:1$ **globe-tl**	46.04%	49.22%	44.29%	70.18%	68.36%	70.43%
en $1:1$ **globe-tl**	48.31%	48.25%	45.27%	69.67%	68.12%	70.65%
en $1:2$ **globe-tl**	48.70%	49.22%	45.33%	69.02%	68.18%	70.46%
en $1:5$ **globe-tl**	49.68%	49.81%	45.91%	68.92%	67.92%	70.33%

Thirdly, although our study consciously separates cultural diversity and multilinguality, investigating their intersection presents an intriguing subject for future research. Fourthly, acknowledging the vagueness of the notion of culture and cultural diversity, we recognize that our experiments are mainly comparing model performance across different countries, regions, or income groups. We do not offer a precise definition of cultural diversity in the context of VLMs and do not claim to cover all aspects of cultures in our analysis. Lastly, we do not address the relationship between cultural diversity and social biases that have been shown to be perpetuated by VLMs. Exploring these connections presents an opportunity to develop more inclusive AI systems.

6 Conclusion

This work highlights the importance of considering cultural diversity when training contrastive VLMs. We recommend that researchers and practitioners move away from training models on English-only image–text pairs. While this approach may seem beneficial when considering performance on popular benchmarks, such as ImageNet and COCO, it discards a vast amount of valuable and culturally diverse training information and disproportionately hurts communities of lower socioeconomic status. Our findings suggest that (i) pretraining on the full dataset followed by *short* fine-tuning on English-only data, or (ii) pretraining on a mixture of data, achieve good performance on standard benchmarks while also promoting cultural awareness. When doing so, it is important to acknowledge that there seems to be a trade-off between optimized performance on standard benchmarks and maintaining cultural diversity. Practitioners should, hence, carefully consider the intended use case and the importance of cultural understanding of the resulting model when deciding their pretraining data mixture, also taking into account unintended biases possibly manifesting in downstream models and applications. In specific scenarios, where downstream use is limited to English and multilingualism is not required, we have found that translating the training data into English is a viable option. However, the latter approach should be applied judiciously to avoid sacrificing valuable cultural context within the data.

Acknowledgement

The authors would like to thank Michael Tschannen and Jeremiah Harmsen from Google DeepMind for their feedback on earlier drafts of this manuscript, as well as Tobias Weyand and Maribeth Rauh from Google DeepMind for the helpful discussions.

References

[1] Agarwal, S., Krueger, G., Clark, J., Radford, A., Kim, J. W., and Brundage, M. (2021). Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*. 9

[2] Alabdulmohsin, I., Wang, X., Steiner, A. P., Goyal, P., D'Amour, A., and Zhai, X. (2023). Clip the bias: How useful is balancing data in multimodal learning? In *The Twelfth International Conference on Learning Representations*. 2, 9

[3] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., and Simonyan, K. (2022). Flamingo: a visual language model for few-shot learning. 9

[4] Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. (2023). Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. 9

[5] Beyer, L., Hénaff, O. J., Kolesnikov, A., Zhai, X., and Oord, A. v. d. (2020). Are we done with imagenet? *arXiv preprint arXiv:2006.07159*. 9

[6] Beyer, L., Zhai, X., and Kolesnikov, A. (2022). Big vision. https://github.com/google-research/big_vision. 3, 16, 19

[7] Birhane, A., Prabhu, V. U., and Kahembwe, E. (2021). Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*. 2, 9

[8] Blasi, D., Anastasopoulos, A., and Neubig, G. (2021). Systematic inequalities in language technology performance across the world's languages. *arXiv preprint arXiv:2110.06733*. 3

[9] Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., et al. (2018). Jax: composable transformations of python+ numpy programs. 16

[10] Bugliarello, E., Cotterell, R., Okazaki, N., and Elliott, D. (2021). Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language berts. *Transactions of the Association for Computational Linguistics*, 9:978–994. 9

[11] Bugliarello, E., Liu, F., Pfeiffer, J., Reddy, S., Elliott, D., Ponti, E. M., and Vuli'c, I. (2022). IGLUE: A benchmark for transfer learning across modalities, tasks, and languages. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, page 2370–2392, Balitmore, MA. PMLR. 9

