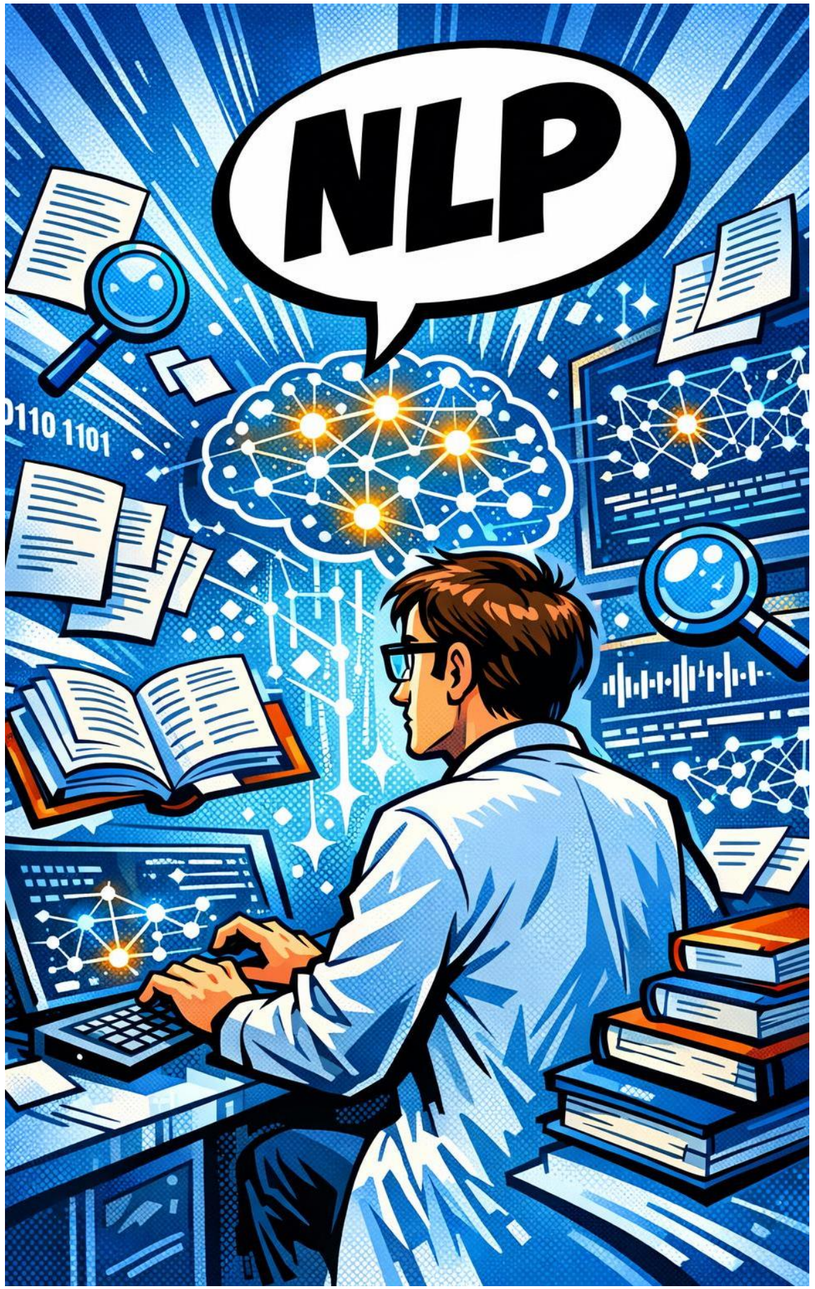
MetaCLIP: Open, Scalable Data Curation for Vision-Language Models

Scott Wen-tau Yih
Meta FAIR

Text-only Research (before 2023)

- Research Interests: NLP, ML & IR
  - Question Answering
  - Neural Retrieval
  - Retrieval-augmented generation
  - Factuality / Hallucination

- Theme: Knowledge Grounding



Multimodal Models Need World Knowledge

What does Bichon Frisé look like..?

A **Bichon Frisé** sitting on the bench.

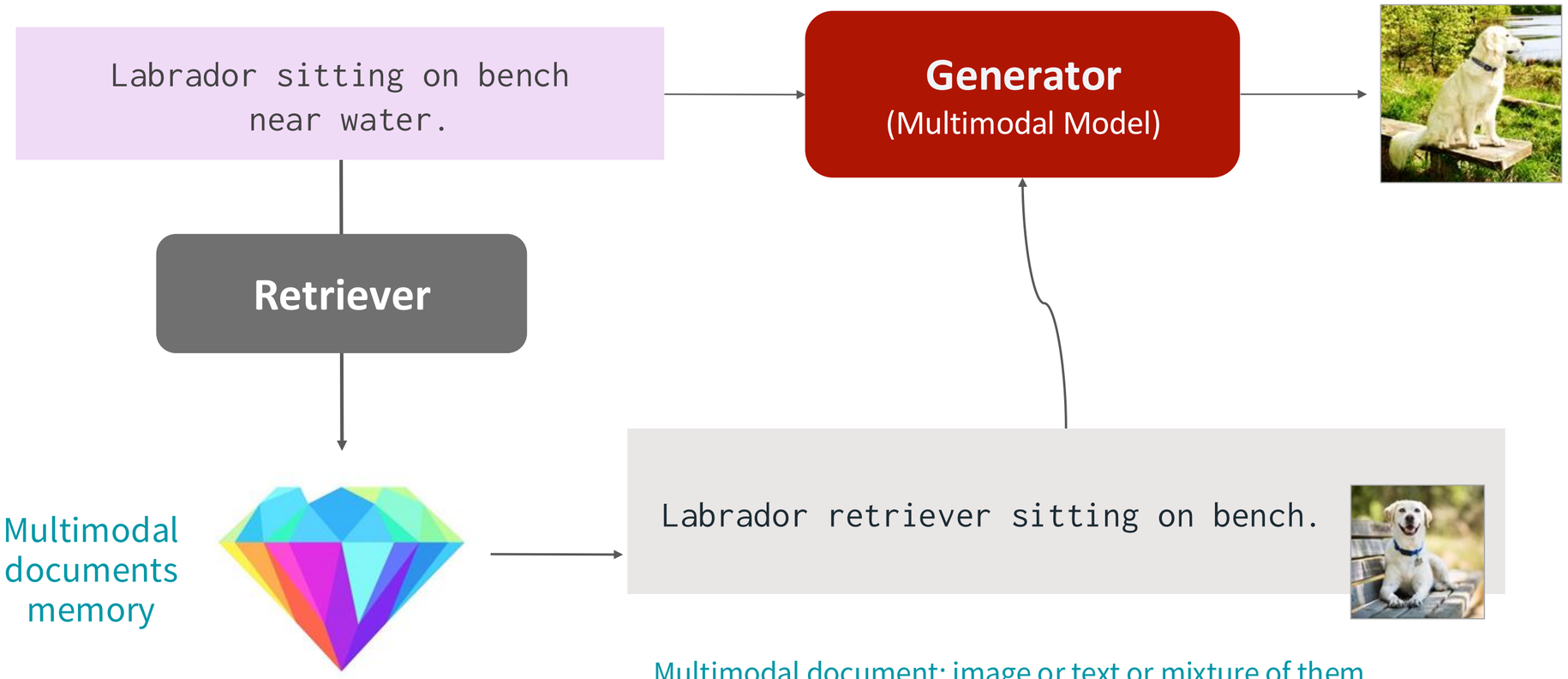
Text to image

What is the name of this place..?

