Demystifying CLIP Data

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma,
Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer and Christoph Feichtenhofer
FAIR, NYU

Meta AI

CLIP: **Noisy** Language Supervision

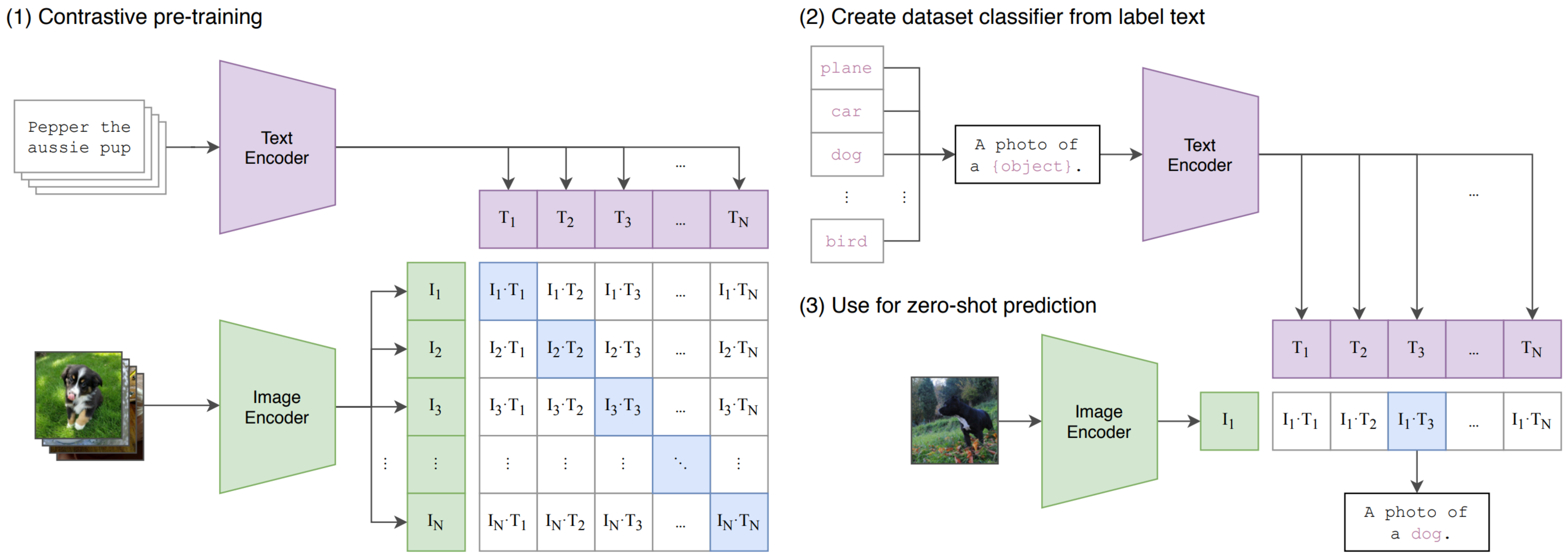


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

