

MetaCLIP 2: A Worldwide Scaling Recipe

Yung-Sung Chuang¹, Yang Li², Dong Wang², Ching-Feng Yeh², Kehan Lyu², Ramya Raghavendra², James Glass¹, Lifei Huang², Jason Weston,² Luke Zettlemoyer², Xinlei Chen², Zhuang Liu³, Saining Xie⁴, Wen-tau Yih², Shang-Wen Li², Hu Xu²

MIT CSAIL¹, Meta FAIR², Princeton³, NYU⁴

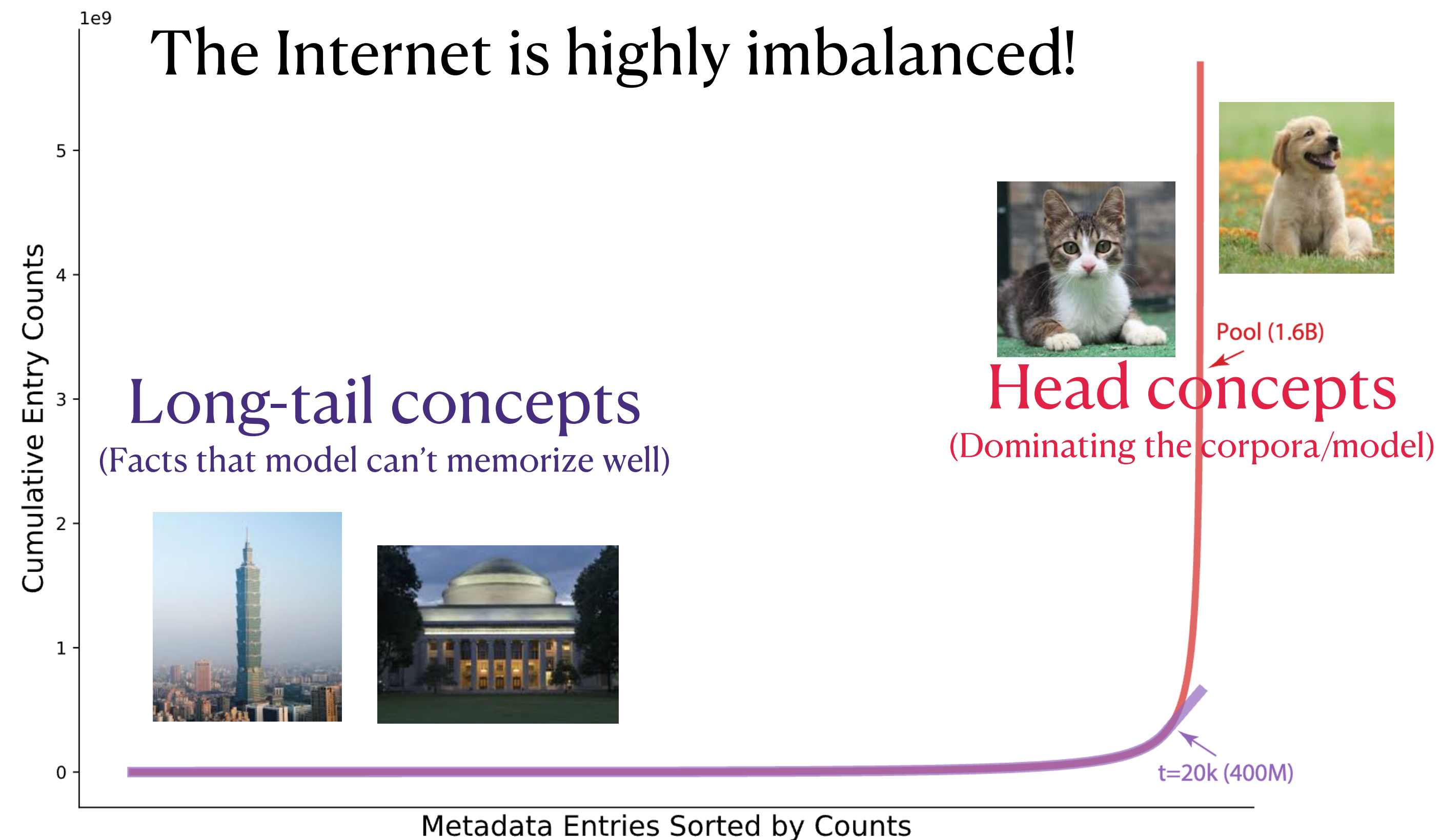


NeurIPS 2025

<https://arxiv.org/abs/2507.22062>

How to curate high-quality multi-modal dataset at scale?