Image to text

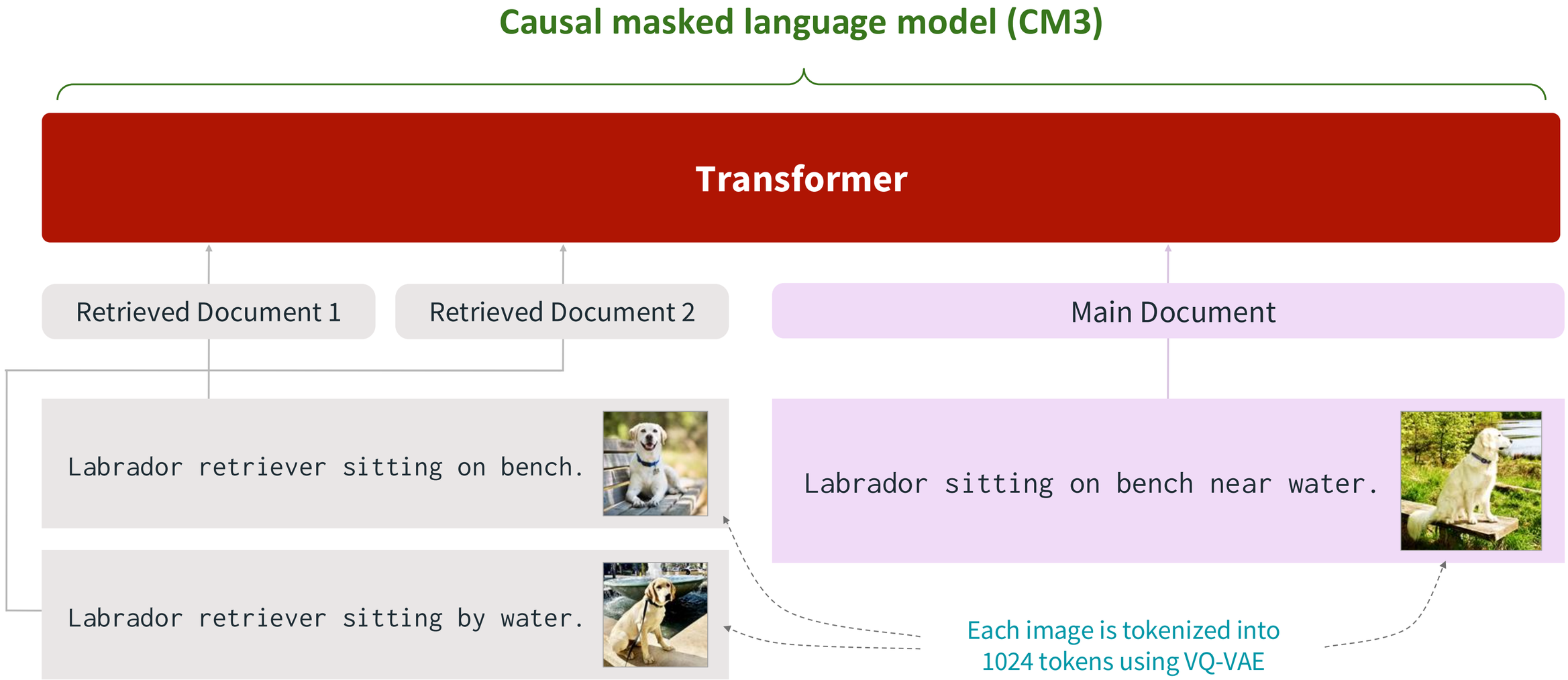
The Dragon and Tiger Pagodas next to fireworks.

Our Idea: Retrieval Augmented Multimodal Model



Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, Wen-tau Yih. Retrieval-Augmented Multimodal Language Modeling. ICML-2023

