Babel-ImageNet: Massively Multilingual Evaluation of Vision-and-Language Representations

Gregor Geigle[1,2] **Radu Timofte**[2] **Goran Glavaš**[1]
[1]WüNLP, [2]Computer Vision Lab, CAIDAS, University of Würzburg,
gregor.geigle@uni-wuerburg.de

Abstract

Vision-and-language (VL) models with separate encoders for each modality (e.g., CLIP) have become the go-to models for zero-shot image classification and image-text retrieval. They are, however, mostly evaluated in English as multilingual benchmarks are limited in availability. We introduce Babel-ImageNet, a massively multilingual benchmark that offers (partial) translations of ImageNet labels to 100 languages, built without machine translation or manual annotation. We instead automatically obtain reliable translations by linking them – via shared WordNet synsets – to BabelNet, a massively multilingual lexico-semantic network. We evaluate 11 public multilingual CLIP models on zero-shot image classification (ZS-IC) on our benchmark, demonstrating a significant gap between English ImageNet performance and that of high-resource languages (e.g., German or Chinese), and an even bigger gap for low-resource languages (e.g., Sinhala or Lao). Crucially, we show that the models' ZS-IC performance highly correlates with their performance in image-text retrieval, validating the use of Babel-ImageNet to evaluate multilingual models for the vast majority of languages without gold image-text data. Finally, we show that the performance of multilingual CLIP can be drastically improved for low-resource languages with parameter-efficient language-specific training. We make our code and data publicly available: https://github.com/gregor-ge/Babel-ImageNet

1 Introduction

CLIP models (Radford et al., 2021; Jia et al., 2021; Pham et al., 2021) have become widely used vision-and-language (VL) models, owing popularity to efficient inference based on separate yet semantically aligned encoders for the two modalities. Their bi-encoder architecture makes them ideal for efficient image-text retrieval (Lin et al., 2014; Plummer et al., 2015) and zero-shot image classification (Radford et al., 2021), or to produce downstream features for supervised tasks such as image generation (Rombach et al., 2022) or cross-modal reasoning (Eichenberg et al., 2022; Li et al., 2023).

Motivated by the observation that performance on ImageNet classification translates well to performance in many other image tasks (Recht et al., 2019; Fang et al., 2023), CLIP models are typically evaluated on zero-shot image classification (ZS-IC), i.e., by comparing the representation of an image with text representations of class labels, whereby ImageNet (Deng et al., 2009) is the most prominent benchmark. With ImageNet class labels available only in English, this supports only evaluation of monolingual English models (i.e., models trained with English captions only). Although most CLIP models are trained on English-only image-caption data, some effort has been put into creating multilingual and monolingual non-English models by (1) training them from scratch (Bianchi et al., 2021; Ilharco et al., 2021; Yang et al., 2022; Jain et al., 2021; Zhai et al., 2023) or (2) distilling them from English models (Carlsson et al., 2022; Chen et al., 2022b; Zhang et al., 2022; Visheratin, 2023), typically using parallel data as supervision. Despite attempts to translate ImageNet labels to other languages (Bianchi et al., 2021; Yang et al., 2022), the language coverage remains very limited. Because of this, multilingual CLIP models have mainly been benchmarked on image-text retrieval datasets (Aggarwal and Kale, 2020; Bugliarello et al., 2022; Thapliyal et al., 2022), which cover only limited sets of mid-to-high resource languages.

Creating *massively multilingual* datasets for VL tasks (e.g., image-text retrieval) is prohibitively expensive. Existing efforts (Aggarwal and Kale, 2020; Bugliarello et al., 2022; Thapliyal et al., 2022) either hire native speakers to write image captions in target languages or resort to machine translation (MT) of English data, followed by manual post-editing by native speakers. The MT approach

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5064–5084

(the cheaper of the two), is, we argue, still too expensive for low-resource languages because MT models are less accurate when translating to those languages, which implies a bigger post-editing effort for bilingual annotators, native in the low-resource language and fluent in English; in addition, such annotators are more difficult to find for low-resource than for high-resource target languages. In this work, we thus seek to create a robust *massively multilingual benchmark* for evaluating the quality of representation spaces of multilingual VL models, without resorting to MT or requiring any manual annotation effort. To be useful, such a benchmark needs to satisfy a crucial requirement: models' performance across languages must be indicative of their performance for the same languages in tasks such as image-text retrieval, for which creating massively multilingual (gold-standard) evaluation datasets is too expensive.

Contributions. With this in mind, we create Babel-ImageNet, a massively multilingual dataset for zero-shot image classification that offers (partial) translations of the 1000 ImageNet classes to 100 languages. To obtain robust translations of ImageNet labels in other languages, we leverage the connection between ImageNet classes, which are derived from WordNet (Miller, 1994) synsets, and BabelNet (Navigli and Ponzetto, 2010), a massively multilingual lexico-semantic network, also (in part) derived from WordNet. Relying on the multilingual BabelNet synsets (and WordNet synset identifiers of ImageNet classes) to pivot between languages, we avoid problems known to occur with machine translation of short phrases without context, e.g., due to polysemy[1]. Exploiting BabelNet allows us to automatically obtain labels for ImageNet concepts in many languages, removing the need for MT and manual annotation.

We evaluate 11 different multilingual CLIP models on Babel-ImageNet, observing that all of them exhibit poor performance for low-resource languages. Crucially, we validate that Babel-ImageNet is a meaningful benchmark for measuring the quality of multilingual VL representations by comparing models' performance on Babel-ImageNet against their performance on established multilingual image-text retrieval datasets. Babel-ImageNet thus allows us to evaluate models in languages not covered by those datasets and it additionally expands the retrieval-focused evaluation with the ZS-IC task in languages included in the established datasets. Finally, we propose a computationally efficient approach for improving multilingual CLIP models for low-resource languages. This modular language specialization approach yields large performance gains (>20% for some of the low-resource languages).

[1]For example, the ImageNet class *walking stick* refers to the *insect* and not the *inanimate object*.

2 Related Work

Multilingual Vision-and-Language Benchmarks. Early multilingual VL models (Gella et al., 2017; Wehrmann et al., 2019; Kim et al., 2020; Burns et al., 2020; Ni et al., 2021; Geigle et al., 2022; Zhou et al., 2021) were often evaluated in image-text retrieval on Multi30k (Elliott et al., 2016, 2017; Barrault et al., 2018), an extension of Flickr30k (Plummer et al., 2015) to German, French, and Czech, as well as on the Japanese (Yoshikawa et al., 2017) and Chinese (Li et al., 2019) translations of MSCOCO (Lin et al., 2014). More recent models were evaluated on multilingual image-text retrieval benchmarks: XTD (Aggarwal and Kale, 2020) (10 languages) and WIT (Srinivasan et al., 2021) (108 languages). Both these benchmarks, however, have prominent shortcomings. XTD predominantly contains examples from Karpathy's training portion of MSCOCO (Karpathy and Fei-Fei, 2017), which is commonly used for pretraining of VL models, which constitutes a case of data leakage (Bugliarello et al., 2022). WIT collects image-caption pairs from Wikipedia(s), which are abundant with named entity mentions that are often identical across a number languages – this artificially equates the difficulty of retrieval across languages (Zhai et al., 2022). More recently, IGLUE (Bugliarello et al., 2022) was introduced as the first benchmark to also include reasoning tasks like visual QA (Pfeiffer et al., 2022), primarily meant to test cross-encoder models that jointly encode image-text pairs (Ni et al., 2021; Zhou et al., 2021; Zeng et al., 2022). IGLUE also introduces the image-caption dataset xFlickrCo, a combination of Flickr30k and MSCOCO with new captions in 7 languages. Another recent dataset, XM3600 (Thapliyal et al., 2022), encompasses 3600 images (balanced by geography of origin) with captions in 36 languages written from scratch by native speakers.

Motivated by monolingual CLIP models in other languages, translations of ImageNet classes have emerged for a handful of high-resource languages:

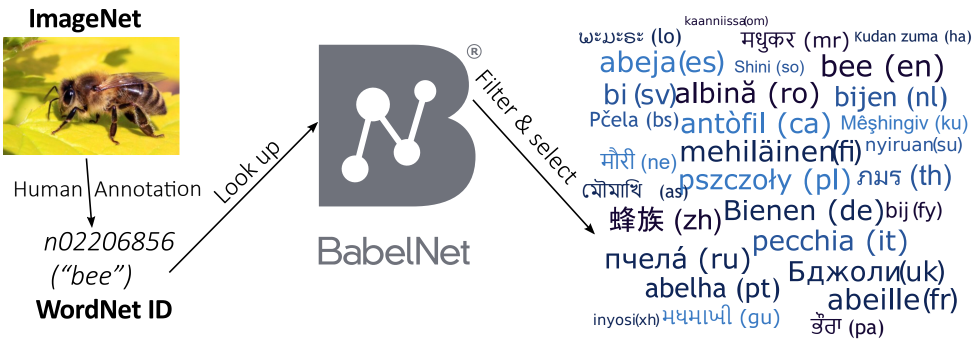


Figure 1: Illustrating the creation of Babel-ImageNet: ImageNet classes correspond to WordNet IDs, which are integrated into BabelNet, a multilingual semantic net. Through this, we look up synonymous word senses in all available languages, perform some cleaning and filtering and select one sense as label.

Italian (obtained with MT) (Bianchi et al., 2021), Chinese (human translations) (Yang et al., 2022), and Japanese[2] and Arabic[3] (undisclosed methods). Extending ImageNet to more languages, however, is not feasible at scale, due to the challenges of finding native speakers to translate or verify machine-translated labels. With Babel-ImageNet, we exploit the linkage between ImageNet and BabelNet through WordNet, to create the first robust massively multilingual translation of ImageNet classes, avoiding the caveats of polysemy associated with automatic translation of concepts.

Multilingual CLIP. CLIP (Radford et al., 2021) is not the first model to embed images and text in a shared representation space, but it has arguably become the most widely used one, owing its effectiveness – especially in ZS-IC – to the immense pretraining corpus. Older models, not exposed to large-scale VL pretraining, e.g., (Faghri et al., 2018) for English and (Gella et al., 2017; Wehrmann et al., 2019; Kim et al., 2020; Burns et al., 2020) multilingually, focused predominantly on text-to-image retrieval and were not shown to exhibit ZS-IC abilities. MURAL (Jain et al., 2021) was the first – albeit not publicly released – multilingual CLIP model, trained on billions of multilingual image-caption pairs. The only publicly available multilingual CLIP models trained "from scratch" are the OpenCLIP models (Ilharco et al., 2021) (trained on the full multilingual LAION5B dataset (Schuhmann et al., 2022)) and mSigLIP (Zhai et al., 2023) (trained on Google's WebLI (Chen et al., 2022a)). Monolingual CLIP models for a few languages other than English (e.g., Italian, Chinese) have also been released (Bianchi et al., 2021; Yang et al., 2022): due to comparatively small pretraining data, they trail the English models. Given the huge computational cost of training a multilingual CLIP from scratch, teacher distillation (Reimers and Gurevych, 2020) has become popular as an efficient alternative (Carlsson et al., 2022; Chen et al., 2022b; Zhang et al., 2022): a multilingual text encoder (e.g., XLM-R (Conneau et al., 2020)) is forced (commonly using parallel sentences) to align its representation space to the text encoder of English CLIP.