From scratch, **NOT** using any “teacher” models as “filters”, i.e., LAION, DFN



Metadata can help!

- A dictionary of words/phrases of real world concepts
e.g., dog, cat, Taipei 101, MIT
- Source can be Wikipedia/WordNet

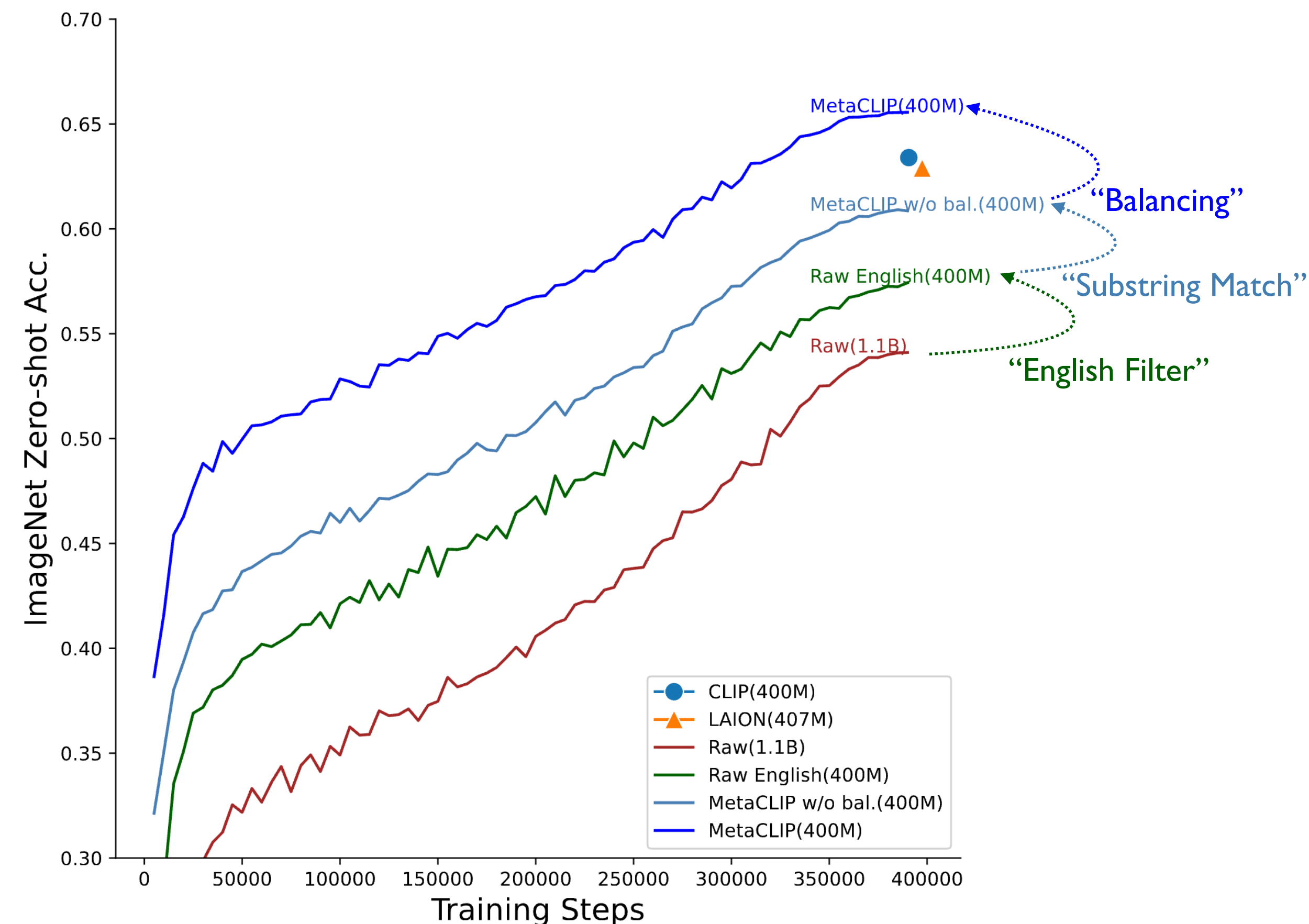
Substring Matching:

- Do substring matching for all image captions
- Exclude the images matched with no metadata

Balancing:

- Constrain the max # of images per concept
e.g. $t = 20k$ for 400M text-image pairs
 $t = 170k$ for 2.5B text-image pairs

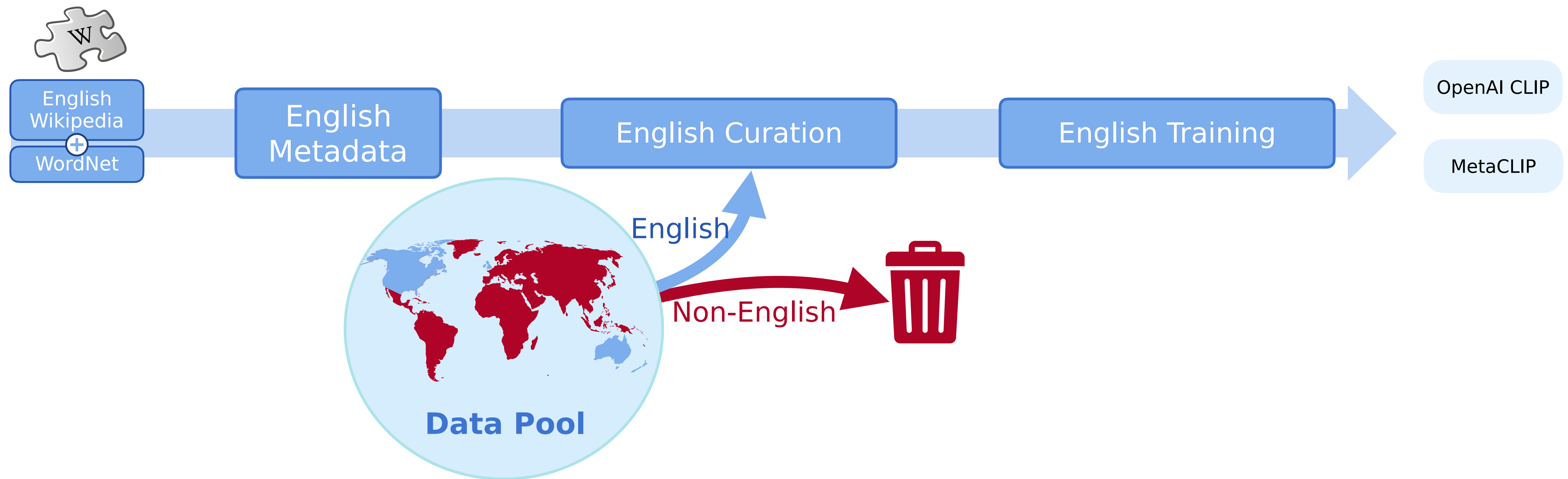
(use downsampling in the data pipeline)



Hu Xu, et al. "Demystifying CLIP Data. (MetaCLIP)" ICLR 2024.