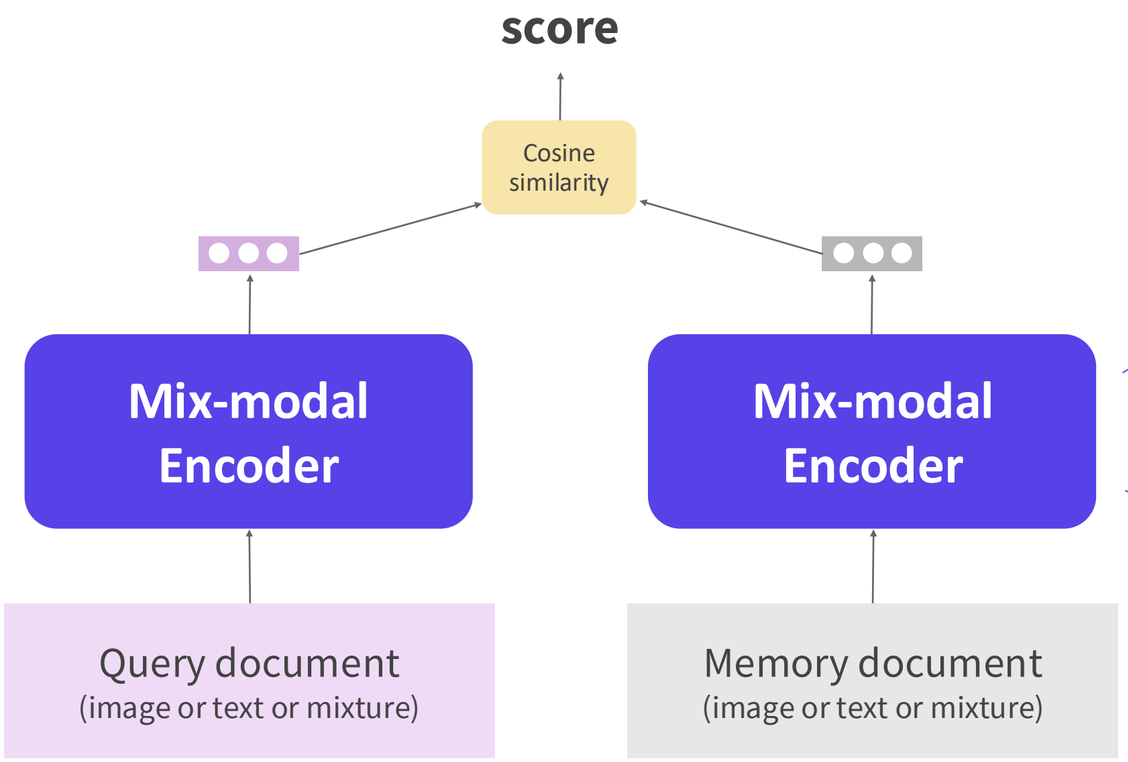
Our Generator: Retrieval Augmented CM3



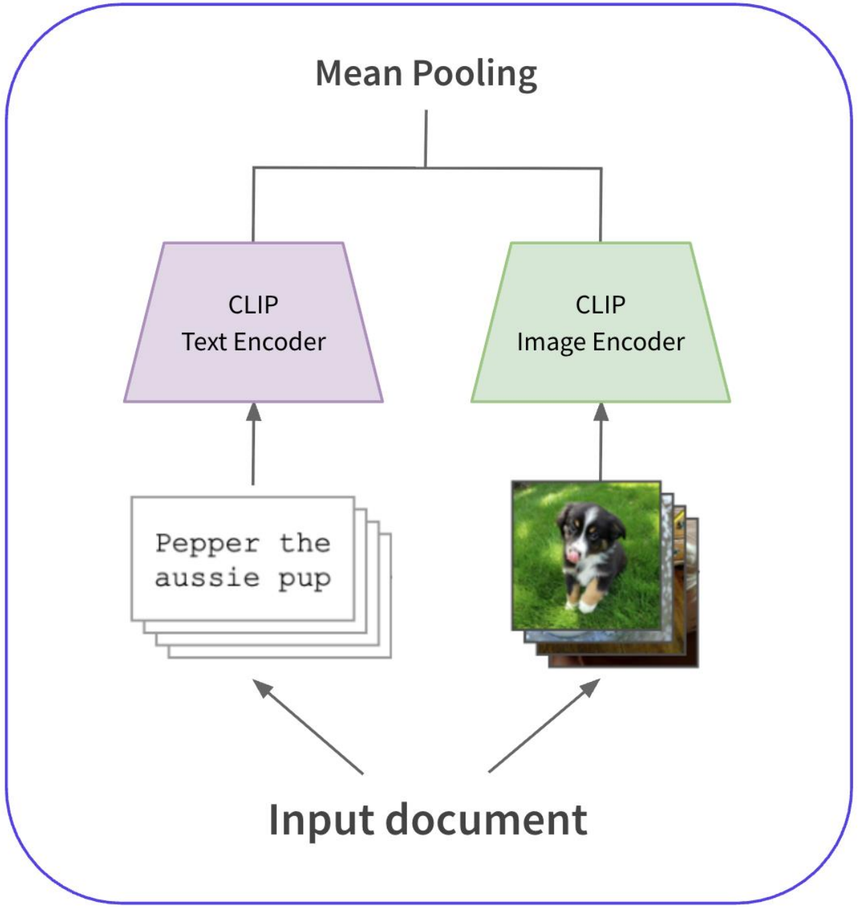
Our Multimodal Retriever

Dense Retriever with Mix-modal Encoder

`f(query, memory) → score`

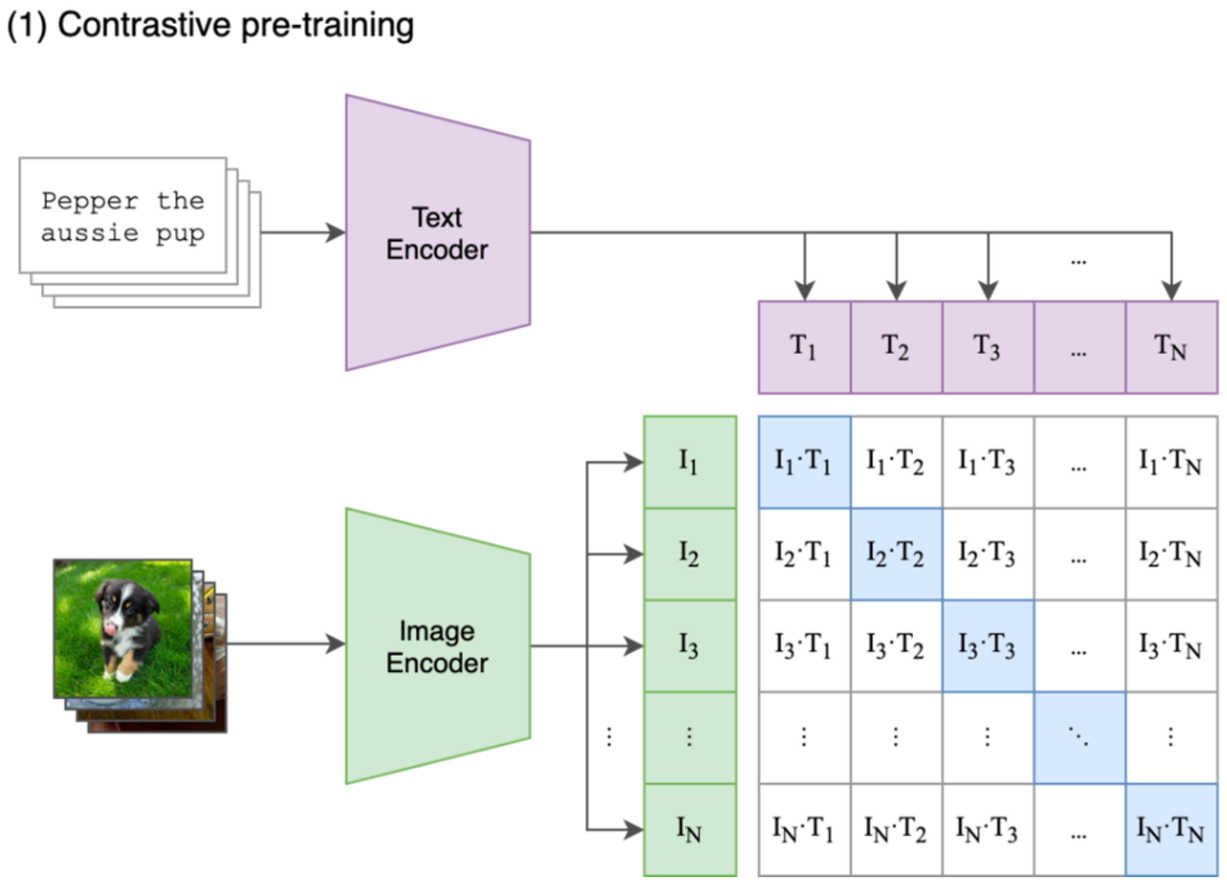


E.g. Extension of CLIP



Background: CLIP

CLIP produces text embeddings and image embeddings in shared vector space.



Radford et al. Learning Transferable Visual Models From Natural Language Supervision

The Problem: Reproducing CLIP

Opaque Training Data

- WIT (400M pairs) is proprietary and unreleased.
- Curation process is only vaguely described.

Black-Box Data Filtering

- OpenAI uses 500K queries to filter CommonCrawl.
- Exact algorithms and thresholds are withheld.

Conflated Quality & Architecture

Cannot determine if CLIP's success stems from its contrastive learning method or the proprietary data curation.

Reproduction Challenges

- "Chicken-and-egg" dependency: using CLIP to filter data for CLIP reproductions.
- Public models (e.g., LAION) fail to match quality.

Introducing Meta CLIP

- CLIP's success: Is it the model, the objective, or the data?
- Our thesis: it's the data — and we can do it without a pre-trained model

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, Christoph Feichtenhofer. **Demystifying CLIP Data**. ICLR 2024 Spotlight.

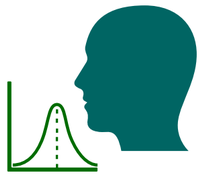
Meta AI

Key Insight: From Curation to Alignment

- CLIP hinted at using 500K queries from WordNet + Wikipedia to **curate** from the Web.
- The real contribution isn't the query list — it's the **distribution** it implies!

Key Insight: From Curation to Alignment

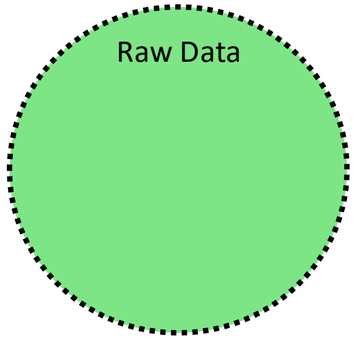
- CLIP hinted at using 500K queries from WordNet + Wikipedia to **curate** from the Web.
- The real contribution isn't the query list — it's the **distribution** it implies!

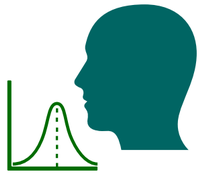


Distribution from Human Experts

Key Insight: From Curation to Alignment

- CLIP hinted at using 500K queries from WordNet + Wikipedia to **curate** from the Web.
- The real contribution isn't the query list — it's the **distribution** it implies!