[12] Cabello, L., Bugliarello, E., Brandl, S., and Elliott, D. (2023). Evaluating bias and fairness in gender-neutral pretrained vision-and-language models. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8465–8483, Singapore. Association for Computational Linguistics. 9

[13] Changpinyo, S., Xue, L., Szpektor, I., Thapliyal, A. V., Amelot, J., Chen, X., and Soricut, R. (2022). Towards multi-lingual visual question answering. *arXiv preprint arXiv:2209.05401*. 9

[14] Chen, X., Wang, X., Beyer, L., Kolesnikov, A., Wu, J., Voigtlaender, P., Mustafa, B., Goodman, S., Alabdulmohsin, I., Padlewski, P., Salz, D., Xiong, X., Vlasic, D., Pavetic, F., Rong, K., Yu, T., Keysers, D., Zhai, X., and Soricut, R. (2023). Pali-3 vision language models: Smaller, faster, stronger. 9

[15] Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., et al. (2022). Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*. 3, 16

[16] 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*, pages 52–59. 1, 4, 9

[17] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *CVPR*. 1, 5, 16

[18] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. 3, 16

[19] Dubey, A., Ramanathan, V., Pentland, A., and Mahajan, D. (2021). Adaptive methods for real-world domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14340–14349. 9

[20] Etxaniz, J., Azkune, G., Soroa, A., de Lacalle, O. L., and Artetxe, M. (2023). Do multilingual language models think better in english? *arXiv preprint arXiv:2308.01223*. 3, 4

[21] Fang, A., Jose, A. M., Jain, A., Schmidt, L., Toshev, A., and Shankar, V. (2023). Data filtering networks. *CoRR*, abs/2309.17425. 3, 5

[22] Gadre, S. Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al. (2024). Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36. 3, 5, 9

[23] Garcia, N., Hirota, Y., Wu, Y., and Nakashima, Y. (2023). Uncurated image-text datasets: Shedding light on demographic bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6957–6966. 9

[24] Goyal, P., Duval, Q., Seessel, I., Caron, M., Misra, I., Sagun, L., Joulin, A., and Bojanowski, P. (2022a). Vision models are more robust and fair when pretrained on uncurated images without supervision. 9

[25] Goyal, P., Soriano, A. R., Hazirbas, C., Sagun, L., and Usunier, N. (2022b). Fairness indicators for systematic assessments of visual feature extractors. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 70–88. 9

[26] Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al. (2022). Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012. 9

[27] Hall, M., Gustafson, L., Adcock, A., Misra, I., and Ross, C. (2023). Vision-language models performing zero-shot tasks exhibit gender-based disparities. *arXiv preprint arXiv:2301.11100*. 2, 9

[28] Heek, J., Levskaya, A., Oliver, A., Ritter, M., Rondepierre, B., Steiner, A., and van Zee, M. (2020). Flax: A neural network library and ecosystem for jax. *Version 0.3*, 3:14–26. 16

[29] Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. (2021). Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271. 9

[30] Ilharco, G., Wortsman, M., Gadre, S. Y., Song, S., Hajishirzi, H., Kornblith, S., Farhadi, A., and Schmidt, L. (2022). Patching open-vocabulary models by interpolating weights. *Advances in Neural Information Processing Systems*, 35:29262–29277. 9