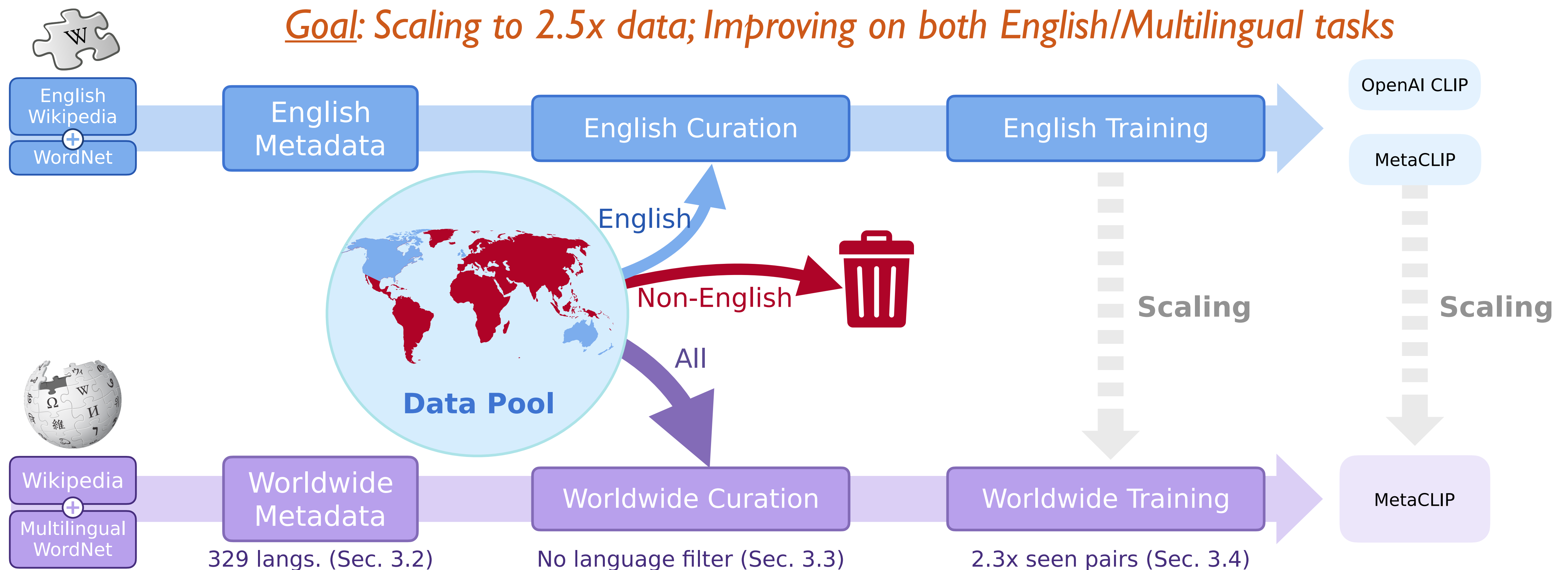
MetaCLIP is limited by the English filter

- About 60% of data are wasted in MetaCLIP because they are non-English



MetaCLIP is limited by the English filter

- About 60% of data are wasted in MetaCLIP because they are non-English
- In MetaCLIP 2, we leverage this 60% data to scale CLIP for visual ability (Beyond multilingual)



Worldwide Data Curation

Substring Matching

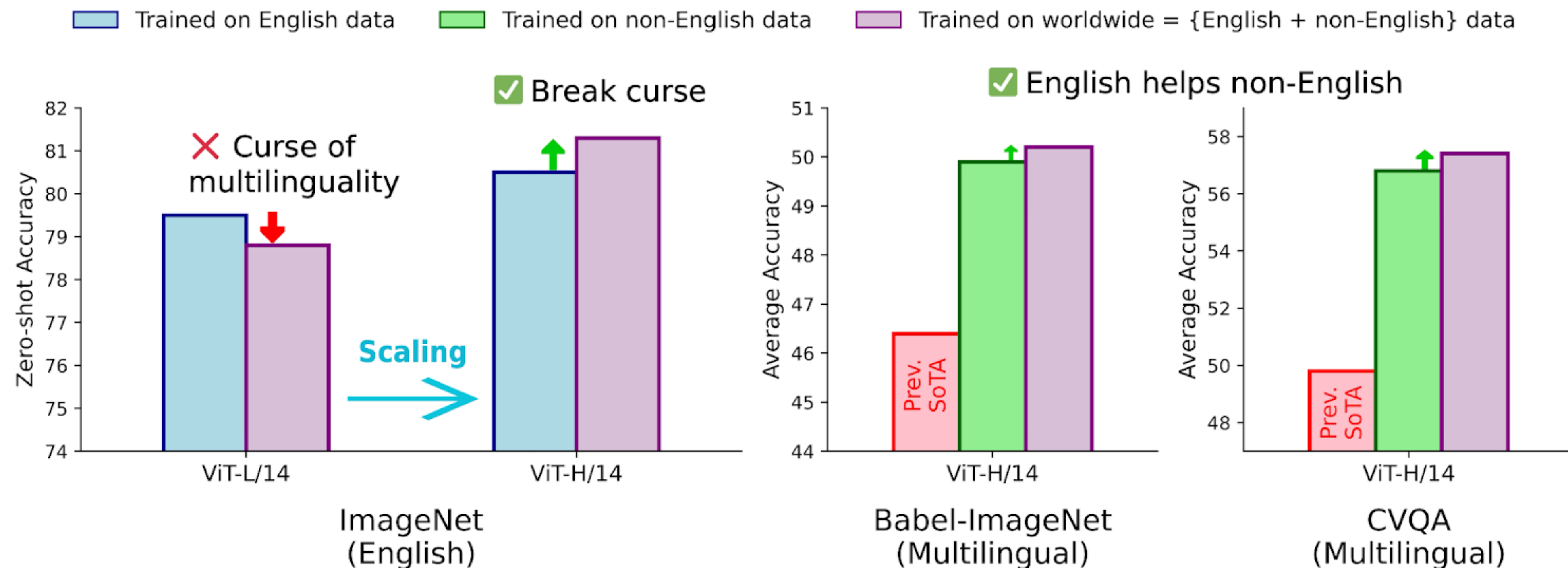
- Build **metadata** for 329 Wikipedia languages: Unigram/Bigram/Title + Multilingual OpenWordNet; use custom tokenizers for 7 special languages, e.g. Chinese/Japanese
- Perform **multilingual substring matching** for each language separately