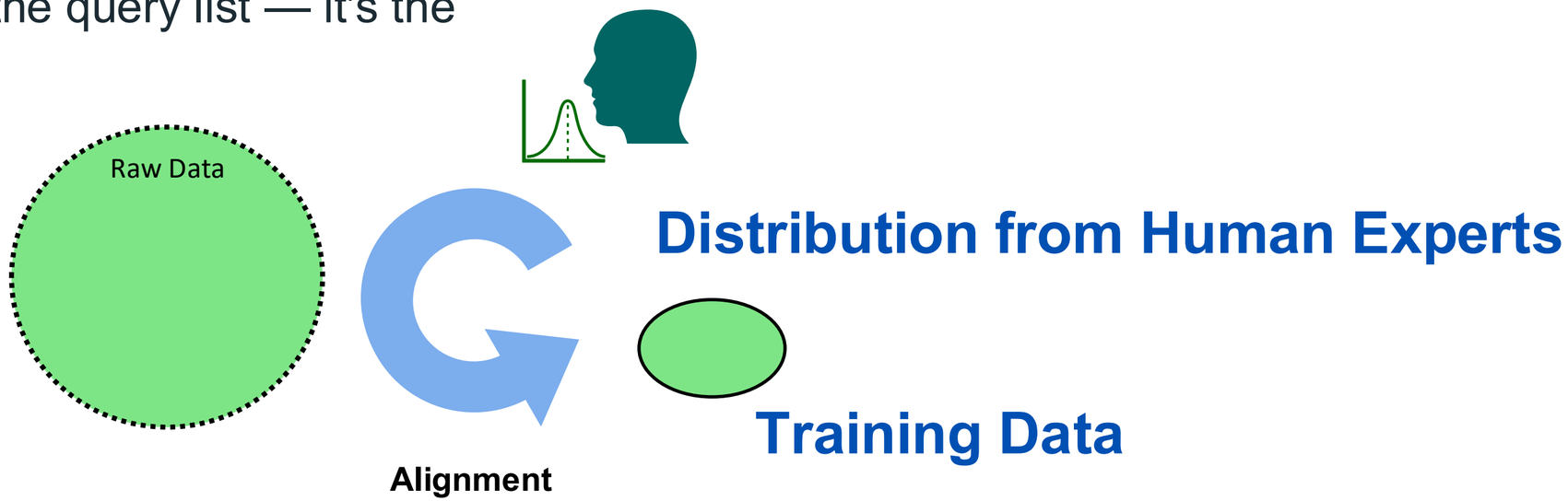




Distribution from Human Experts

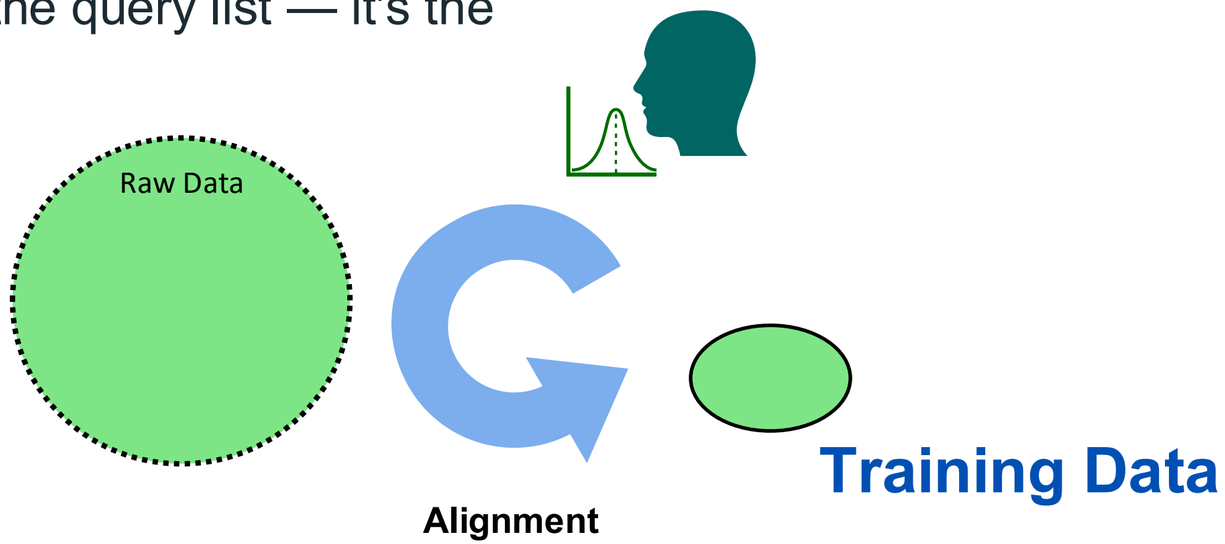
Key Insight: From Curation to Alignment

- CLIP hinted at using 500K queries from WordNet + Wikipedia to **curate** from the Web.
- The real contribution isn't the query list — it's the **distribution** it implies!



Key Insight: From Curation to Alignment

- CLIP hinted at using 500K queries from WordNet + Wikipedia to **curate** from the Web.
- The real contribution isn't the query list — it's the **distribution** it implies!



- Have a known distribution before training.

Metadata-**C**urated **L**anguage-**I**mage **P**re-training

Step 1: Build Metadata

~500K entries from WordNet synsets + Wikipedia unigrams and bigrams

Step 2: Match

Substring-match alt-text of CommonCrawl image-text pairs against metadata entries

Step 3: Balance

Cap the count per metadata entry to prevent long-tail dominance

No neural network in the loop. Purely data-driven. Fully reproducible.

Balancing Matters

Without Balancing

- Common concepts (e.g., "photo", "image") dominate
- Rare concepts are drowned out

With Balancing (freq. threshold=20K)

- Every concept gets fair representation

Key Difference from “Filtering”

It's about controlling the distribution, not just removing noise!

Main Results: ImageNet Zero-Shot Accuracy

Same architecture, same training recipe — only the data differs

ViT-B/16 Results

CLIP (WIT, 400M): **68.3%**

MetaCLIP (400M): **70.8% (+2.5%)**

MetaCLIP (2.5B): **72.4% (+4.1%)**

Scaling Across Model Sizes

ViT-L/14: 76.2% (400M) → 79.2% (2.5B)

ViT-H/14: 80.5% (2.5B)— no bells and whistles

MetaCLIP consistently outperforms standard CLIP across data scales and model architectures.

Why Not Model-Based Filtering?

The LAION Approach

Uses a trained CLIP model to keep pairs with high similarity scores.

Circular Dependency

You need a good model to get good data, just to train a good model. This creates a closed-loop bottleneck.

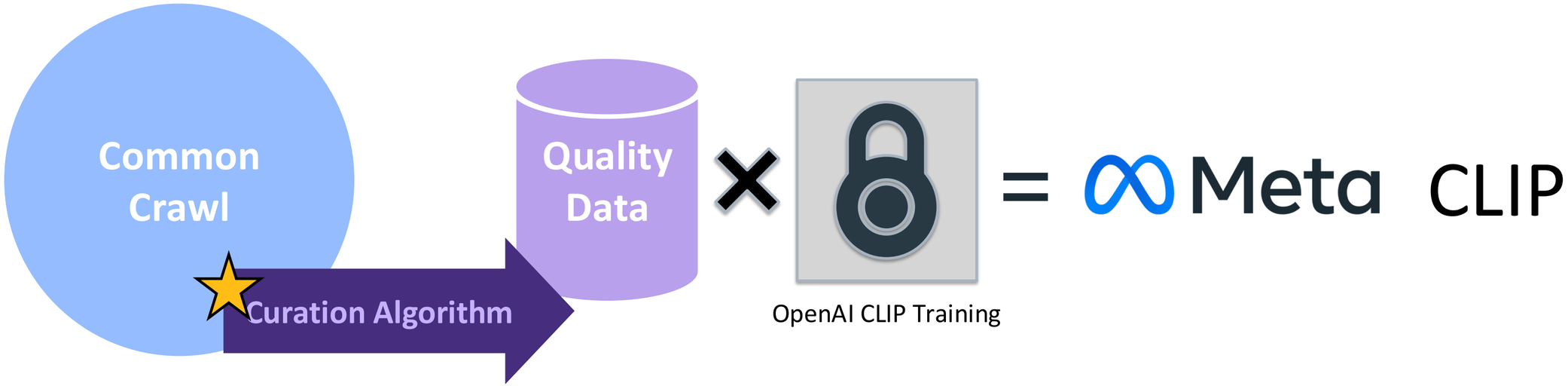
Distribution Bias

Model-filtered data reflects the model's own internal biases rather than the true diversity of the world's distribution.

The MetaCLIP Advantage

Avoids model-in-the-loop entirely — metadata is the only signal used for curation.

Data Alignment at Scale



Foundation for Research and Production at Meta

CLIP

Multimodal LLM

Segmentation

Vision Encoding

MovieGen

Recommendation

Foundation for Research and Production at Meta

CLIP

Llama 3

SAM 3

DINO/Perception Encoder

MovieGen

Recommendation

MetaCLIP Recap

Goal: Open-Source CLIP's Data Curation

Revealing the "secret sauce": it is the data, not the model.