[31] Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR. 1, 5

[32] Jouppi, N. P., Young, C., Patil, N., Patterson, D. A., Agrawal, G., Bajwa, R., Bates, S., Bhatia, S., Boden, N., Borchers, A., Boyle, R., Cantin, P., Chao, C., Clark, C., Coriell, J., Daley, M., Dau, M., Dean, J., Gelb, B., Ghaemmaghami, T. V., Gottipati, R., Gulland, W., Hagmann, R., Ho, C. R., Hogberg, D., Hu, J., Hundt, R., Hurt, D., Ibarz, J., Jaffey, A., Jaworski, A., Kaplan, A., Khaitan, H., Koch, A., Kumar, N., Lacy, S., Laudon, J., Law, J., Le, D., Leary, C., Liu, Z., Lucke, K., Lundin, A., MacKean, G., Maggiore, A., Mahony, M., Miller, K., Nagarajan, R., Narayanaswami, R., Ni, R., Nix, K., Norrie, T., Omernick, M., Penukonda, N., Phelps, A., Ross, J., Salek, A., Samadiani, E., Severn, C., Sizikov, G., Snelham, M., Souter, J., Steinberg, D., Swing, A., Tan, M., Thorson, G., Tian, B., Toma, H., Tuttle, E., Vasudevan, V., Walter, R., Wang, W., Wilcox, E., and Yoon, D. H. (2017). In-datacenter performance analysis of a tensor processing unit. *CoRR*, abs/1704.04760. 3

[33] Kalluri, T., Xu, W., and Chandraker, M. (2023). Geonet: Benchmarking unsupervised adaptation across geographies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15368–15379. 9

[34] Kim, Z., Araujo, A., Cao, B., Askew, C., Sim, J., Green, M., Yilla, N., and Weyand, T. (2021). Towards a fairer landmark recognition dataset. *arXiv preprint arXiv:2108.08874*. 4

[35] Kim, Z., Araujo, A., Cao, B., Askew, C., Sim, J., Green, M., Yilla, N., and Weyand, T. (2022). Improving fairness in large-scale object recognition by crowdsourced demographic information. *arXiv preprint arXiv:2206.01326*. 3

[36] 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. 4

[37] Li, J., Li, D., Savarese, S., and Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. 9

[38] 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 *ECCV*. 1, 5, 9, 16

[39] Liu, F., Bugliarello, E., Ponti, E. M., Reddy, S., Collier, N., and Elliott, D. (2021). Visually grounded reasoning across languages and cultures. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *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. 3, 4, 16

[40] Lu, J., Clark, C., Lee, S., Zhang, Z., Khosla, S., Marten, R., Hoiem, D., and Kembhavi, A. (2023). Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. 9

[41] Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., et al. (2022). Simple open-vocabulary object detection. In *European Conference on Computer Vision*, pages 728–755. Springer. 2

[42] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229. 16

[43] Nwatu, J., Ignat, O., and Mihalcea, R. (2023). Bridging the digital divide: Performance variation across socio-economic factors in vision-language models. *arXiv preprint arXiv:2311.05746*. 2, 9

[44] Pfeiffer, J., Geigle, G., Kamath, A., Steitz, J.-M. O., Roth, S., Vulić, I., and Gurevych, I. (2021). xgqa: Cross-lingual visual question answering. *arXiv preprint arXiv:2109.06082*. 9

[45] Qiu, C., Oneață, D., Bugliarello, E., Frank, S., and Elliott, D. (2022). Multilingual multimodal learning with machine translated text. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4178–4193, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. 9

[46] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR. 1, 5

[47] Ramaswamy, V. V., Lin, S. Y., Zhao, D., Adcock, A., van der Maaten, L., Ghadiyaram, D., and Russakovsky, O. (2024). Geode: a geographically diverse evaluation dataset for object recognition. *Advances in Neural Information Processing Systems*, 36. 2, 3, 4, 9, 16

[48] Richards, M., Kirichenko, P., Bouchacourt, D., and Ibrahim, M. (2024). Does progress on object recognition benchmarks improve real-world generalization? In *ICLR*. 9

[49] Rojas, W. A. G., Diamos, S., Kini, K. R., Kanter, D., Reddi, V. J., and Coleman, C. (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*. 2, 3, 4, 9, 16

[50] 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. 9

[51] Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., and Jitsev, J. (2022). LAION-5B: an open large-scale dataset for training next generation image-text models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*. 3

[52] Sellergren, A. B., Chen, C., Nabulsi, Z., Li, Y., Maschinot, A., Sarna, A., Huang, J., Lau, C., Kalidindi, S. R., Etemadi, M., et al. (2022). Simplified transfer learning for chest radiography models using less data. *Radiology*, 305(2):454–465. 2