[2] github.com/rinnakk/japanese-clip
[3] github.com/LAION-AI/CLIP_benchmark/pull/68

3 Babel-ImageNet

Why (massively) multilingual ZS-IC? With class labels in a particular language we can evaluate VL models in language-specific ZS-IC. Note that the goal is not to improve the classification performance – using labels in any other language yields worse performance compared to using English labels. Instead, we argue that a model's language-specific ZS-IC performance is a good estimate of the quality of its multilingual VL representation space for the language, and thus a good predictor of the model's performance for that language.

In addition, prior work evaluated models mainly with retrieval. For the languages, where such data exists, we provide a more comprehensive evaluation: ImageNet covers a far more diverse set of concepts than image captions usually contain.

WordNet as a matchmaker for ImageNet and BabelNet. Unlike in most image classification datasets (e.g., CIFAR10, Oxford Pets (Parkhi et al., 2012), Flowers102 (Nilsback and Zisserman, 2008)), where image classes are *words*, ImageNet (Deng et al., 2009) links images to *concepts*, represented with sets of synonyms (synsets) from English WordNet (Miller, 1994). BabelNet (Navigli and Ponzetto, 2010) is a massively multilingual lexico-semantic network, automatically created by merging and consolidating numerous lexico-semantic resources: from WordNets in dozens of languages (e.g., (Hamp and Feldweg, 1997; Pianta et al., 2002)) to (massively multilingual) Wikipedia and WikiData (Vrandečić, 2012).[4] Crucially for our efforts, BabelNet is (1) also organized in (multilingual) synsets, containing synonyms across many languages and (2) each of its synsets has an explicit link to the corresponding (English) WordNet synset (if such exists). With WordNet as the seam between ImageNet and BabelNet, we are able to create a massively multilingual ZS-IC benchmark, without resorting to manual annotation or MT.

Class Translation and Cleaning Process. We illustrate the general process in Figure 1. For every ImageNet class, using its WordNet synset ID, we fetch all synonyms in every language including English (where available). Next, we remove words that match an English word (of the same class) because models use their high-quality English representations for these words, distorting results and leading to misleadingly optimistic estimates of models' multilingual abilities. Next, we eliminate all words that were added to BabelNet via machine translation, removing the potential negative effects of context-agnostic MT from our benchmark. Finally, we select for every class in every language the first remaining word (according to the order in BabelNet) as our final language-specific class label. Classes with no word are removed.

Language Selection. We find that of the 520 languages in BabelNet v5.2, 298 have at least 10 classes using our process, the majority of them being low- and very-low resource languages. We limit our evaluation to 100 non-English languages for more balanced selection of low-, mid- and high-resource languages. To this end, we combine the 92 languages (counting unique ISO codes) covered by the pretraining corpora of XLM-R (Conneau et al., 2020) with 8 manually selected languages not covered by machine translation (neither Google, Microsoft, nor by NLLB (Costa-jussà et al., 2022)).

Grouping Languages in Evaluation. Comparing models' performance over 100 languages (+English) is still unwieldy but averaging performance across all languages is too reductive and consequently not particularly informative. We thus opt for the middle ground: we group Babel-ImageNet languages in four buckets based on their number of classes (<101, 101 to 333, 334 to 667, and 668 to 1000). We argue that the number of classes is a reasonable proxy for general "resourceness" of a language (see §B.1 for the full list of languages and corresponding numbers of classes) and accordingly designate the four groups as *very-low-*, *low-*, *mid-*, and *high-resource*, encompassing 17, 32, 35, and 16 languages, respectively. Additionally, given that ZS-IC becomes easier with fewer classes, averaging results across languages with more comparable numbers of classes makes more sense than averaging them across all languages.

Source	Objective	#Data	#Langs
`OpenAI`	CLIP	400M	1
`OpenCLIP`	CLIP/LiT	5B	>100
`M-CLIP`	distill	3M	69
`M-CLIP`	distill	7M	48
`SentenceTransformer`	distill	>50M	49
`AltCLIP`	distill+LiT	50M+100M	9
`SigLIP`	CLIP	900M	100
`NLLB-SigLIP`	LiT	20M	201

Table 1: CLIP variants benchmarked with: (i) the source (who trained the model), (ii) the training objective (CLIP: contrastive training as in Radford et al. (2021), LiT: locked image tuning Zhai et al. (2022), distill: MSE teacher distillation (Reimers and Gurevych, 2020)), (iii) training data (for "distill" the number of caption pairs, for CLIP/LiT the number of image-text pairs), and (iv) the number of languages seen in training.

Verification. BabelNet mappings are automatically created and not error-free even though, for example, over 90% of Wikipedia-WordNet mappings were manually validated with a precision over 99.5% (Navigli et al., 2021). Still, we also manually verify our mappings on 4 languages with native speakers[5] and on the 3 smallest languages (*xh, om, ha*) with online dictionaries; we find on average 5.4% errors. We believe this to be a very acceptable error rate, considering (i) that around 6% of ImageNet images are mislabeled (Northcutt et al., 2021) and (ii) that there are also erroneous mappings between ImageNet images and WordNet synsets (Nielsen, 2018; Radford et al., 2021).

4 Benchmarking CLIP Models

Models. We present the CLIP variants we benchmark on Babel-ImageNet (overview in Table 1). All use Vision Transformers (ViT) (Dosovitskiy et al., 2021) albeit of different sizes and with dif-

[4]BabelNet v5.2 consolidates 53 sources

[5]The authors and colleagues. Languages: *de, ur, tr, hr*

Model	Param.	v-low	low	mid	high	en
OpenAI B-32	151m	4.28	3.79	5.02	9.23	63.35
ST mBERT B-32	286m	6.23	9.72	15.33	17.44	39.45
M-CLIP mBERT B-32	330m	10.16	15.42	19.63	19.26	29.97
OpenCLIP XLMR-B B-32	366m	12.00	18.29	30.86	39.52	62.32
mSigLIP	371m	17.33	29.05	**48.20**	**56.66**	**75.12**
NLLB-SigLIP-base	507m	**34.11**	**34.58**	32.17	29.37	39.75
M-CLIP XLMR-L B-32	712m	18.52	26.40	33.47	34.11	44.06
M-CLIP XLMR-L B-16+	769m	18.92	27.62	34.98	36.46	47.02
AltCLIP XLMR-L L-14	864m	12.67	16.98	21.32	33.97	73.36
M-CLIP XLMR-L L-14	988m	19.80	29.70	38.17	40.07	52.34
OpenCLIP XLMR-L H-14	1193m	13.77	23.57	41.03	**52.23**	**76.95**
NLLB-SigLIP-large	1195m	**40.61**	**43.22**	**42.78**	39.75	51.96

Table 2: ZS-IC performance on Babel-ImageNet: average results for very-low-/low-/mid-/high-resource languages and English. **Bold**: best result in each column, both between models with base (B) and large (L/H) image encoders.

ferently sized input patches (e.g., B-32 = `Base` Transformer with 32×32-pixel patches).

`OpenAI`: The original CLIP (Radford et al., 2021), trained fully from scratch on 400M English image-caption pairs.

`OpenCLIP`: OpenCLIP (Ilharco et al., 2021) aims to replicate the OpenAI models using the public LAION datasets (Schuhmann et al., 2021, 2022). Two multilingual models have been trained: the B-32 model with the text encoder initialized with the weights of XLM-R-Base and the H-14 model initialized with XLM-R-Large. The B-32 variant is trained with the original contrastive CLIP objective, whereas the H-14 model was trained via locked image tuning (LiT, (Zhai et al., 2022)) in which the pretrained image encoder of the English H-14 OpenCLIP model is frozen and only the parameters of the text encoder are updated.

`SentenceTransformer (ST)`: One of the first distillation-based multilingual CLIP-like models using the approach of Reimers and Gurevych (2020),[6] with over 50M EN-X parallel sentences (X being one of 49 other languages) as supervision. They distill a (distilled) mBERT student (Devlin et al., 2019) from the English OpenAI B-32 teacher.

`M-CLIP`: Multilingual-CLIP (M-CLIP) (Carlsson et al., 2022) is a model distilled using mBERT and OpenAI B-32 with translations of 3M English image captions to 69 languages as parallel supervision. Post-publication, they released a set of models[7] with XLM-R-Large as the student and OpenAI B-32, L-14, and OpenCLIP B-16+ as teachers. These models were trained on 7M captions by Li et al. (2022), translated to 48 languages.

[6] CLIP-VIT-B-32-MULTILINGUAL-V1

[7] github.com/FreddeFrallan/Multilingual-CLIP

`AltCLIP`: This model by Chen et al. (2022b) distills an XLM-R-Large student with OpenAI L-14 as teacher, targeting 9 languages and using as training data a mix of machine-translated captions, multilingual captions sampled from LAION5B, and aligned English-X sentence pairs. After distillation, the authors additionally fine-tune the model via LiT using selected image-text pairs from LAION5B in the 9 target languages.

`SigLIP`: The multilingual mSigLIP (Zhai et al., 2023) is a B-16-size CLIP model trained from scratch on Google's multilingual WebLI dataset (Chen et al., 2022a). They replace the softmax function used in the original CLIP with a sigmoid for more compute-efficient training.

`NLLB-SigLIP`: The NLLB-CLIP models (Visheratin, 2023) are a suite of CLIP models trained with LiT using 100k image-caption pairs translated fully for all 200 languages supported by NLLB (Costa-jussà et al., 2022) for 20M examples in total. They use SigLIP B-16 and SO400M-14 as image encoders with the encoders of NLLB-600M-distilled and NLLB-1.3B-distilled as text encoders, which we call *base* and *large*, respectively.

Zero-Shot Image Classification Setup. We adopt the ZS-IC setup of Radford et al. (2021): for an image-label pair, the image embedding is obtained directly from the image encoder; the label is inserted into 80 different prompt templates (from (Radford et al., 2021)), each of which is independently embedded by the text encoder – the final label representation is then the mean of prompt embeddings. The class with the label embedding that is most similar to the image embedding (according to cosine similarity) is taken as the prediction; accuracy (top-1) is the evaluation metric.

Translating Prompts. We translate the 80 English prompts used by Radford et al. (2021) to our 100 languages using NLLB (Costa-jussà et al., 2022) (`nllb-200-distilled-1.3B`; see §B.2 for details). We show (see §C.1) that translated, language-specific prompts lead to better performance compared to using only the class labels or inserting them into the original English prompts. Moreover (see §C.2), we show that translated prompts yield similar performance as human-crafted prompts in *ar*, *it*, *ja*, and *zh*.