Key Idea: Data Alignment

- WordNet + Wikipedia as quality signal (~500K entries)
- Sub-select from CommonCrawl to match balanced distribution
- No pre-trained CLIP model needed for filtering

Algorithm: CLIP Curation

- Build metadata from WordNet synsets + Wiki n-grams
- Match CC image-text pairs via substring matching
- Balance: cap counts (e.g., 20K) to avoid long-tail dominance

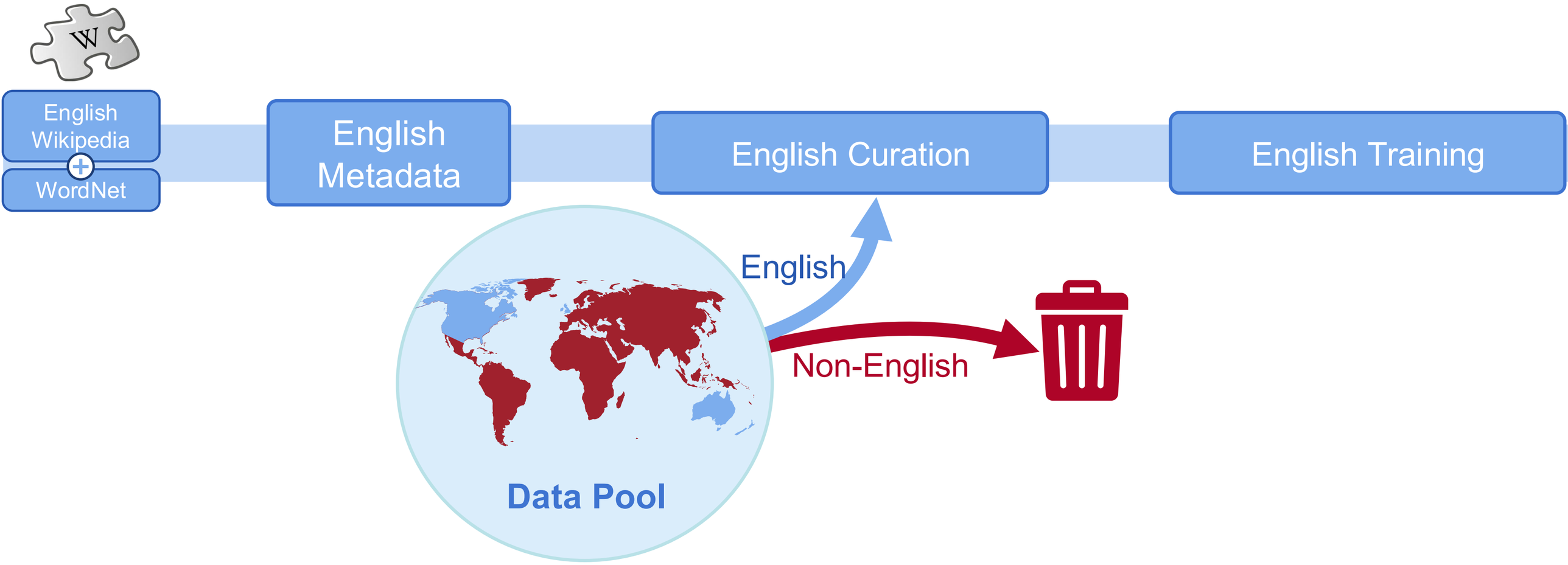
Results & Impact

- Trained on 400M pairs: matches or exceeds OpenAI CLIP on benchmarks
- Scales to 2.5B pairs: achieving new state-of-the-art
- Fully open: pipeline, metadata, and models released

Meta CLIP 2: A Worldwide Scaling Recipe

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, James Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, Xinlei Chen, Zhuang Liu, Saining Xie, Wen-tau Yih, Shang-Wen Li, Hu Xu.
Meta CLIP 2: A Worldwide Scaling Recipe. NeurIPS 2025 Spotlight.

Meta CLIP



- MetaCLIP solved English CLIP data curation.
- But the world is not English-only!

Two Key Challenges in Curating Multilingual Data

Challenge 1: No curation for non-English web data

MetaCLIP's metadata (WordNet + English Wikipedia) is English-only.

Requirements:

- Multilingual metadata
- Language-aware matching

Challenge 2: The curse of multilinguality

Adding non-English data typically hurts English performance.

Past Limitations:

- Reliant on translation
- Special architectural changes

Our Goal: No translation, no special architecture — just better data curation.

Meta CLIP 2: Worldwide

- Wikipedia: 329 language editions
  - Extract unigrams, bigrams (with PMI), and article titles per language
  - Language-specific tokenization (e.g., Chinese segmentation, Thai, Japanese)
- Multilingual WordNet
  - Synsets mapped across languages

Wikipedia

\+

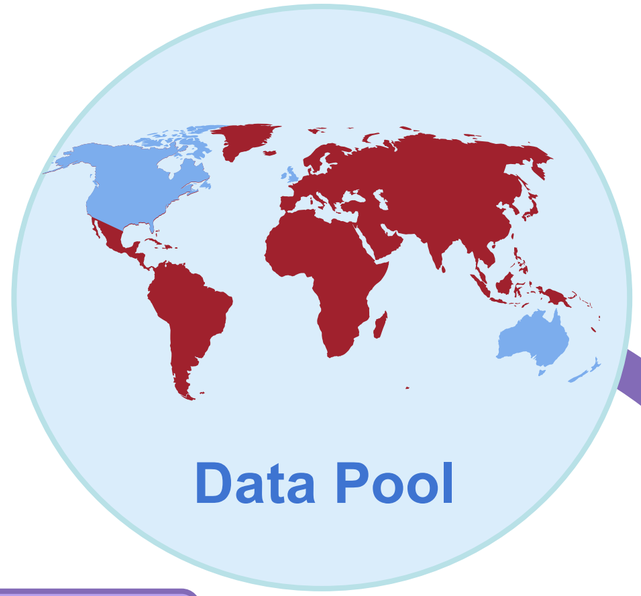
Multilingual WordNet

Worldwide Metadata

329 langs.

Meta CLIP 2: Worldwide

- Per-language metadata limits capped at English counts
  - Unigrams: top 10% by frequency (max ~251K)
  - Bigrams: top by PMI (max ~101K)
  - Titles: top 76% by pageviews (max ~61K)



All

Wikipedia + Multilingual WordNet

Worldwide Metadata

329 langs.