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

CLIP: **Noisy** Language Supervision

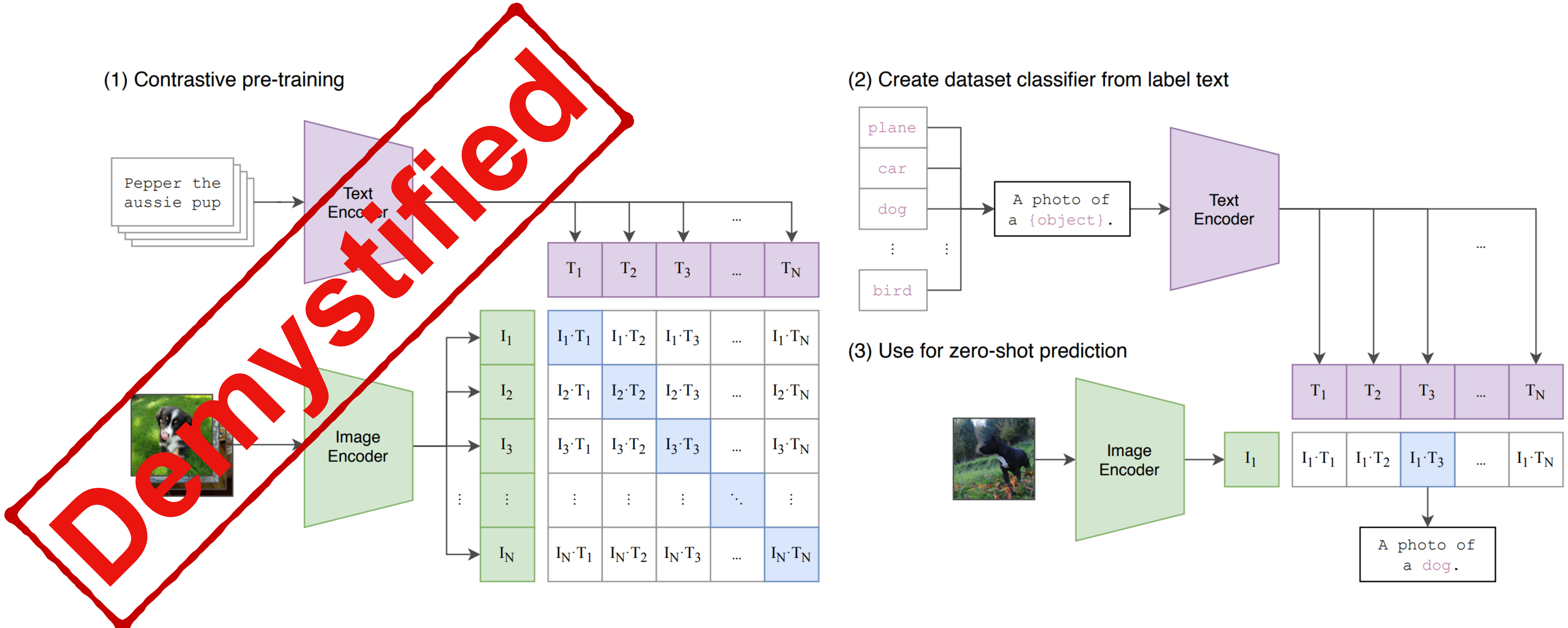


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

Meta AI

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

What's CLIP's main contribution?

Data !

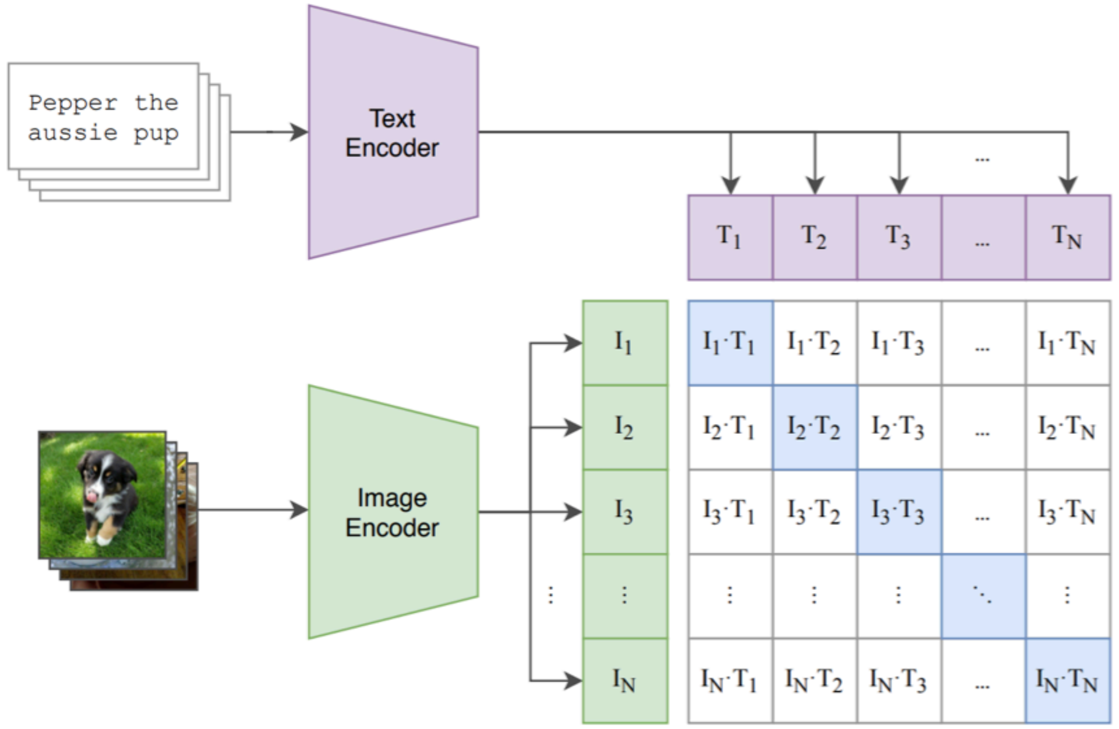
What's CLIP's main contribution?