ZS-IC Results. Table 2 summarizes the results for the four language groups, alongside the English performance. The full results for all 100 languages can be found in the Appendix (Table 8).

These results make it abundantly clear that multilingual CLIP models perform dramatically worse (i) for high-resource languages than for English, and (ii) for low- and mid-resource languages than for high-resource languages. Note that this is *despite* the classification tasks *a priori* being easiest for low-resource languages (under 333 classes) and hardest for English, where models must distinguish between all 1000 classes of ImageNet.

The English ImageNet performance of the models is *not* indicative of their ZS-IC performance for other languages, especially low-resource ones: for example, OpenCLIP XLMR-L H-14 outperforms M-CLIP XLMR-L L-14 by 25 accuracy points on English ImageNet, yet trails it 8.6 points on average for low-resource languages. We believe that this points to the "curse of multilinguality" of the text encoder – namely that, under a fixed model capacity, an improvement of representation quality for some language(s) comes at the expense of representational deterioration for others. This phenomenon has been well-documented in particular for XLM-R (Conneau et al., 2020; Pfeiffer et al., 2020b). Among the model variants obtained with the same training procedure (e.g., four variants of M-CLIP), English performance does seem to correlate with the performance on other languages.

The OpenCLIP models and mSigLIP, both trained on massive web-crawled corpora, yield good results for high- and mid-resource languages but perform poorly (in comparison to M-CLIP and NLLB-SigLIP) for low-resource languages. In §C.4, we demonstrate that OpenCLIP performance strongly correlates with the distribution of languages in LAION5B: this would suggest that contrastive training (i.e., CLIP and LiT) leads to poor cross-language generalization.

In contrast, the better performance of (XLM-R-based) M-CLIP models on low-resource languages suggests that distillation-based training offers stronger cross-lingual generalization (and yields best performance even for languages unseen in distillation training, see §C.3). We hypothesize that by aligning representations of captions in all other languages to the representations of corresponding English captions results in a more language-agnostic representation space. At the same time, in line with the "curse of multilinguality", this improved generalization is paid with reduced quality of representations of high-resource languages, where M-CLIP models fall well behind.

The NLLB-SigLIP models, which combine a strong image encoder with the NLLB encoders trained on 200 languages, show impressive results for low- to mid-resource languages but underperform for the high-resource languages, suggesting again a trade-off between languages.

The trade-off between cross-lingual generalization and per-language performance is best exemplified with AltCLIP: the model is exceptionally good for the 9 languages present in its large-scale distillation training (§C.3), yet performs (comparatively) poorly for most other languages – training on a very large dataset for only a few languages simply overwrites the XLM-R's knowledge of other languages, obtained in its original pretraining.

The two mBERT-based models significantly underperform all other models. This is in part due to mBERT being generally a weaker multilingual text encoder (Hu et al., 2020; Lauscher et al., 2020). On top of that, M-CLIP mBERT variants have been trained on less data than XLM-R-based counterparts (3M vs. 7M captions) and ST is distilled with parallel sentences that are *not* image captions.

5 Validating Babel-ImageNet

We perform two additional analyses that establish the validity of Babel-ImageNet as a benchmark: (1) how different number of classes affects performance and findings across languages and (2) how ZS-IC performance on Babel-ImageNet relates to multilingual image-text retrieval performance. We provide further analyses in the Appendix.

Effect of number of classes on ZS-IC accuracy. Babel-ImageNet is an incomplete translation of the ImageNet classes (see §3). Intuitively, classification with fewer classes is easier and results in

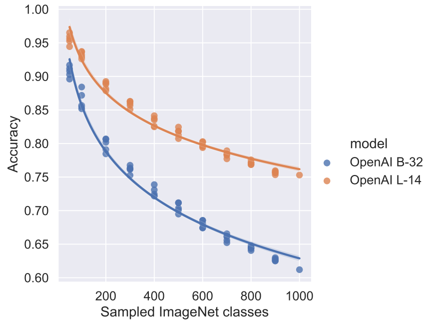


Figure 2: English ImageNet results with random subset of the 1k classes (5 random seeds each).

Model	v-low	low	mid	high	en
ST mBERT B-32	6.23	11.83	24.36	32.65	64.79
M-CLIP mBERT B-32	10.16	18.30	29.86	34.49	54.21
OpenCLIP XLMR-B B-32	12.00	20.84	42.15	60.20	87.56
mSigLIP	17.33	32.20	**59.83**	**75.83**	**93.58**
NLLB-SigLIP-base	**34.11**	**38.93**	45.53	47.78	63.18
M-CLIP XLMR-L B-32	18.52	30.10	45.36	53.21	68.70
M-CLIP XLMR-L B-16+	18.92	30.91	45.75	54.14	70.31
AltCLIP XLMR-L L-14	12.68	19.42	29.20	49.49	93.35
M-CLIP XLMR-L L-14	19.80	32.99	49.43	58.13	75.02
OpenCLIP XLMR-L H-14	13.77	26.32	52.65	**71.65**	**94.09**
NLLB-SigLIP-large	**40.61**	**47.55**	**56.25**	58.84	74.59

Table 3: ZS-IC results with the same number of classes for each language (we report averages over 10 random subsets of 100 classes per language).

higher absolute performance for all models. Our analysis of how the number of classes affects the ZS-IC performance on the English ImageNet for OpenAI CLIP models (B-32 and L-14) confirms this. Figure 2 shows that the task difficulty (i.e., ZS-IC performance) is log-linear in the number of classes: this makes intuitive sense – moving from 50 to 100 classes increases the difficulty more than going from 900 to 950 classes.

We next fix the number of classes to 100 for all Babel-ImageNet languages (except for languages with <100 classes, for which we make no changes) and report the performance in Table 3 (for each language, we average the results over 10 different randomly selected subsets of 100 classes). While in absolute terms the ZS-IC performance increases compared to full class sets (Table 2), and gaps between the language groups widen , our observations do not change: NLLB-SigLIP still exhibits the best performance for low-resource languages, whereas OpenCLIP and mSigLIP are still best for high-resource languages. This renders the (full) Babel-ImageNet a reliable benchmark for directly comparing multilingual VL models.

Multilingual ZS-IC & multilingual image-text retrieval. The existing body of work commonly evaluates multilingual VL models in image-text retrieval. One goal of Babel-ImageNet, which measures multilingual ZS-IC performance, is to reflect (or, at least, give an estimate of) the quality of the multilingual embedding spaces of VL model. We thus compare how models' performance on Babel-ImageNet correlates with their performance on three different multilingual image-text retrieval datasets: xFlickrCo (Bugliarello et al., 2022), XTD (Aggarwal and Kale, 2020), and XM3600 (Thapliyal et al., 2022), covering 7, 10, and 34[8] languages. We use R@1 in text-to-image retrieval as the evaluation metric: it captures the percentage of examples where a correct image is top-ranked for a given caption. We report the full retrieval results in the Appendix (Tables 10, 11, and 12).

Figure 3 displays the retrieval results on xFlickrCo, XM3600, and XTD, respectively, against the ZS-IC accuracy on Babel-ImageNet: each dot represents one model-language combination. The plots reveal high correlation between the Babel-ImageNet and text-to-image retrieval scores across model-language pairs: 0.67 for xFlickrCo, 0.75 for XM3600, and 0.66 for XTD. It is particularly positive that Babel-ImageNet shows the highest correlation with XM3600, with which it intersects in most languages (34). These results confirm that Babel-ImageNet is a sensible benchmark for comparing proficiency of VL models for a multitude of languages not covered by existing benchmarks.

The evaluation also reveals model-specific idiosyncrasies that consideration of either task alone would not have shown. For example, models that use a higher image resolution (M-CLIP B-16+, NLLB-SigLIPs) perform relatively better for retrieval, where the increased detail helps, than for ZS-IC. mSigLIP, potentially due to its small text encoder (compared to, e.g., XLM-R-large), is not as strong for retrieval as in ZS-IC. With the addition of ZS-IC to formerly retrieval-only multilingual evaluation, Babel-ImageNet allows for a more comprehensive evaluation of models.

6 Improving Multilingual CLIP for Low-Resource Languages

Finally, we improve the performance for low-resource languages by resorting to parameter-

[8] For correlation analysis, we exclude *mi* and *quz* as they are not under the 100 Babel-ImageNet languages.

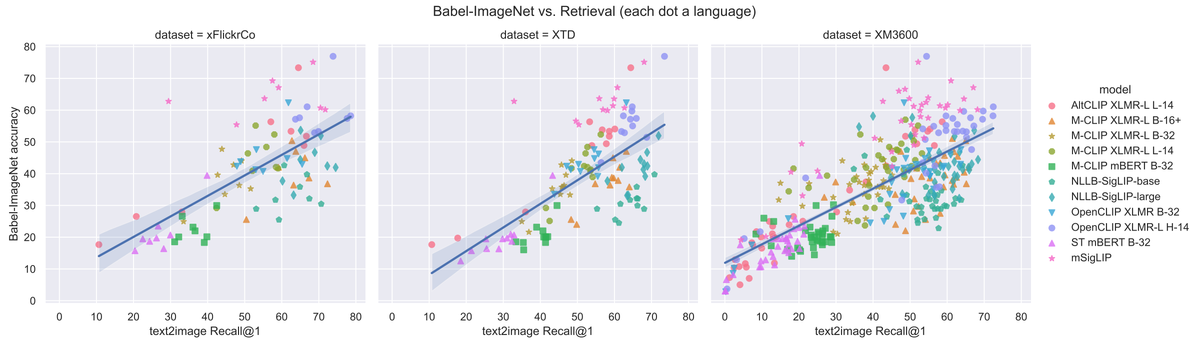


Figure 3: R@1 text-to-image retrieval results on three datasets plotted against Babel-ImageNet performance (each dot denotes the performance of one model for one language) together with a linear regression estimate (95% CI).