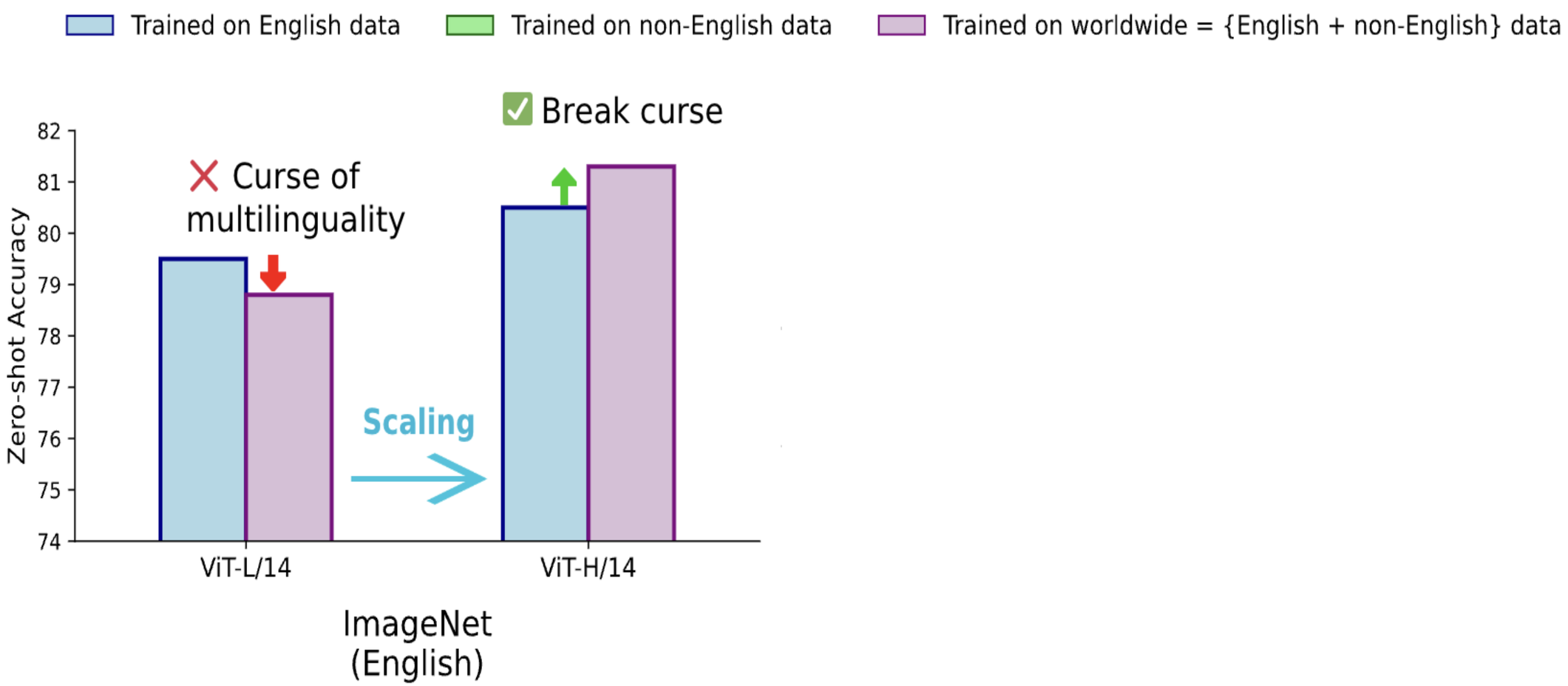
Worldwide Curation

No language filter

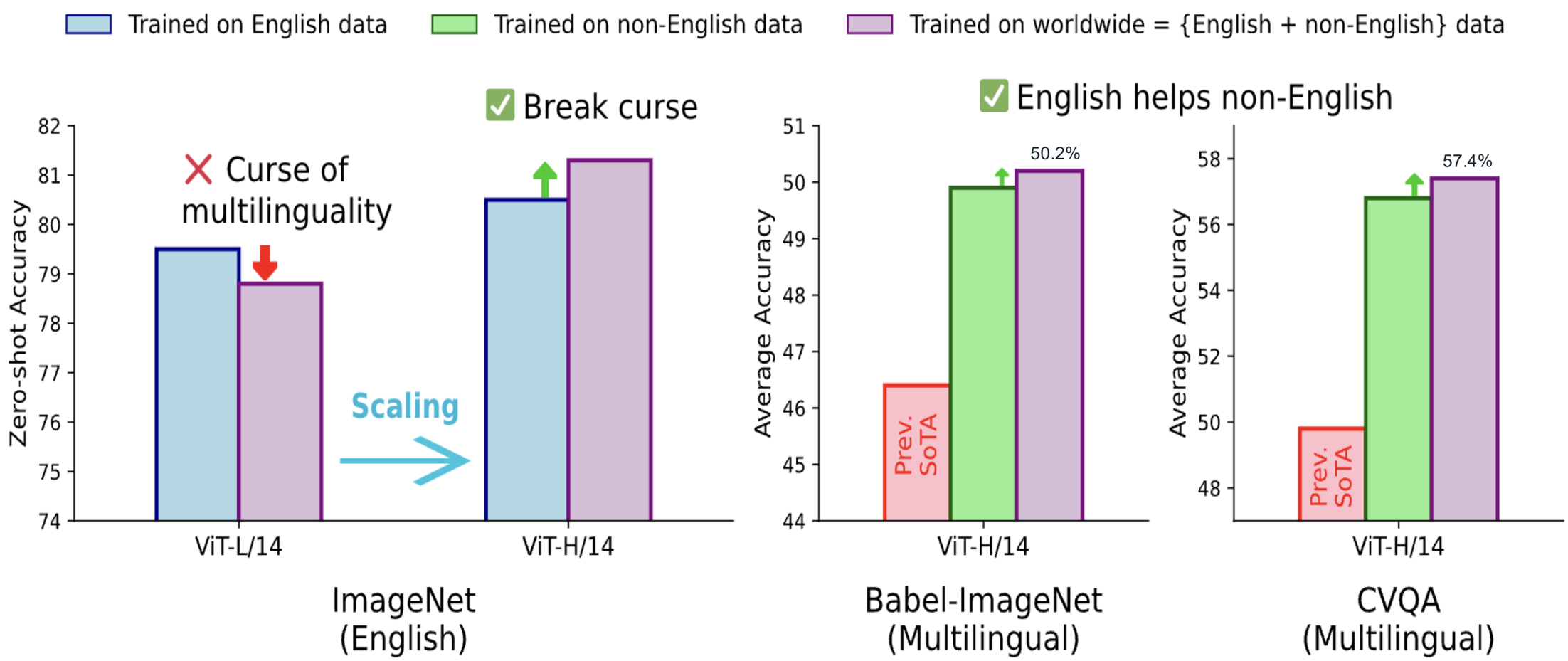
Worldwide Training

2.3x seen pairs

Meta CLIP 2: Worldwide

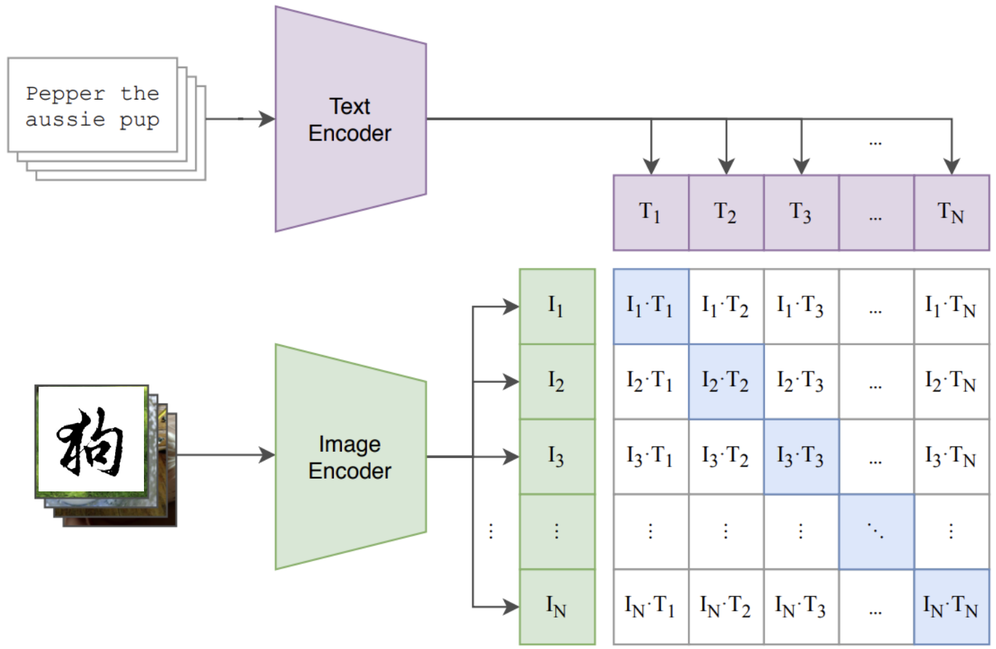


Meta CLIP 2: Worldwide

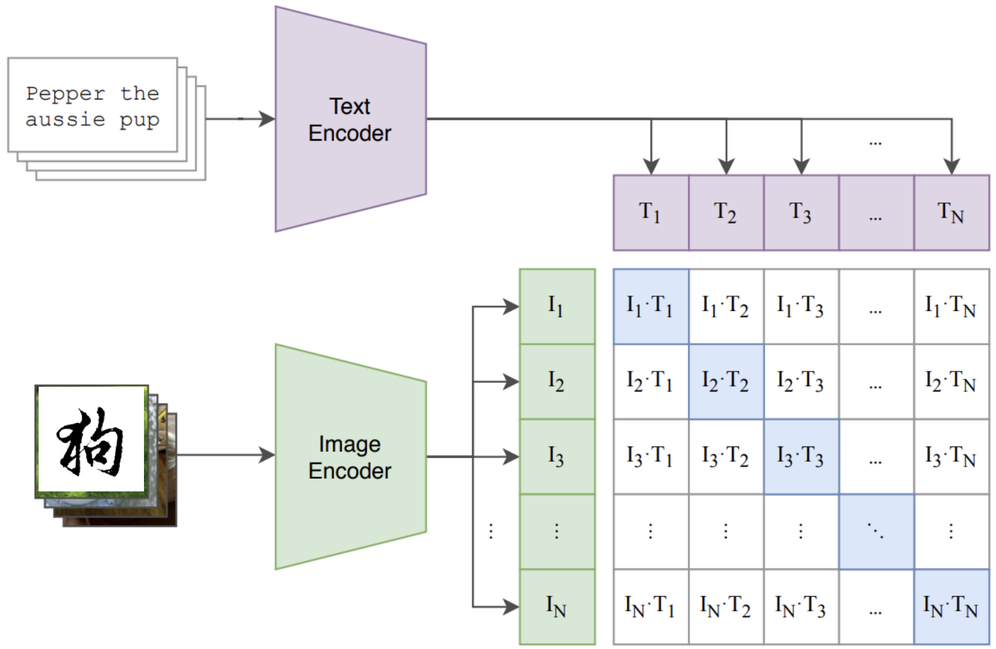


English and non-English data benefit each other.