Challenges in balancing for different languages

- The threshold t is tuned on English data; it's too high for other smaller languages
- Impossible to tune the threshold t for all 300+ languages
- We instead keep the “**tail concepts ratio** = 6%” **invariant** among languages
b.c. the 6% ratio is also invariant when scaling the model size of English MetaCLIP

Curse of Multilinguality: non-English data hurts English ability

- Curated a dataset contains 5B text-images pairs (~2B are English)
- Models $\leq$ ViT-L/14 suffers from the *curse of multilinguality*, i.e., mSigLIP < SigLIP
- Scaling to ViT-H/14 breaks the curse



Cultural Diversity Evaluation

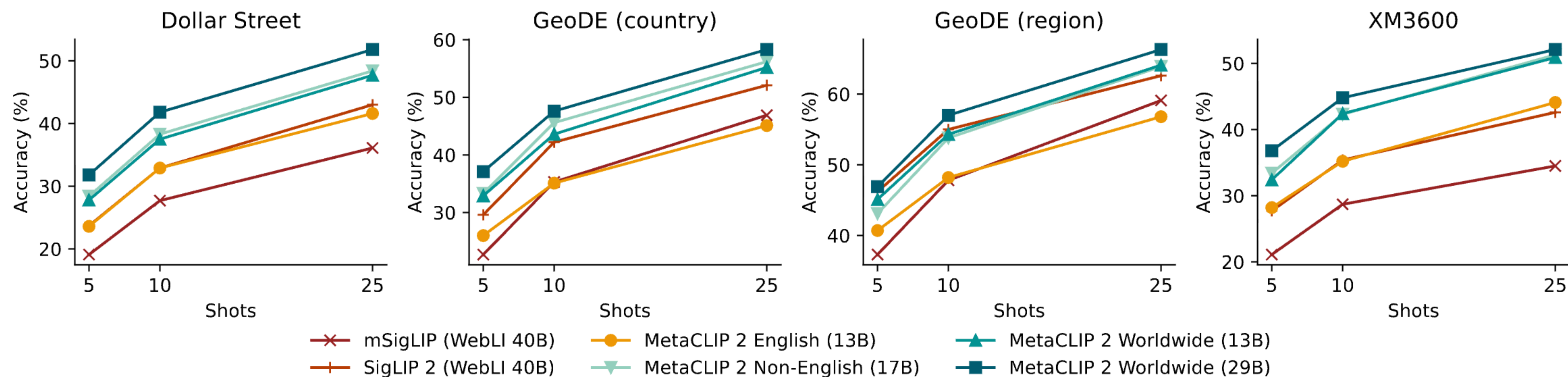
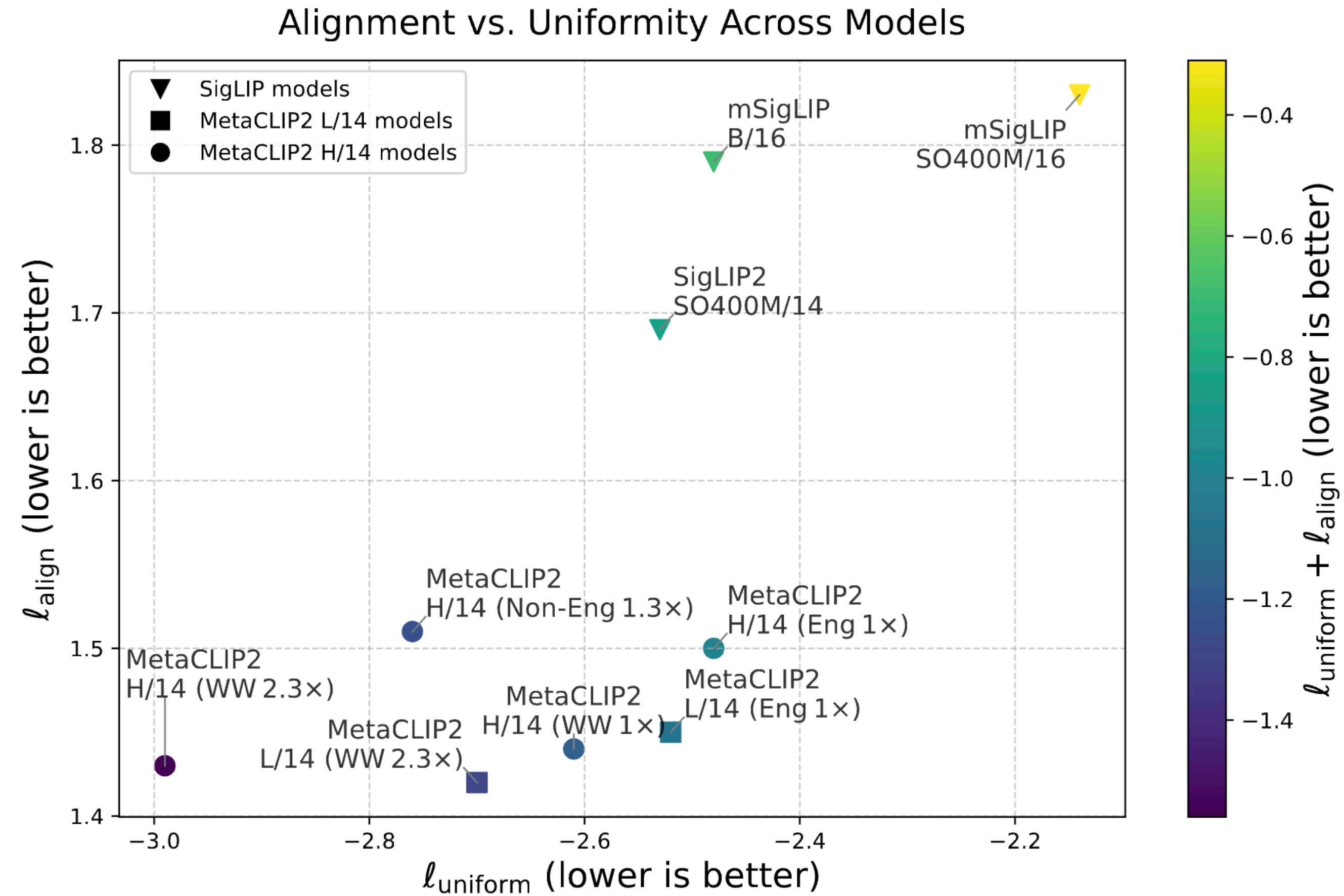


Figure 3 Few-shot geo-localization accuracy on cultural diversity benchmarks.

Embedding Quality Evaluation



Training on worldwide data improves the alignment & uniformity of embedding space

MetaCLIP 2: A Worldwide Scaling Recipe



GitHub:

[github.com/facebookresearch/
MetaCLIP](https://github.com/facebookresearch/MetaCLIP)



Website:

meta-clip.github.io/



HF Models:

[huggingface.co/models?
other=metaclip_2](https://huggingface.co/models?other=metaclip_2)