Model	Loss	xh	si	lo	ur	my	hi	ms	et	sk	lt	eu	ar	ko	fa	de	zh
M-CLIP XLMR-L B-32	No training	17.7	33.6	12.5	29.4	14.6	36.4	36.6	41.4	39.7	27.5	18.3	30.1	21.4	25.0	38.7	32.7
	Text Contrastive	49.0	46.1	23.4	38.4	33.3	36.2	34.0	34.6	35.8	30.5	27.1	25.3	23.4	26.1	31.4	28.8
	LiT	44.5	49.9	24.7	40.8	29.4	37.3	36.1	33.6	35.2	31.2	29.0	25.9	24.7	27.0	33.6	28.1
	MSE	46.8	53.3	26.9	43.0	37.2	42.5	39.3	38.5	41.0	35.3	34.8	29.1	29.9	31.8	38.1	31.3
OpenCLIP XLMR B-32	No training	24.4	3.1	0.7	25.8	5.8	25.8	37.4	29.8	45.1	35.2	17.1	24.6	33.8	32.7	47.8	40.9
	Text Contrastive	44.0	26.3	16.0	38.7	26.8	32.0	37.8	30.0	39.0	31.0	26.8	21.0	25.0	26.6	39.0	33.8
	LiT	47.7	33.7	19.7	37.5	23.0	30.6	35.5	28.2	38.7	30.4	28.9	22.7	27.4	27.7	39.5	30.7
	MSE	47.6	38.7	24.7	42.8	30.4	39.0	44.0	34.6	44.2	36.7	36.2	26.9	31.9	33.0	45.9	33.6
# Classes		35	97	141	220	232	342	419	496	509	535	625	636	648	682	738	885
# XLM-R Tokens		1.11	2.39	1.23	2.86	1.85	3.23	3.12	2.93	3.55	3.26	2.43	3.46	3.75	4.12	4.01	2.64
# LAION5B Examples		6.71	4.11	4.07	6.13	4.49	7.18	7.05	7.01	7.06	6.98	6.73	7.35	7.01	7.32	8.18	8.16
M-CLIP Distilled		T	F	F	T	F	T	F	T	F	F	F	T	F	F	T	T

Table 4: Results of adapter-based language adaptation of M-CLIP and OpenCLIP with three objectives (Text Contrastive, Text MSE, and LiT). Comparison against the model before adaptation. Colors denote the size of change in performance w.r.t. original model: ≤ -5, ≤ 0, ≤ 5, ≤ 10, ≤ 20, > 20 (best viewed in color). We additionally report language statistics: the number of classes, the number of tokens in XLM-R pre-training (in millions, log10), the number of examples in LAION5B (log10) and whether the language was used in M-CLIP training (True/False).

efficient fine-tuning with adapters (Houlsby et al., 2019; Pfeiffer et al., 2020b), trainable bottleneck layers that we insert into the text encoder. We only update adapter parameters, keeping the original CLIP parameters frozen. We train a separate adapter for each language.

Setup. We train language-specific adapters on top of (a) OpenClip B-32 model (trained from scratch) and (b) M-CLIP XLMR-L B-32 (obtained via distillation). We experiment with three training objectives: English-target language distillation with (i) MSE and (ii) contrastive loss, and (iii) LiT on image-caption pairs. The former two require parallel data, whereas the latter requires images paired with target-language captions. For comparability between languages, we sample 100K captions (with corresponding images) from the synthetic dataset provided by Li et al. (2022) and translate them automatically to all target languages with NLLB. We perform adapter-based specialization for 16 languages. One run (i.e., one model-language-objective combination) takes 3h on a single Nvidia RTX 3090 card (see §B.3 for details).

Results. Table 4 displays the results. For low-resource languages (*xh*, *si*, *lo*, *my*, and *eu*), we observe massive improvements. In contrast, the adaptation brings performance losses for high-resource languages (e.g., *de* and *zh*). We hypothesize that constraining the representation space of a target language to English representations is beneficial for low-resource languages with semantically poor initial representations, but detrimental for high-resource languages with semantically accurate initial representations. For both OpenCLIP and M-CLIP, adaptation with the MSE objective on parallel sentences yields the best results. Overall, the trends in performance changes from language adaptation are very similar between OpenCLIP and M-CLIP, despite the fact that they were obtained using very different training procedures and trained on datasets with different language distributions. This suggests that this commonality in language adaptation behavior stems from the initialization of the text encoder with XLM-R weights.

7 Conclusion

We introduced Babel-ImageNet, the first massively multilingual translation of the ImageNet classes to 100 languages. We leverage the WordNet synsets as the link between ImageNet and BabelNet to obtain high-quality translations without relying on MT or human annotators. Using Babel-ImageNet, we carried out the most comprehensively multilingual comparative evaluation of 11 public CLIP models on zero-shot image classification, demonstrating that all models fail for low(er)-resource languages. Crucially, we validate our benchmark by showing that models' text-to-image retrieval performance (on three datasets) strongly correlates with their ZS-IC performance on Babel-ImageNet for the corresponding languages. Finally, we proposed a parameter-efficient fine-tuning procedure that drastically improves the performance of multilingual CLIP models for low-resource languages.

The wide range of languages encompassed by our benchmark reveals that the theoretical "multilinguality" of CLIP models is practically very limited and points to the need for methods that derive robust VL encoders with much stronger performance especially for low-resource languages: e.g., better distillation procedures that retain more of the impressive performance of English CLIP.

8 Limitations

While Babel-ImageNet greatly improves language coverage for evaluation of multilingual VL models, there are some limitations of our work:

For one, ImageNet classes tend to be Anglo-centric due to inherited biases from WordNet (Shankar et al., 2017; DeVries et al., 2019; Liu et al., 2021) so while our benchmark evaluates the performance on languages from all over the globe, we do not evaluate the model performance on concepts specific (or even unique) to cultures in which those languages are spoken. As a result, Babel-ImageNet may overestimate the actual usability of an VL model for real-world uses in some cultures and geographies.

Further, we select for Babel-ImageNet the 92 languages used in XLM-R pretraining along with 8 more manually chosen languages. This selection reinforces research focus on those languages to the detriment of other (mainly extremely low-resource) languages. However, we release our code, as well as data for labels of 298 languages and encourage future research to consider an even wider set of languages.

Acknowledgements

This work was in part supported by the Alexander von Humboldt Foundation.

References

Pranav Aggarwal and Ajinkya Kale. 2020. Towards Zero-shot Cross-lingual Image Retrieval. *CoRR*, abs/2012.05107. ArXiv: 2012.05107.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the Third Shared Task on Multimodal Machine Translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, Belgium, Brussels. Association for Computational Linguistics.

Federico Bianchi, Giuseppe Attanasio, Raphael Pisoni, Silvia Terragni, Gabriele Sarti, and Sri Lakshmi. 2021. Contrastive Language-Image Pre-training for the Italian Language. *CoRR*, abs/2108.08688. ArXiv: 2108.08688.

Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulic. 2022. IGLUE: A benchmark for transfer learning across modalities, tasks, and languages. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 2370–2392. PMLR.

Andrea Burns, Donghyun Kim, Derry Wijaya, Kate Saenko, and Bryan A. Plummer. 2020. Learning to Scale Multilingual Representations for Vision-Language Tasks. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IV*, volume 12349 of *Lecture Notes in Computer Science*, pages 197–213. Springer.

Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. 2022. Cross-lingual and Multilingual CLIP. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 6848–6854. European Language Resources Association.

Xi Chen, Xiao Wang, Soravit Changpinyo, A. J. Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil

Houlsby, and Radu Soricut. 2022a. PaLI: A Jointly-Scaled Multilingual Language-Image Model. *CoRR*, abs/2209.06794. ArXiv: 2209.06794.

Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Fulong Ye, Qinghong Yang, and Ledell Wu. 2022b. AltCLIP: Altering the Language Encoder in CLIP for Extended Language Capabilities. *CoRR*, abs/2211.06679. ArXiv: 2211.06679.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *CoRR*, abs/2207.04672. ArXiv: 2207.04672.

J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Terrance DeVries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. 2019. Does Object Recognition Work for Everyone? In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 52–59. Computer Vision Foundation / IEEE.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Constantin Eichenberg, Sidney Black, Samuel Weinbach, Letitia Parcalabescu, and Anette Frank. 2022. MAGMA - multimodal augmentation of generative models through adapter-based finetuning. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2416–2428. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description. In *Proceedings of the Second Conference on Machine Translation*, pages 215–233, Copenhagen, Denmark. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German Image Descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 12. BMVA Press.

Alex Fang, Simon Kornblith, and Ludwig Schmidt. 2023. Does progress on ImageNet transfer to real-world datasets? *CoRR*, abs/2301.04644. ArXiv: 2301.04644.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6894–6910. Association for Computational Linguistics.

Gregor Geigle, Jonas Pfeiffer, Nils Reimers, Ivan Vulic, and Iryna Gurevych. 2022. Retrieve Fast, Rerank Smart: Cooperative and Joint Approaches for Improved Cross-Modal Retrieval. *Transactions of the Association for Computational Linguistics*, 10:503–521.

Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. 2017. Image Pivoting for Learning Multilingual Multimodal Representations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2839–2845. Association for Computational Linguistics.

Birgit Hamp and Helmut Feldweg. 1997. Germanet-a lexical-semantic net for german. In *Automatic information extraction and building of lexical semantic resources for NLP applications*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. OpenCLIP.

Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. 2021. MURAL: Multimodal, Multitask Representations Across Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 3449–3463. Association for Computational Linguistics.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

A. Karpathy and L. Fei-Fei. 2017. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676.

Donghyun Kim, Kuniaki Saito, Kate Saenko, Stan Sclaroff, and Bryan A. Plummer. 2020. MULE: Multimodal Universal Language Embedding. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 11254–11261. AAAI Press.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *CoRR*, abs/2301.12597. ArXiv: 2301.12597.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. 2019. COCO-CN for Cross-Lingual Image Tagging, Captioning, and Retrieval. *IEEE Trans. Multim.*, 21(9):2347–2360.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually Grounded Reasoning across Languages and Cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 10467–10485. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

George A. Miller. 1994. WordNet: A Lexical Database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Roberto Navigli, Michele Bevilacqua, Simone Conia, Dario Montagnini, and Francesco Cecconi. 2021.

Ten years of babelnet: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4559–4567. ijcai.org.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a Very Large Multilingual Semantic Network. In *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden*, pages 216–225. The Association for Computer Linguistics.

Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongdong Zhang, and Nan Duan. 2021. M3P: Learning Universal Representations via Multitask Multilingual Multimodal Pre-Training. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 3977–3986. Computer Vision Foundation / IEEE.

Finn Årup Nielsen. 2018. Linking ImageNet WordNet Synsets with Wikidata. In *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*, pages 1809–1814. ACM.

Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated Flower Classification over a Large Number of Classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing, ICVGIP 2008, Bhubaneswar, India, 16-19 December 2008*, pages 722–729. IEEE Computer Society.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. 2012. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 3498–3505. IEEE Computer Society.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin O. Steitz, Stefan Roth, Ivan Vulic, and Iryna Gurevych. 2022. xGQA: Cross-Lingual Visual Question Answering. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2497–2511. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulic, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. AdapterHub: A Framework for Adapting Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 46–54. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulic, Iryna Gurevych, and Sebastian Ruder. 2020b. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7654–7673. Association for Computational Linguistics.

Hieu Pham, Zihang Dai, Golnaz Ghiasi, Hanxiao Liu, Adams Wei Yu, Minh-Thang Luong, Mingxing Tan, and Quoc V. Le. 2021. Combined Scaling for Zero-shot Transfer Learning. *CoRR*, abs/2111.10050. ArXiv: 2111.10050.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *First international conference on global WordNet*, pages 293–302.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2641–2649.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do ImageNet Classifiers Generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400. PMLR.

Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4512–4525. Association for Computational Linguistics.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton

Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS*.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. *CoRR*, abs/2111.02114. ArXiv: 2111.02114.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 1351–1361. Association for Computational Linguistics.

Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. 2017. No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World. *CoRR*, abs/1711.08536. Arxiv: 1711.08536.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pages 2443–2449, New York, NY, USA. Association for Computing Machinery.

Ashish V. Thapliyal, Jordi Pont-Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A Massively Multilingual Multimodal Evaluation Dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 715–729. Association for Computational Linguistics.

Alexander A. Visheratin. 2023. NLLB-CLIP - train performant multilingual image retrieval model on a budget. *CoRR*, abs/2309.01859.

Denny Vrandečić. 2012. Wikidata: A new platform for collaborative data collection. In *Proceedings of the 21st international conference on world wide web*, pages 1063–1064.

Jonatas Wehrmann, Maurício Armani Lopes, Douglas M. Souza, and Rodrigo C. Barros. 2019. Language-Agnostic Visual-Semantic Embeddings. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 5803–5812. IEEE.

An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022. Chinese CLIP: Contrastive Vision-Language Pretraining in Chinese. *CoRR*, abs/2211.01335. ArXiv: 2211.01335.

Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. 2017. STAIR Captions: Constructing a Large-Scale Japanese Image Caption Dataset. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 417–421, Vancouver, Canada. Association for Computational Linguistics.

Yan Zeng, Wangchunshu Zhou, Ao Luo, and Xinsong Zhang. 2022. Cross-View Language Modeling: Towards Unified Cross-Lingual Cross-Modal Pre-training. *CoRR*, abs/2206.00621. ArXiv: 2206.00621.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid Loss for Language Image Pre-Training. *CoRR*, abs/2303.15343. ArXiv: 2303.15343.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. LiT: Zero-Shot Transfer with Locked-image text Tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18102–18112. IEEE.

Liang Zhang, Anwen Hu, and Qin Jin. 2022. Generalizing Multimodal Pre-training into Multilingual via Language Acquisition. *CoRR*, abs/2206.11091. ArXiv: 2206.11091.

Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. 2021. UC2: Universal Cross-Lingual Cross-Modal Vision-and-Language Pre-Training. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 4155–4165. Computer Vision Foundation / IEEE.

A License

Babel-ImageNet is a processed version of BabelNet v5.2 downloaded from `https://babelnet.org`, made available with the BabelNet Non-Commercial License (see `https://babelnet.org/full-license`).

B Data and Training Details

B.1 Babel-ImageNet

Table 5 lists the 100 Babel-ImageNet languages with their corresponding number of classes.

Figure 4 visualizes the relationship between the number of classes of a language in Babel-ImageNet and the number of tokens for the language in the XLM-R pretraining corpus (which we use as a proxy for the language "resourceness"). We see that the two are generally correlated (Spearman rank correlation of 0.78), albeit with some expected outliers, e.g., Chinese is "token-compact" so the token count does not reflect its high-resourceness well.

B.2 Prompts

We use NLLB (Costa-jussà et al., 2022) (*nllb-200-distilled-1.3B*) to translate the 80 prompts used by Radford et al. (2021) to our 100 languages. The exceptions are *fy, la, br, wuu, nv, cv, diq, chr, ce, hak, nah*, which are not supported by NLLB; for those languages we report the better results between: (1) using only the language-specific labels and (2) inserting 'labels' into English prompts. For the ISO 639-1 languages corresponding to macro languages, there is only one corresponding ISO 639-3 language in NLLB, except for *no* where we choose Bokmål and for *az* where we choose North Azerbaijani. We translate the prompts in their template form with the { } placeholders. We use a range of different methods like HTML tags or other special characters to increase the likelihood of preserving the placeholders during translation and then select the first successful approach. If no method worked, we append { } to the end of the sentence. We perform no language-specific adaptions like combining prompt variants with definite and indefinite articles for languages where this distinction does not exist (or articles do not exist at all) nor do we account for the grammatical gender of the classes when inserting them in the template.

B.3 Training

Training Data: For the language-specific adaptation training in §6, we leveraged the BLIP (Li et al., 2022) image-caption dataset CCS_SYNTHETIC_FILTERED_LARGE.JSON[9].

Hyperparameters: We train with AdamW (Loshchilov and Hutter, 2019), 0.1 weight decay, a linear learning rate schedule with 20% warmup, learning rate 1e-3 (chosen with sweep over 1e-3, 5e-4, 3e-4, 1e-4), batch size 512 (OpenCLIP)/ 192 (M-CLIP), for 100 epochs (OpenCLIP)/ 15 epochs (M-CLIP; longer training yielded no improvements). Hyperparameters are chosen based on results on Sinhala. We perform no early stopping and use the last epoch for evaluation. The temperature for the contrastive loss is a trainable parameter as in Radford et al. (2021) but we freeze it for text contrastive loss (training it resulted in worse results). The maximum text sequence length is 70. For adapters, we use the Pfeiffer architecture (Pfeiffer et al., 2020b) (task adapters, not language adapters) with reduction factor 16 with the implementation from AdapterHub (Pfeiffer et al., 2020a). We pre-encode images and English captions; i.e. the English embeddings for MSE and contrastive loss are not computed by the trained model but come from the model before training. We do not use any type of image augmentation.

Negative results: We experimented with the following methods but did not pursue them further due to not-better or poor results.

1. Training with MSE loss using aligned English-X sentences from WikiMatrix (Schwenk et al., 2021), similar to the ST and (in part) AltCLIP models, resulted in a performance decrease throughout (except for *si* with OpenCLIP) as Table 6 shows. This suggests that it is important to use *"visually-descriptive"* parallel data (i.e., parallel image captions), rather than *any* parallel data.

2. LoRA fine-tuning (Hu et al., 2022) ($\alpha = 8, r = 16$, lr 1e-3 after sweep) significantly ($>$10% on *si*) underperformed adapter-based fine-tuning.

3. SimCSE loss (a self-supervised objective) (Gao et al., 2021) based only on target-language captions yielded no improvements

[9] `https://github.com/salesforce/BLIP#pre-training-datasets-download`

very low	*om* (18), *xh* (35), *ha* (47), *so* (58), *sd* (61), *nah* (62), *hak* (63), *mg* (64), *sa* (66), *or* (71), *ce* (73), *chr* (83), *am* (85), *diq* (89), *si* (97), *su* (98), *as* (98)
low	*ku* (101), *gu* (106), *ug* (106), *ps* (112), *pa* (128), *ne* (134), *cv* (137), *mr* (140), *lo* (141), *fy* (155), *bs* (156), *km* (167), *kn* (175), *yi* (175), *jv* (183), *mn* (201), *te* (202), *gd* (217), *sw* (220), *ur* (220), *my* (232), *ky* (247), *uz* (254), *nv* (257), *tl* (272), *sq* (273), *la* (276), *wuu* (278), *ml* (281), *bn* (282), *br* (297), *af* (303)
mid	*hi* (342), *ta* (346), *hr* (347), *az* (365), *kk* (365), *lv* (392), *sl* (393), *cy* (407), *is* (409), *be* (415), *ms* (419), *ka* (438), *mk* (453), *hy* (454), *id* (463), *sr* (468), *gl* (473), *et* (496), *ga* (502), *sk* (509), *vi* (523), *lt* (535), *tr* (559), *el* (572), *hu* (594), *no* (599), *bg* (602), *eo* (603), *da* (610), *cs* (615), *eu* (625), *ar* (636), *uk* (640), *ko* (648), *he* (648)
high	*pt* (667), *fa* (682), *ro* (687), *sv* (699), *ja* (733), *de* (738), *ru* (748), *nl* (749), *ca* (767), *it* (773), *pl* (778), *fr* (799), *es* (845), *zh* (885), *th* (896), *fi* (973)

Table 5: The 100 languages of Babel-ImageNet in their respective groups. Number of classes in parentheses.

compare to the initial model, i.e., without any additional language-specialization training (experimented with OpenCLIP and batch size 256).

4. Multitask training with both LiT and MSE distillation objectives produced no gains compared to training only with the MSE objective.

C Further Experiments and Analysis

C.1 Experimental Validation of Machine-translated Prompts

We show in Figure 5 that our translated prompts produce better results (on average, across all languages), compared to (i) using just the labels and (ii) inserting the translated labels into the original English prompts created by Radford et al. (2021). With the translated prompts, we get gains of over 2 points for low-resource languages and up to 5 points for high-resource languages.

C.2 Comparison with Existing ImageNet Translations

Prior work has created full translations of the 1k ImageNet classes into *ar, zh, jp, it* along with human-written prompts for those languages. We use those translations to validate our BabelNet-derived labels and MT prompts: We evaluate models on the subset of ImageNet classes available for each language in Babel-ImageNet and compare a) only labels and b) human-created templates vs. our MT prompts. Results are shown in Figure 6. While results for *ar* and *it* are slighly higher in absolute numbers on the existing translation, the relative order of models on the Babel-ImageNet benchmarks of those languages is nearly identical to their relative ranking on the respective benchmarks with manually translated ImageNet labels.

We observe that the human-written prompts do not result in a relative improvement over our MT prompts (i.e. no down-shift parallel to the $x = y$ line). In fact, for *it*, our MT prompts even close the gap slightly compared to the label-only setup.

C.3 Performance Differences between Distilled and Not-Distilled Languages

With teacher distillation, one would expect the performance in the languages seen in the distillation data to be better than in other languages, not used for distillation. With the wide language selection of our benchmark, we can analyze in-depth how performance on "distilled" languages differs from the performance on "non-distilled" languages.

We compare results for distilled languages on the low/mid/high-resource language groups for M-CLIP and AltCLIP in Table 7; we use the OpenCLIP H-14 model as reference for an expected 'baseline' Δ-difference in performance between the distilled/not-distilled language groups that is due to other factors inherent to the specific languages and not the distillation. For AltCLIP, we see that the the performance on the 8 distilled languages is significantly better than on the non-distilled languages. Moreover, the performance on its distilled languages is even comparable to that of the larger H-14 model. For M-CLIP, the performance on the distilled languages is only slightly better than on the non-distilled low- and mid-resource languages when compared to the OpenCLIP model and the gap is even smaller for high-resource languages. Interestingly, the performance on non-distilled low-resource languages is still noticeably better for M-CLIP than for the OpenCLIP H-14 model. We speculate that the shorter training of M-CLIP compared to OpenCLIP might retain more of the language-specific competences for low-resource languages, obtained in XLM-R pretraining.