Meta CLIP 2: Worldwide



Meta CLIP 2: Worldwide



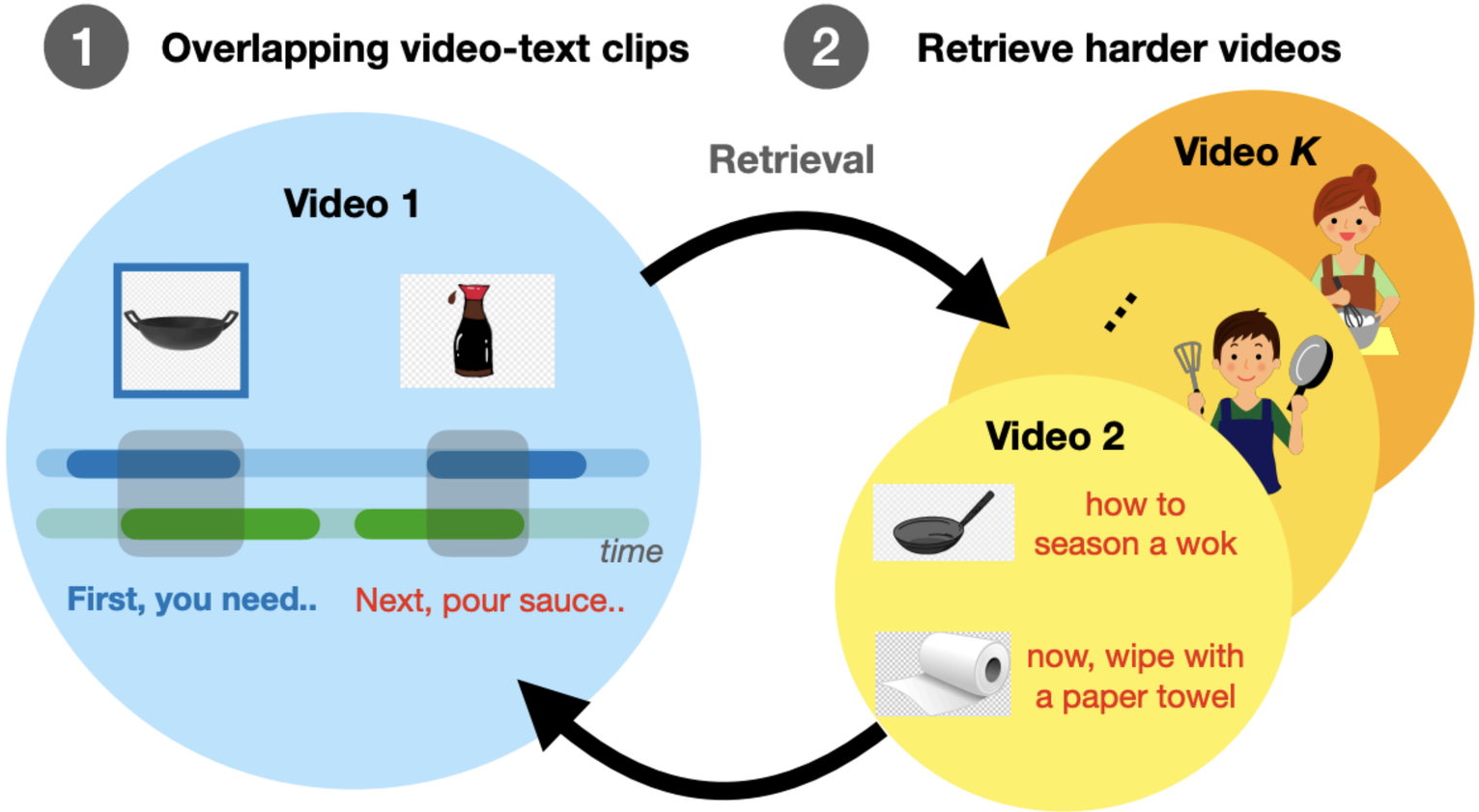
Word	Description	Cosine Sim.
狗	“dog” in Chinese (exactly visualized on image)	0.54325
犬	“dog” in Chinese, literary/ancient usage	0.04636
猫	“cat” in Chinese	0.00025
豺	“jackal” / ‘wild dog” in Chinese	0.03427
狼	“wolf” in Chinese	0.01405
dog	English “dog”	0.08239
diagram	Unrelated word	0.00143
cat	English “cat”	0.00005
puppy	English “puppy”	0.02826
hound	English “hound”	0.05586
いぬ	“dog” in Japanese	0.19320
ねこ	“cat” in Japanese	0.00064

Cross-lingual OCR Transfer

Meta CLIP Series

Meta AI

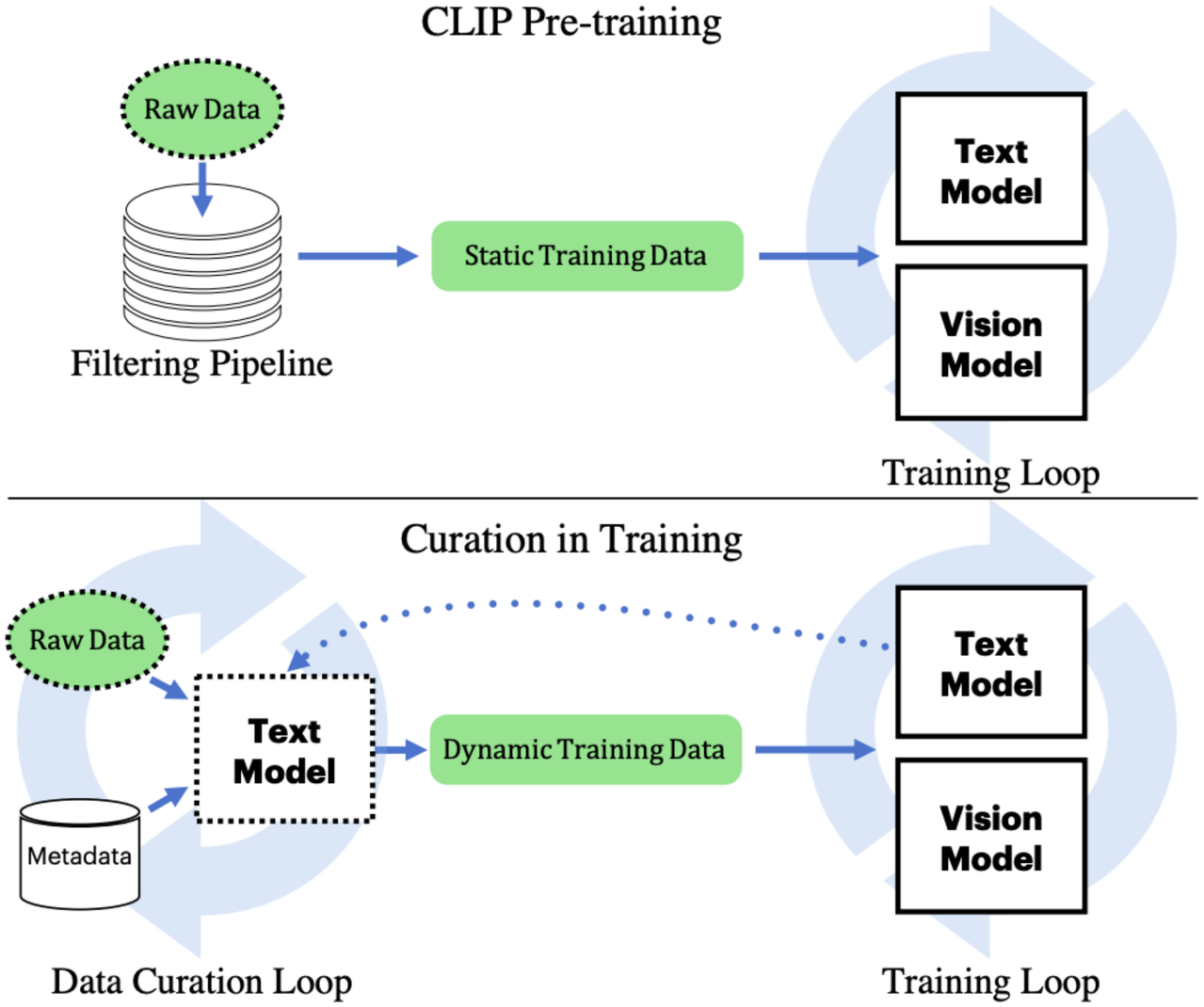
VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding



VideoCLIP: Contrastive learning with hard-retrieved negatives and overlapping positives for video-text pre-training.