Data ! Data !!

What's CLIP's main contribution?

Data ! Data !! Data at Scale !!!

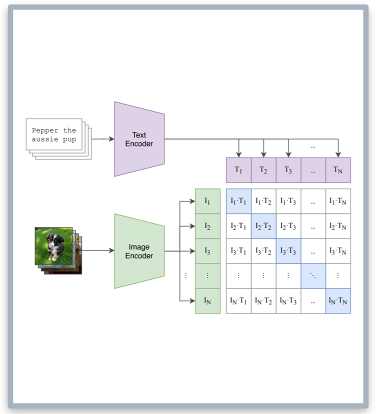
CLIP Model Training



OpenAI CLIP Training

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

From Model to Data Quality



OpenAI CLIP Training

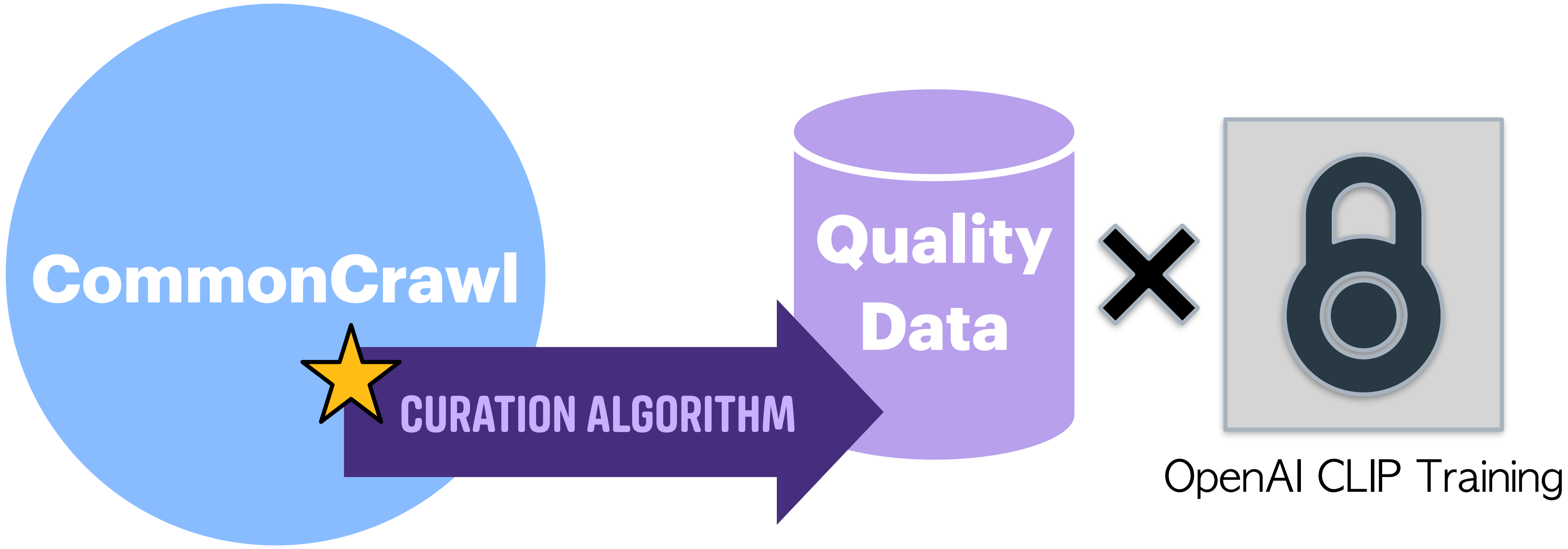
From Model to Data Quality

OpenAI CLIP Training

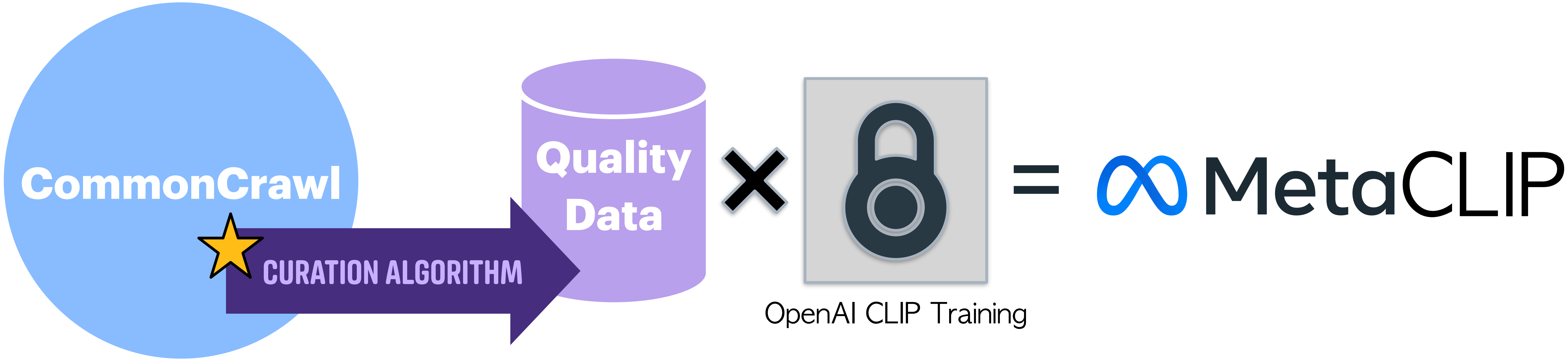
From Model to Data Quality

From Model to Data Quality

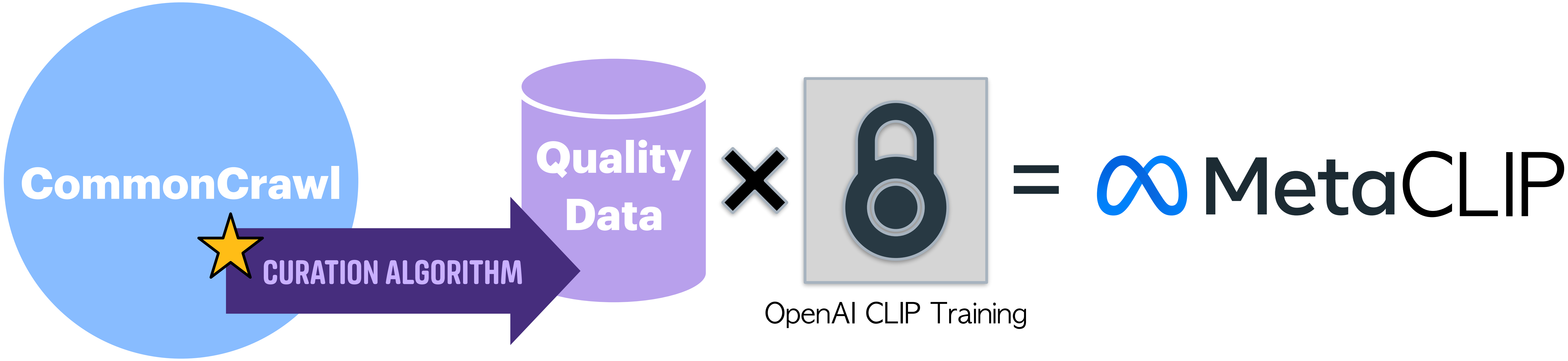
From Model to Data Quality



From Model to Data Quality



From Model to Superhuman Level Data Quality



Towards Superhuman Level Quality

Towards Superhuman Level Quality

Search for visual concepts (metadata)

Towards Superhuman Level Quality

Search for visual concepts (metadata)
that an average human do not know.

Towards Superhuman Level Quality

Search for visual concepts (metadata)
that an average human do not know.

Data with hard information.

Towards Superhuman Level Quality

alt text

Lizard
Chameleon
jacksons chameleon

Curation Probability

Curate!

Towards Superhuman Level Quality

alt text	Curation Probability
Lizard	❌ curation_prob.=0.13
Chameleon	❌ curation_prob.=0.20
jacksons chameleon	✅ curation_prob.=1.00

Curate!