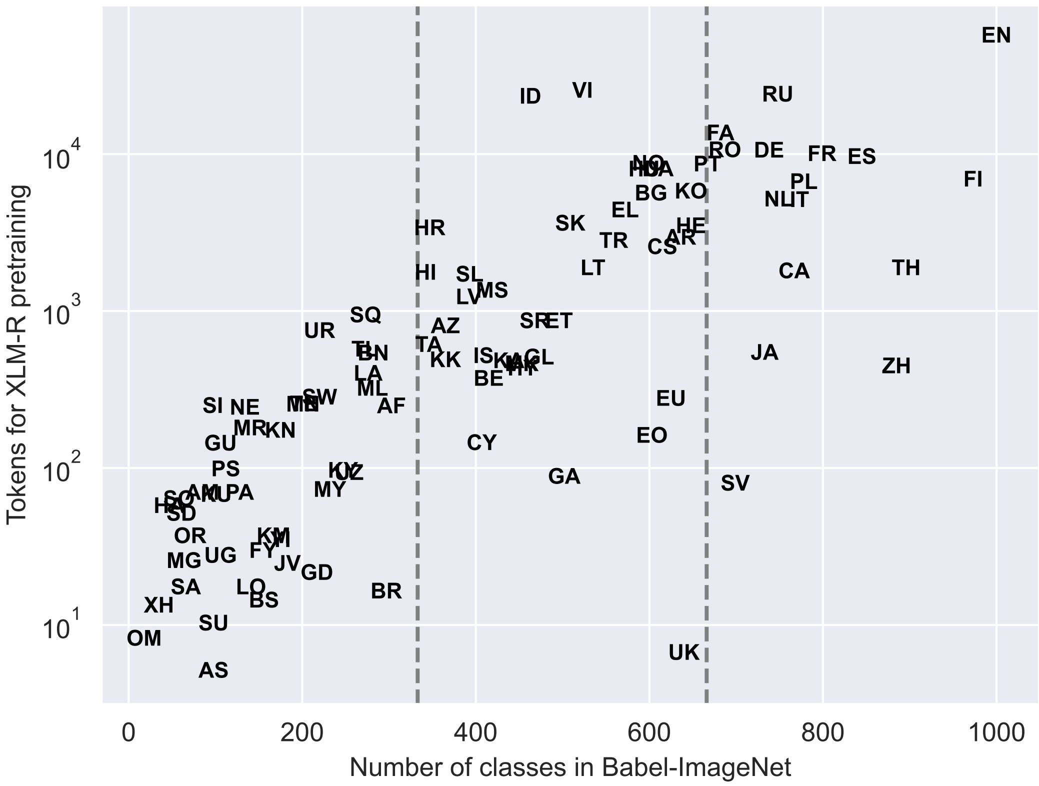


Figure 4: Number of classes in Babel-ImageNet plotted against the number of tokens (millions, log10) in the XLM-R pretraining corpus. When taking the XLM-R tokens as proxy for "resourceness" of a language, we see that this generally correlates with the number of classes. Vertical lines indicate the grouping of languages for evaluation.

C.4 LAION5B: Language Distribution and Performance

The distillation-based models evaluated in our benchmark use the same number of training examples for every language. The OpenCLIP models, on the other hand, are trained on LAION5B which follows a more 'natural' distribution of image-caption pairs across languages, as found on the web: Figure 7b shows that over half the data is English, 7 high-resource languages account for another 25% of the data, whereas all remaining languages "share" the remaining 25%.

We can see in Figure 7a that the OpenCLIP's ZS-IC performance for Babel-ImageNet languages highly depends on the number of instances of those languages in the LAION5B dataset. The Spearman rank correlation between the number of language-specific LAION5B examples and the respective Babel-ImageNet accuracy for the language is 0.76. This suggests that pure image-text contrastive pretraining results in poor generalization and limited cross-lingual gains to languages unseen in pretraining. Additional training objectives that aim to better align the multilingual space for example using paired text like in MURAL (Jain et al., 2021) might be necessary to improve results of OpenCLIP-like models (trained from scratch) for low-resource languages.

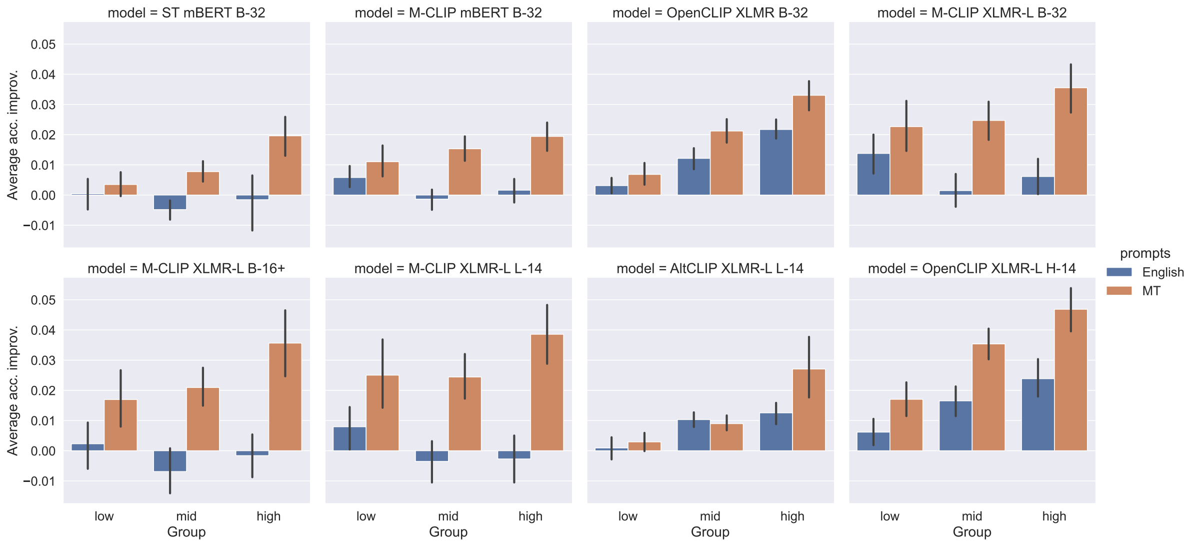


Figure 5: Average increase within the low/mid/high language groups (with 95% CI) over only labels using English prompts (with non-English labels) and our machine-translated prompts.

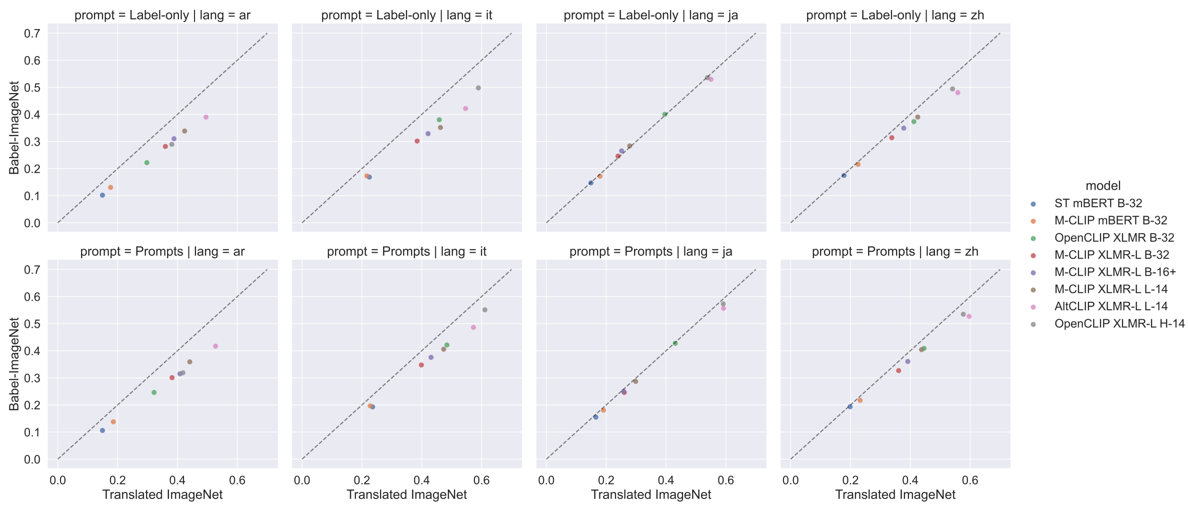


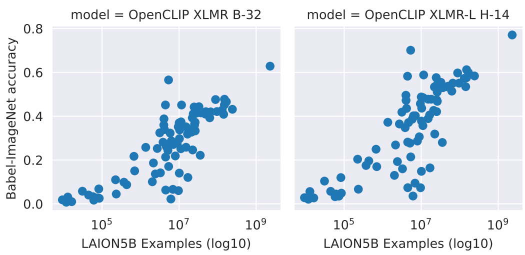
Figure 6: Results on Babel-ImageNet against results on existing ImageNet translations ("labels only" in top row and with our MT prompts vs. human-created prompts in bottom row) for four languages: *ar, zh, jp, it*. Relative model ranking is nearly identical between Babel-ImageNet and manually translated ImageNet benchmarks of respective languages.

Model	Loss	xh	si	lo	ur	my	hi	ms	et	sk	lt	eu	ar	ko	fa	de	zh
M-CLIP XLMR-L B-32	No training	17.7	33.6	12.5	29.4	14.6	36.4	36.6	41.4	39.7	27.5	18.3	30.1	21.4	25.0	38.7	32.7
	MSE (WikiData)	—	25.0	—	—	—	22.2	—	21.3	23.3	21.8	16.9	24.0	16.3	22.5	31.4	24.3
OpenCLIP XLMR B-32	No training	24.4	3.1	0.7	25.8	5.8	25.8	37.4	29.8	45.1	35.2	17.1	24.6	33.8	32.7	47.8	40.9
	MSE (WikiData)	—	17.1	—	—	—	18.1	—	18.5	25.0	23.0	17.7	21.1	14.7	23.0	38.6	28.5

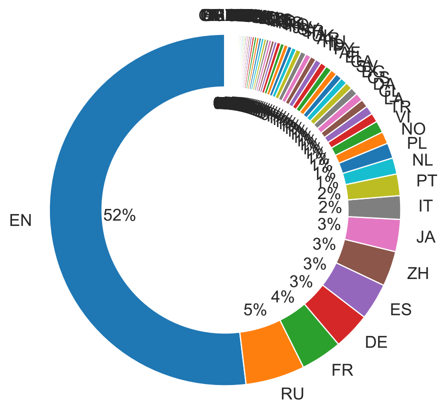
Table 6: Results of adapter-based language adaptation of M-CLIP and OpenCLIP with TextMSE loss using aligned sentences from WikiMatrix. Colors denote the size of change in performance w.r.t. original CLIP model: ≤ -5, ≤ 0, ≤ 5, ≤ 10, ≤ 20, > 20 (best viewed in color).

Model	low	Δlow	mid	Δmid	high	Δhigh
M-CLIP XLMR-L L-14	37.6	+14.4	45.9	+16.8	41.6	+11.3
OpenCLIP XLMR-L H-14	26.1	+10.1	47.5	+13.2	55.3	+15.8
AltCLIP XLMR-L L-14	—	—	47.5	+28.0	51.7	+28.9
OpenCLIP XLMR-L H-14	—	—	37.8	-3.5	57.0	+7.4

Table 7: Average results for the "distilled" languages in the low/mid/high-resource language groups and the Δ difference to the other "non-distilled" languages of the groups. OpenCLIP H-14 serves as control for language-specific differences in performance not caused by distillation. For M-CLIP, 14/41, 18/35, and 13/16 languages per group are distilled; For AltCLIP, 0/41, 2/35, and 6/16 are distilled.