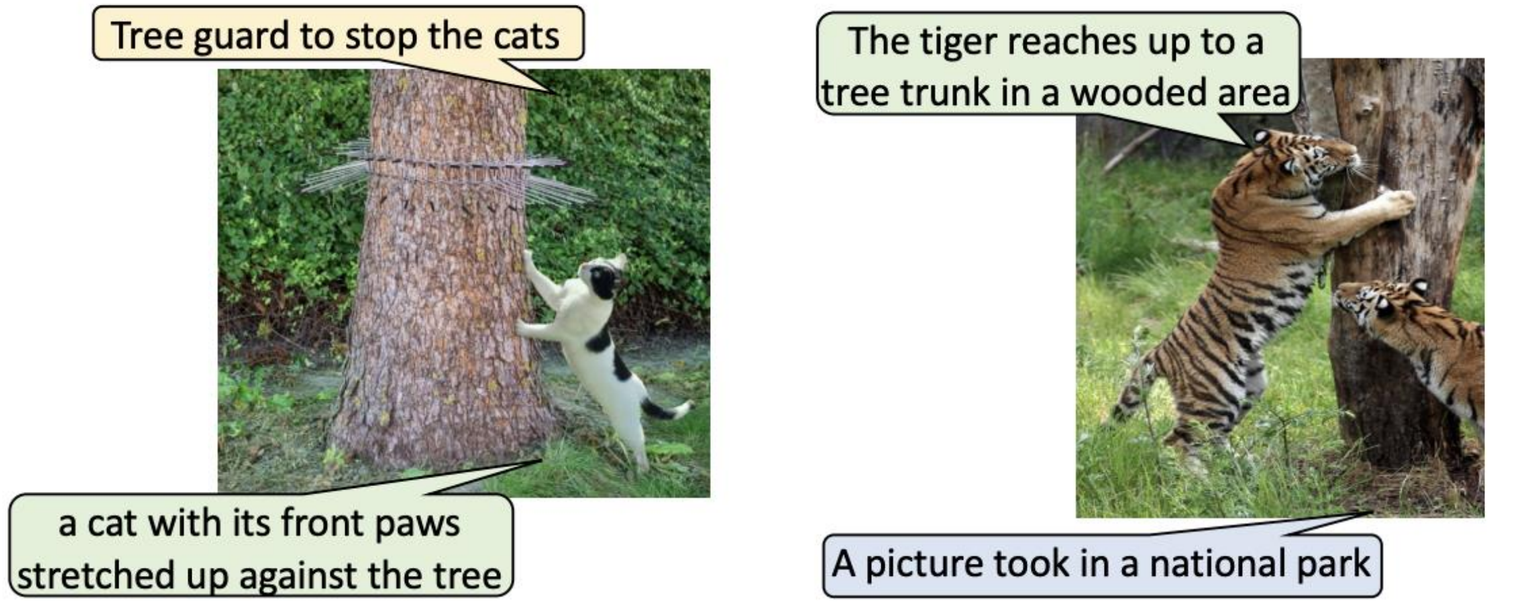
Meta AI

CiT: Curation in Training for Effective Vision-Language Data

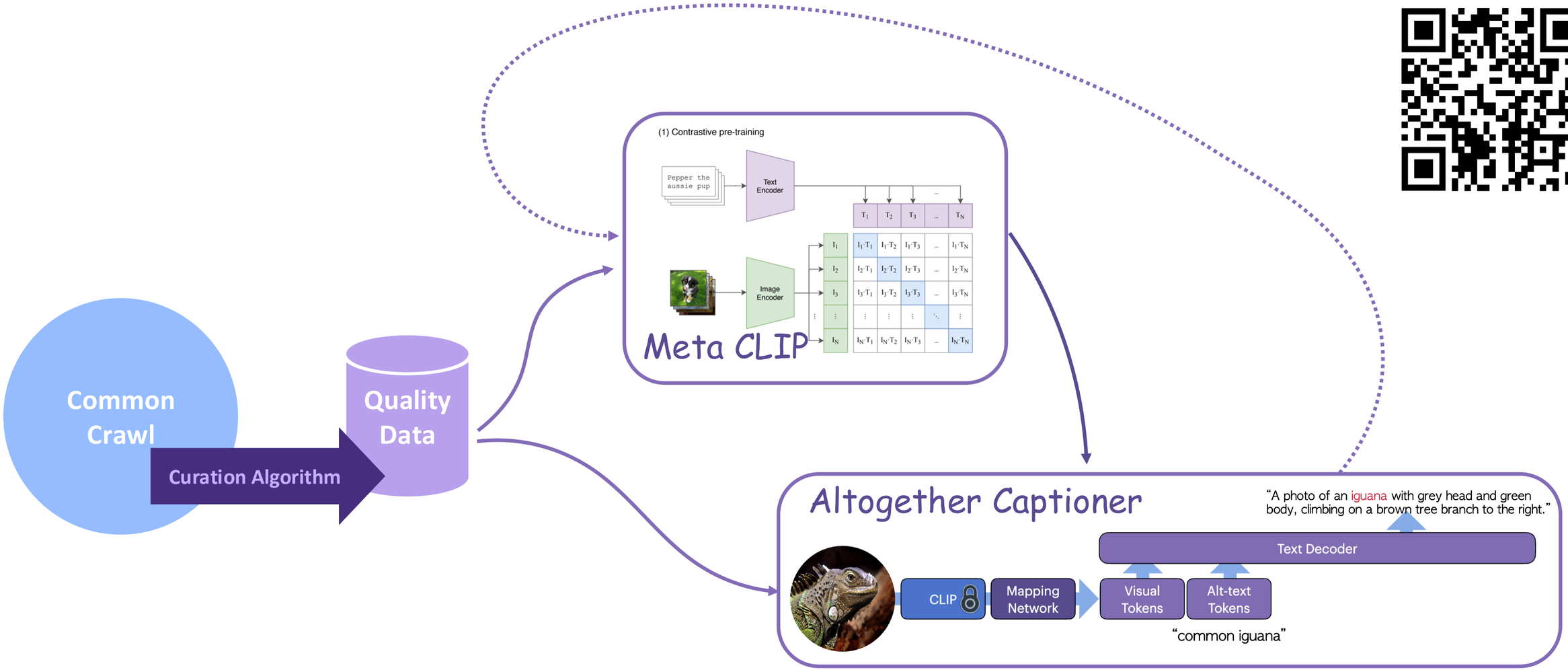


Meta AI

MoDE: CLIP Data Experts via Clustering



Altogether: Image Captioning via Re-aligning Alt-text



Team

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh

Kehan Lyu, Ramya Raghavendra, Lifei Huang,

Shang-Wen Li, **Hu Xu**

Advisor

James Glass, Jason Weston, Luke Zettlemoyer, Xinlei Chen,

Zhuang Liu, Saining Xie, Wen-tau Yih

MetaCLIP Papers

- [ICLR-2024] Demystifying CLIP Data
- [CVPR-2024] MoDE: CLIP Data Experts via Clustering
- [EMNLP-2024] Altogether: Image Captioning via Re-aligning Alt-text
- [NeurIPS-2025] Meta CLIP 2: A Worldwide Scaling Recipe

- Metadata, Code and Model:
  - https://github.com/facebookresearch/MetaCLIP
  - https://meta-clip.github.io