[53] 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*. 1, 4, 7, 9

[54] Shazeer, N. and Stern, M. (2018). Adafactor: Adaptive learning rates with sublinear memory cost. In *ICML*. 3

[55] Smith, B., Farinha, M., Hall, S. M., Kirk, H. R., Shtedritski, A., and Bain, M. (2023). Balancing the picture: Debiasing vision-language datasets with synthetic contrast sets. *arXiv preprint arXiv:2305.15407*. 9

[56] Sun, Q., Fang, Y., Wu, L., Wang, X., and Cao, Y. (2023). EVA-CLIP: improved training techniques for CLIP at scale. *CoRR*, abs/2303.15389. 3, 5

[57] Sun, Q., Wang, J., Yu, Q., Cui, Y., Zhang, F., Zhang, X., and Wang, X. (2024). EVA-CLIP-18B: scaling CLIP to 18 billion parameters. *CoRR*, abs/2402.04252. 3, 5

[58] Thapliyal, A. V., Pont-Tuset, J., Chen, X., and Soricut, R. (2022). Crossmodal-3600: A massively multilingual multimodal evaluation dataset. *arXiv preprint arXiv:2205.12522*. 2, 3, 4, 5, 9, 16

[59] Tschannen, M., Kumar, M., Steiner, A., Zhai, X., Houlsby, N., and Beyer, L. (2023). Image captioners are scalable vision learners too. In *NeurIPS*. 9

[60] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30. 3, 16

[61] Wan, B., Tschannen, M., Xian, Y., Pavetic, F., Alabdulmohsin, I., Wang, X., Pinto, A. S., Steiner, A., Beyer, L., and Zhai, X. (2024). Locca: Visual pretraining with location-aware captioners. 9

[62] Wang, J., Chen, D., Wu, Z., Luo, C., Zhou, L., Zhao, Y., Xie, Y., Liu, C., Jiang, Y.-G., and Yuan, L. (2022a). Omnivl:one foundation model for image-language and video-language tasks. 9

[63] Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., and Wang, L. (2022b). Git: A generative image-to-text transformer for vision and language. 9

[64] Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., and Yang, H. (2022c). Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. 9

[65] Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., and Cao, Y. (2022d). Simvlm: Simple visual language model pretraining with weak supervision. 9

[66] Weyand, T., Araujo, A., Cao, B., and Sim, J. (2020). Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2575–2584. 1, 2, 3, 4, 9, 16

[67] Xu, H., Xie, S., Tan, X. E., Huang, P., Howes, R., Sharma, V., Li, S., Ghosh, G., Zettlemoyer, L., and Feichtenhofer, C. (2023a). Demystifying CLIP data. *CoRR*, abs/2309.16671. 3, 5

[68] Xu, H., Ye, Q., Yan, M., Shi, Y., Ye, J., Xu, Y., Li, C., Bi, B., Qian, Q., Wang, W., Xu, G., Zhang, J., Huang, S., Huang, F., and Zhou, J. (2023b). mplug-2: A modularized multi-modal foundation model across text, image and video. 9

[69] Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2020). mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*. 3

[70] Yin, D., Li, L. H., Hu, Z., Peng, N., and Chang, K.-W. (2021). Broaden the vision: Geo-diverse visual commonsense reasoning. *arXiv preprint arXiv:2109.06860*. 9

[71] Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78. 9

[72] Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. (2022). Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*. 2

[73] Yuan, L., Chen, D., Chen, Y.-L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al. (2021). Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*. 2

[74] Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. (2022). Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113. 3

[75] Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. (2023). Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*. 1, 2, 3, 5, 16

[76] Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., et al. (2023). Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2(3):6. 2

[77] Zou, X., Dou, Z.-Y., Yang, J., Gan, Z., Li, L., Li, C., Dai, X., Behl, H., Wang, J., Yuan, L., Peng, N., Wang, L., Lee, Y. J., and Gao, J. (2022). Generalized decoding for pixel, image, and language. 9

[78] Zunair, H., Khan, S., and Hamza, A. B. (2024). Rsud20k: A dataset for road scene understanding in autonomous driving. *arXiv preprint arXiv:2401.07322*. 9

A Appendix

A.1 Model Card

Model details following Mitchell et al. [42].