(a) Accuracy of the LAION5B-trained OpenCLIP models for the 92 Babel-ImageNet languages plotted against the number of examples for each language in LAION5B (log10).



(b) Distribution of languages in LAION5B (exluding the 1.3B no-language examples) shows that most examples are either English or one of a few other languages.

Figure 7: The relationship between the LAION5B language distribution with the performance of OpenCLIP models trained on that data.

D Full Results

We report full results for all languages on (i) Babel-ImageNet and (ii) each of the three image-text retrieval datasets: xFlickrCo, XM3600, and XTD.

lang	AC L-14	MC mB B-32	MC B-16+	MC B-32	MC L-14	NS-base	NS-large	OAI B-32	OC B-32	OC H-14	mSigLIP	ST B-32
en	73.36	29.97	47.02	44.06	52.34	39.75	51.96	63.35	62.32	76.95	75.12	39.45
af	22.6	25.26	50.33	47.48	55.32	41.11	53.89	10.65	36.48	47.53	47.17	14.94
am	7.13	0.99	27.22	29.25	29.11	50.12	55.79	1.51	1.04	2.56	14.56	0.85
ar	42.44	13.99	32.06	30.68	36.3	29.52	40.1	0.56	24.51	31.7	42.27	10.69
as	6.53	17.22	21.69	22.84	24.92	39.94	49.22	1.18	5.43	9.53	23.14	2.92
az	16.4	23.71	26.57	25.69	28.81	34.77	44.24	7.27	25.83	36.98	47.82	12.68
be	27.52	14.08	35.35	32.66	36.85	36.72	48.4	0.64	27.18	40.12	32.05	10.02
bg	34.48	21.61	42.63	40.57	47.13	32.23	42.43	0.8	41.26	54.02	58.99	19.56
bn	7.25	25.99	31.17	29.77	35.55	47.48	57.68	0.17	8.61	19.57	49.42	3.06
br	12.02	3.99	10.71	9.64	11.61	8.82	12.28	7.14	13.02	15.23	14.96	3.47
bs	29.45	39.15	60.18	59.74	65.27	49.55	58.69	13.04	55.99	69.88	79.49	35.35
ca	32.02	19.9	37.48	34.94	40.33	28.34	38.67	11.54	33.37	46.78	42.96	17.5
ce	16.38	6.9	15.73	11.37	13.53	19.04	20.22	1.18	21.21	22.19	25.53	4.36
chr	1.2	1.2	1.2	1.16	1.47	1.18	2.19	1.47	1.33	0.48	1.54	1.64
cs	20.03	22.04	38.44	36.59	41.89	31.26	42.06	6.36	41.8	54.11	66.02	19.5
cv	29.93	11.75	23.64	19.78	22.99	17.55	24.67	0.58	31.8	34.67	32.96	10.32
cy	12.8	19.81	11.39	11.31	12.99	25.8	33.45	5.04	12.08	16.52	12.72	5.62
da	24.96	22.79	46.92	44.45	51.84	34.66	46.87	12.39	42.92	55.5	62.46	22.65
de	27.96	20.07	42.01	39.59	46.39	31.74	44.43	15.83	47.58	61.05	67.1	18.7
diq	14.81	6.49	20.67	19.19	18.47	17.62	22.11	8.2	18.72	21.42	25.96	9.08
el	5.03	16.58	41.73	39.03	45.75	33.3	45.72	0.9	37.21	50.77	51.16	15.29
eo	24.84	13.04	25.22	24.64	28.58	35.12	48.63	5.77	21.82	28.5	21.65	10.9
es	51.78	20.13	38.64	36.31	42.07	32.14	42.42	16.98	44.78	57.4	60.72	20.67
et	15.54	19.79	45.19	42.22	48.73	29.21	39.91	4.64	29.68	37.84	53.5	13.51
eu	17.72	9.63	18.92	18.72	21.25	34.16	45.22	7.43	16.87	21.51	21.89	9.0
fa	20.13	20.69	28.24	25.55	29.26	27.6	37.34	0.46	32.66	42.48	52.3	13.89
fi	10.56	14.45	27.49	25.79	30.27	22.82	32.19	3.96	25.0	35.69	46.88	11.25
fr	54.06	20.17	38.84	35.84	42.39	32.21	43.05	21.08	46.21	59.72	63.54	20.93
fy	23.73	6.83	27.9	28.72	30.09	22.7	29.64	11.99	34.13	40.65	36.21	6.15
ga	8.36	3.12	9.42	8.29	9.76	20.12	27.98	2.72	6.71	9.45	5.55	2.24
gd	6.01	3.46	8.25	7.52	10.53	25.43	30.48	3.05	5.98	7.3	4.94	2.63
gl	43.92	23.08	41.53	40.26	45.05	35.13	47.57	15.47	43.82	54.66	47.37	23.91
gu	11.34	28.77	28.11	26.94	32.32	49.3	63.42	0.4	6.94	12.02	20.72	7.64
ha	3.23	19.36	4.94	4.17	5.19	41.96	49.74	1.62	2.47	3.57	3.06	2.0
hak	33.27	18.7	32.29	26.03	28.1	21.94	27.9	3.97	32.6	35.08	31.14	11.08
he	7.03	16.68	22.04	21.02	23.7	26.07	35.16	0.3	26.64	35.87	48.99	12.86
hi	12.89	20.99	37.37	37.04	41.61	41.47	54.75	0.09	25.85	38.65	40.51	17.58
hr	21.75	30.22	51.08	50.67	57.0	40.19	51.75	8.63	44.82	58.2	66.58	25.64
hu	16.55	19.91	45.47	42.69	50.3	31.38	42.39	5.57	40.93	53.25	63.9	17.1
hy	5.3	16.89	18.65	17.97	18.26	26.7	36.7	0.17	9.91	17.5	35.09	7.43
id	26.52	26.6	50.39	47.73	55.14	42.63	53.63	16.09	43.78	57.61	69.25	23.49
is	10.82	18.91	41.84	39.96	45.56	29.15	38.46	2.93	13.51	19.25	30.18	6.04
it	49.43	19.97	37.85	35.81	41.27	29.95	41.32	15.55	41.88	55.03	59.89	19.49
ja	56.33	18.55	25.57	24.93	29.2	28.91	37.71	4.18	42.35	57.14	62.79	15.74
jv	21.8	19.51	33.85	32.89	37.27	42.74	52.2	14.89	31.81	40.44	50.4	15.08
ka	9.0	15.48	22.83	22.05	24.26	26.69	34.27	0.2	11.09	20.44	41.13	8.66
kk	31.61	19.96	26.59	25.59	28.44	34.44	44.46	0.53	28.12	34.52	47.7	10.8
km	6.72	0.91	16.57	17.68	19.77	25.78	30.16	0.79	3.31	3.16	31.54	0.38
kn	10.97	21.17	27.09	27.36	29.57	43.3	53.95	1.04	4.16	5.68	16.3	2.16
ko	53.8	16.02	24.02	21.61	25.12	24.48	33.15	0.42	33.52	43.51	56.58	12.52
ku	11.23	8.95	12.75	14.2	14.79	14.73	21.27	5.09	14.1	15.88	15.5	8.73
ky	32.11	18.48	28.79	28.7	31.69	38.53	47.82	0.65	32.56	38.55	41.6	14.01
la	23.06	4.7	16.43	11.23	14.38	12.8	21.94	10.77	21.14	26.65	26.35	4.2
lo	9.86	0.74	12.45	12.75	13.05	32.13	37.59	0.77	0.85	2.21	10.26	0.71
lt	18.7	21.69	29.9	28.06	31.96	29.09	39.78	5.94	34.68	45.79	55.12	18.58
lv	21.42	26.01	33.9	34.14	37.14	31.99	40.98	7.74	38.52	47.3	56.7	23.65
mg	13.25	8.62	13.06	14.19	14.12	38.47	47.72	4.06	13.94	14.62	12.34	9.06
mk	31.41	23.55	47.78	46.53	51.4	33.18	43.89	1.33	36.08	49.47	47.31	20.45
ml	7.64	20.05	38.69	38.8	44.01	37.69	49.56	0.16	1.69	4.45	15.31	1.23
mn	20.44	28.29	20.86	19.37	21.78	25.09	28.6	1.02	18.58	26.49	44.96	14.06
mr	17.46	25.9	45.89	45.41	47.83	59.39	68.63	1.13	32.29	41.66	42.24	29.66
ms	21.25	22.12	39.23	37.86	43.59	35.71	46.72	14.94	37.08	48.35	61.1	19.26
my	9.47	2.49	14.72	14.49	16.2	38.34	46.23	0.49	5.82	10.27	17.16	10.35
nah	8.13	5.52	6.52	7.42	9.03	7.13	7.97	4.35	10.68	11.35	9.9	2.23
ne	16.51	17.07	37.57	40.16	39.21	49.78	60.39	0.54	25.0	36.39	40.16	16.9
nl	25.03	18.57	39.09	37.33	44.56	29.67	41.41	13.05	41.71	54.93	57.96	18.88
no	23.47	21.13	43.4	40.66	47.56	33.24	43.78	9.4	41.98	53.57	61.21	19.55
nv	0.65	0.39	0.6	0.38	0.16	0.66	0.29	0.35	0.44	0.41	0.25	0.54
om	8.0	3.44	11.11	9.67	9.67	29.22	30.44	4.44	5.44	13.0	10.22	8.78
or	12.62	1.75	30.62	31.18	36.11	57.24	70.56	1.75	3.27	1.61	1.61	1.77

Table 8: Babel-ImageNet results for all languages (sorted alphabetically expect for English). To save space, we shorten sources (OAI: OpenAI; OC: OpenCLIP; MC: M-CLIP; ST: SentenceTransformer, AC: AltCLIP, NS: NLLB-SigLIP) and remove the text model if possible.