- **Model Architecture**: The model architecture contains two towers, a vision transformer encoder [18] and a language transformer encoder [60], both of size B. Models are trained using a contrastive pretraining technique with sigmoid loss [75].
- **Inputs**: The vision encoder takes an image reshaped to 256×256 as input. The text encoder takes tokenized text cropped to the first 64 tokens as input.
- **Outputs**: The vision and text encoders both output a d-dimensional feature vector where $d = 768$.
- **Intended Use**: The primary use is to conduct research on multimodal applications, such as zero-shot classification and retrieval. We use the models to study the impact of training data filtering on cultural diversity.
- **Known Caveats**: As noted in several prior works, multimodal systems can pick up societal biases. While we demonstrate some of those issues in this work, our analysis is necessarily limited in scope.
- **System Description**: Models are analyzed in a stand-alone setting and not used as part of a larger system.
- **Upstream Dependencies**: None
- **Downstream Dependencies**: None
- **Hardware & Software**: Models are developed using JAX [9] and Flax [28] in the Big Vision [6] codebase. They are trained on Google Cloud TPUs.
- **Compute Requirements**: Each model is trained on 16×16 TPU chips on 10B seen image-text pairs. A typical training run takes 3.3 days.
- **Model Initialization**: The model is trained from a random initialization.
- **Model Size**: Each SigLIP model has a ViT B/16 image encoder and a size B text encoder.
- **Training Dataset**: We use different subsets of WebLI [15], which consists of images with alt-texts from the public web.
- **Evaluation Datasets**: We evaluate the models on ImageNet-ILSRCV2012 [17], MS COCO [38], Dollar Street [49], Google Landmarks Dataset v2 [66], GeoDE [47], MaRVL [39] and Crossmodal-3600 [58].

B Impact of Quality Filters

To assess the impact of quality filters on our findings, we train two models, **en** and **globe-tl**, on 1B image-text pairs, following the same training setup used in the paper. Table 4 shows a summary of these results. We observe that our empirical findings also hold in this setting. For instance, quality-filtered **globe-tl** performs better than quality-filtered **en** on 0-shot Dollar Street, GLDv2, and GeoDE but performs worse on Western-oriented benchmarks, such as ImageNet and COCO retrieval. In addition, the improvement in few-shot geo-localization for quality-filtered **globe-tl** over quality-filtered **en** is particularly significant.

Table 4: Applying quality filters to SigLIPs does not change the primary conclusions. Filtering training data to English image-text pairs continues to negatively impact cultural diversity even though it improves performance on standard benchmarks.

	en	**globe-tl**	**en** vs. **globe-tl**
Culturally diverse zero-shot evaluations			
Dollar Street	46.05%	48.28%	+2.23%
GLDv2	28.21%	30.67%	+2.46%
GeoDE	90.53%	90.57%	+0.04%
Prevalent Western-oriented benchmarks			
0-shot ImageNet	66.96%	66.32%	−0.64%
COCO I→T R@1	56.72%	52.80%	−3.92%
COCO T→I R@1	37.33%	34.46%	−2.87%
10-shot geo-localization			
Dollar Street (country)	9.40%	9.82%	+0.42%
GeoDE (country)	10.10%	14.73%	+4.63%
GeoDE (region)	21.97%	28.22%	+6.25%

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.

- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We use the SigLIP trainer as well as a range of pre-trained SigLIP models that have been released as part of the Big Vision codebase [6]. While we are unable to provide access to the training data, all evaluation datasets are publicly available.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments are executed on TPUs. A single model is trained for approximately 40K TPUv2 core-hours.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper advocates for the evaluation of VLMs on diverse datasets, aiming to potentially yield positive societal impacts by promoting more inclusivity in AI research and development.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original owners or datasets, models, and codebase used in our experiments.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.