lang	AC L-14	MC mB B-32	MC B-16+	MC B-32	MC L-14	NS-base	NS-large	OAI B-32	OC B-32	OC H-14	mSigLIP	ST B-32
pa	11.42	7.97	33.89	32.09	31.69	60.27	70.47	1.53	1.77	2.56	3.05	4.09
pl	19.72	18.34	35.83	33.33	38.95	28.89	39.18	7.63	39.15	51.46	61.38	16.32
ps	18.05	10.86	19.48	18.98	23.21	33.38	41.8	1.57	21.66	25.04	24.52	11.88
pt	37.41	21.57	43.24	40.06	47.36	34.42	46.26	14.43	47.21	59.62	61.7	23.03
ro	23.92	18.78	40.45	37.78	44.26	29.74	41.21	10.19	35.17	47.6	51.22	16.22
ru	48.87	18.34	36.82	35.25	41.7	30.47	41.96	0.54	42.95	58.21	60.16	16.39
sa	11.18	9.45	17.36	17.21	18.97	32.36	37.09	0.91	11.55	11.73	23.15	9.55
sd	15.93	20.98	28.13	27.9	29.87	43.34	51.84	2.0	14.56	16.49	15.61	9.41
si	19.4	2.19	33.32	33.75	36.39	55.38	65.48	2.64	3.22	5.32	29.13	2.12
sk	22.97	25.31	43.22	40.56	46.11	35.56	46.68	8.65	44.99	58.68	66.17	22.72
sl	17.39	24.61	45.05	42.98	48.71	34.15	47.76	6.01	36.93	48.65	59.23	22.46
so	3.76	17.55	10.38	12.55	12.14	29.97	38.86	2.48	6.66	7.41	7.07	5.17
sq	21.71	26.21	49.82	48.12	54.07	38.39	48.67	8.24	33.99	43.35	43.77	24.86
sr	30.64	19.18	45.99	44.58	49.92	32.35	44.06	1.18	33.93	47.7	45.43	17.49
su	18.73	16.04	28.16	29.22	29.33	37.39	44.04	12.06	27.71	30.61	35.9	11.67
sv	21.06	21.1	45.72	42.0	49.44	31.93	42.26	9.17	42.88	55.2	61.67	19.68
sw	10.37	17.3	37.05	36.85	39.38	31.87	38.76	4.72	10.25	12.91	20.17	6.63
ta	5.53	18.67	18.78	18.09	20.98	30.98	41.14	0.35	4.54	6.73	30.21	2.12
te	10.75	24.95	32.56	31.75	34.44	45.62	58.13	0.51	2.83	3.71	19.17	3.02
th	11.85	15.55	29.69	27.77	32.77	25.63	33.35	1.33	28.71	40.16	40.84	10.62
tl	18.25	16.17	33.43	32.51	38.11	26.43	33.62	7.86	17.54	21.73	33.05	8.18
tr	17.67	23.16	44.98	42.78	48.4	35.84	46.72	7.61	41.25	52.88	63.65	19.47
ug	9.06	2.17	13.19	11.68	12.75	33.42	40.3	1.19	3.06	4.64	6.19	2.3
uk	34.82	17.61	36.67	35.77	41.22	28.99	39.5	0.63	39.19	53.18	57.54	15.13
ur	18.48	27.24	27.15	29.85	31.17	47.62	58.92	0.59	26.01	37.12	29.48	17.75
uz	17.24	21.4	22.57	22.54	24.13	36.09	48.54	6.13	21.83	28.45	36.42	9.8
vi	11.69	19.09	39.59	37.94	44.58	29.77	39.13	6.94	40.92	53.39	59.9	17.86
wuu	72.53	24.4	49.64	28.58	48.78	31.35	41.76	3.43	59.02	70.81	72.22	19.69
xh	21.77	16.34	19.2	17.71	20.11	57.6	69.2	18.86	24.11	27.14	24.8	14.23
yi	5.29	1.04	18.37	18.9	19.38	39.22	52.81	0.71	2.49	4.73	3.82	1.33
zh	53.35	21.99	36.41	33.5	40.9	25.5	33.26	1.85	40.73	53.28	55.42	19.75

Table 9: Babel-ImageNet results for all languages (sorted alphabetically expect for English). To save space, we shorten sources (OAI: OpenAI; OC: OpenCLIP; MC: M-CLIP; ST: SentenceTransformer, AC: AltCLIP) and remove the text model if possible.

model	en	de	es	id	jp	ru	tr	zh	average
AltCLIP XLMR-L L-14	64.50	33.05	66.65	20.75	57.05	66.00	10.65	62.50	45.24
M-CLIP mBERT B-32	42.45	32.20	39.80	33.15	31.10	39.10	36.05	36.60	35.43
M-CLIP XLMR-L B-16+	63.80	59.10	67.35	62.75	50.45	72.35	66.25	63.20	63.06
M-CLIP XLMR-L B-32	49.15	42.70	48.55	43.85	33.50	51.60	47.00	44.65	44.55
M-CLIP XLMR-L L-14	58.00	50.85	58.45	52.95	42.40	59.00	55.85	52.10	53.09
NLLB-SigLIP-base	66.80	58.30	67.45	60.65	53.10	70.65	64.90	59.25	62.04
NLLB-SigLIP-large	70.55	65.00	72.30	65.10	60.20	74.55	68.90	63.20	67.04
OpenCLIP XLMR B-32	61.80	53.15	61.45	48.90	47.90	65.50	53.00	54.95	54.98
OpenCLIP XLMR-L H-14	73.85	66.85	77.65	64.80	63.70	78.60	68.85	69.90	70.05
mSigLIP	68.45	59.15	70.45	57.50	29.45	71.75	55.35	47.80	55.92
ST mBERT B-32	39.85	24.35	29.95	26.65	20.40	28.15	22.50	26.20	25.46

Table 10: xFlickrCo T2I R@1. Average is without English.

lang	AC L-14	MC mB B-32	MC B-16+	MC B-32	MC L-14	NS-base	NS-large	OC B-32	OC H-14	mSigLIP	ST B-32
en	43.49	24.63	47.47	31.69	36.56	48.25	50.14	48.38	54.43	52.17	25.58
average	21.65	21.76	50.98	33.79	39.69	52.07	56.03	42.6	50.52	44.55	13.31
ar	43.56	18.03	51.12	34.34	38.93	53.02	55.0	39.26	47.73	45.13	9.75
bn	1.25	10.53	34.53	20.92	22.61	49.11	50.58	2.22	5.22	20.75	0.19
cs	9.8	21.76	50.23	33.63	38.13	49.81	53.78	45.08	53.41	46.95	14.57
da	12.0	27.11	61.85	42.86	50.56	58.41	63.35	52.29	62.73	54.32	18.57
de	30.17	28.69	65.79	45.13	53.2	63.79	68.32	63.68	72.44	67.06	19.62
el	4.08	19.64	51.76	35.51	41.45	49.07	55.02	45.0	54.08	40.44	12.49
es	48.75	24.24	55.36	37.6	43.55	54.41	58.22	54.02	61.52	59.53	17.49
fa	13.04	23.3	50.89	32.71	38.0	53.04	56.65	48.31	56.69	51.36	13.02
fi	5.79	24.16	58.1	38.04	44.02	57.68	63.25	42.95	57.15	45.03	13.43
fil	7.29	16.92	45.98	29.26	34.87	43.71	47.9	7.27	9.54	20.95	2.24
fr	55.4	28.4	62.51	42.2	50.36	62.24	66.63	60.79	69.72	64.53	21.3
he	6.57	23.9	48.75	30.39	37.82	56.97	61.44	48.17	57.93	51.01	10.99
hi	2.76	8.39	29.9	15.01	18.52	34.8	36.21	17.12	21.8	17.85	4.16
hr	8.53	29.31	61.41	41.96	48.72	57.51	62.01	48.82	59.68	48.63	18.97
hu	9.95	23.31	62.83	43.35	50.98	56.93	63.29	49.74	62.71	54.88	14.3
id	18.52	28.91	65.0	44.88	52.13	64.12	67.3	56.2	66.29	61.03	21.5
it	51.82	26.33	61.35	41.25	48.28	59.26	64.18	58.92	67.19	63.49	16.49
ja	58.65	27.1	54.15	33.83	40.38	58.29	61.64	59.14	69.55	42.66	16.19
ko	51.32	20.17	46.07	29.28	35.83	52.08	56.59	44.51	53.44	47.58	9.74
mi	0.23	0.02	0.34	0.13	0.27	23.99	27.68	0.51	0.49	0.32	0.11
nl	21.04	23.59	53.11	36.25	42.42	51.74	56.6	48.36	56.37	52.8	16.93
no	12.91	26.36	58.1	39.46	46.32	57.35	61.93	49.98	59.68	50.95	15.53
pl	13.68	26.26	59.03	41.27	48.45	56.62	62.74	55.31	65.2	58.03	17.98
pt	38.08	25.18	57.08	38.98	46.62	55.65	60.97	55.2	64.46	60.15	20.02
quz	2.69	0.81	2.14	1.19	1.61	16.63	18.6	2.93	3.5	3.03	0.75
ro	17.35	25.19	64.59	42.82	51.38	59.85	65.96	56.32	66.59	53.74	16.64
ru	56.54	28.83	65.17	44.24	51.92	61.83	65.75	63.61	72.36	68.18	20.17
sv	12.83	25.38	59.29	40.71	47.63	55.29	60.36	51.57	59.66	53.13	16.44
sw	2.5	12.52	39.02	22.75	27.59	41.98	44.78	2.5	3.18	11.1	0.41
te	3.86	13.18	26.83	14.54	17.24	37.71	39.99	0.25	0.58	4.35	0.06
th	13.38	20.43	52.29	33.21	38.72	53.87	56.49	44.44	54.29	25.01	9.38
tr	7.51	22.98	55.25	37.65	44.37	54.27	57.83	46.41	57.68	49.79	13.11
uk	33.67	26.32	61.75	42.83	49.6	57.44	61.98	55.37	65.18	54.87	17.24
vi	5.52	24.9	59.41	38.84	45.59	55.99	59.09	53.56	64.2	52.67	17.25
zh	54.74	26.44	56.97	37.59	44.05	51.98	54.96	55.34	61.92	50.32	16.67

Table 11: XM3600 T2I R@1 results. Average is without English.

model	en	de	es	fr	it	jp	ko	pl	ru	tr	zh	average
AltCLIP XLMR-L L-14	64.4	36.0	58.6	60.1	57.8	53.3	56.7	17.7	53.9	10.7	58.6	46.34
M-CLIP mBERT B-32	44.5	40.9	41.4	42.0	41.5	33.4	35.5	41.4	35.4	39.0	40.7	39.12
M-CLIP XLMR-L B-16+	63.2	61.4	59.8	59.3	61.0	48.3	49.8	64.0	54.8	59.6	58.8	57.68
M-CLIP XLMR-L B-32	48.5	46.9	46.4	46.1	45.8	35.0	36.9	48.0	43.2	45.7	45.4	43.94
M-CLIP XLMR-L L-14	56.3	52.2	52.7	51.8	53.6	41.5	42.5	54.1	48.4	52.7	53.5	50.30
NLLB-SigLIP-base	70.8	64.2	66.3	66.0	66.2	55.3	61.2	68.0	61.6	66.0	60.2	63.50
NLLB-SigLIP-large	71.9	67.0	68.9	68.0	67.8	58.1	63.4	68.3	62.0	68.8	63.7	65.60
OpenCLIP XLMR B-32	63.2	54.5	54.6	55.7	55.7	47.1	43.8	55.5	50.3	48.2	50.8	51.62
OpenCLIP XLMR-L H-14	73.5	64.8	65.9	64.7	64.9	64.3	56.4	68.7	62.4	62.7	61.9	63.67
mSigLIP	68.0	60.8	62.7	59.9	58.1	32.9	49.6	59.6	56.9	55.9	50.4	54.68
ST mBERT B-32	43.3	32.3	32.9	32.0	29.8	21.4	18.5	28.8	25.5	25.3	31.0	27.75

Table 12: XTD T2I R@1 results. Average is without English.