Towards Superhuman Level Quality

"To address this, we constructed a new dataset of 400 million (image, text) pairs collected from a variety of publicly available sources on the Internet. To attempt to cover as broad a set of visual concepts as possible, we *search* for (image, text) pairs as part of the construction process whose text includes one of a set of *500,000 queries* We approximately class balance the results by including *up to 20,000 (image, text) pairs per query*."

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

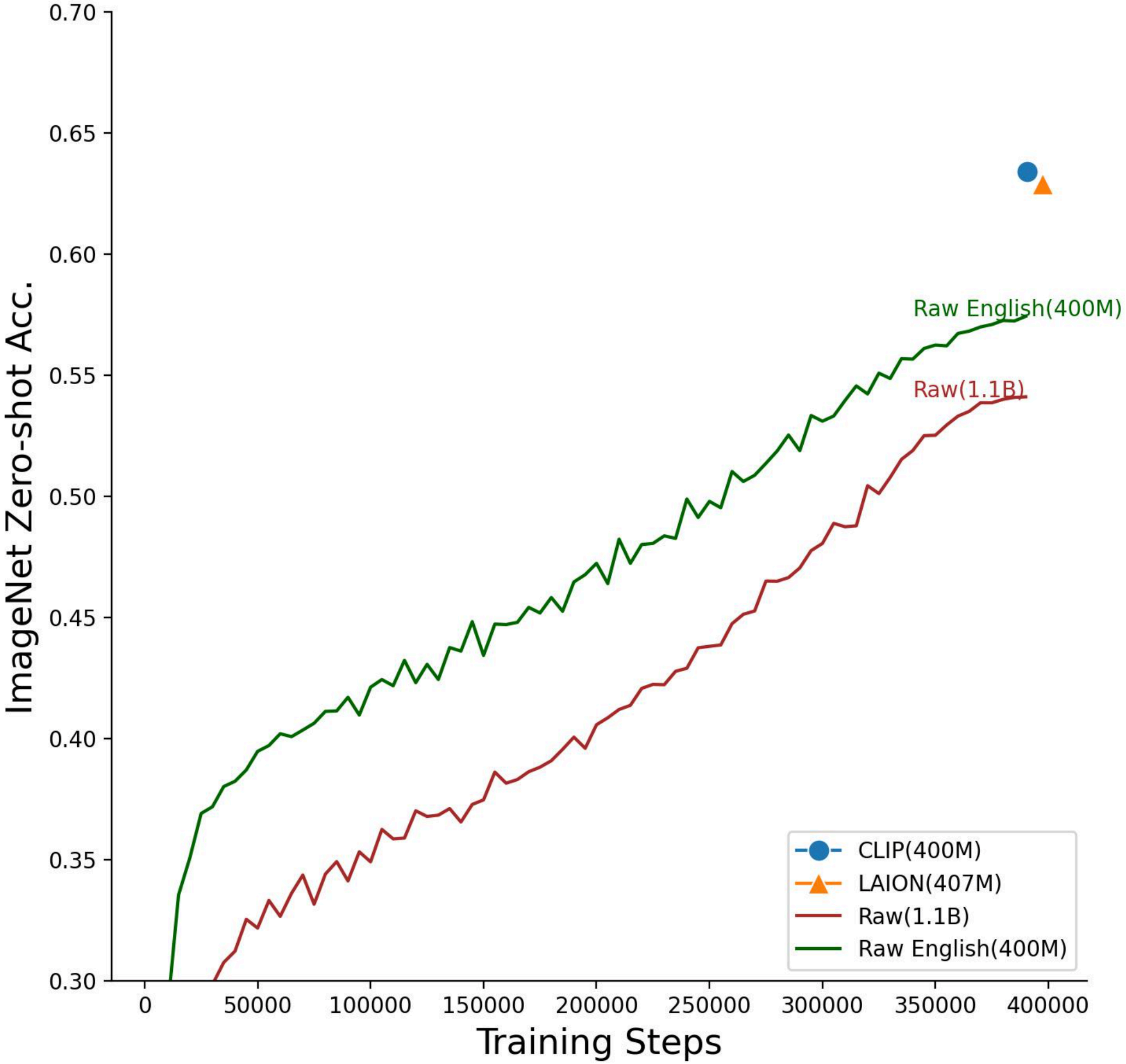
Our Contribution

A scalable algorithm:

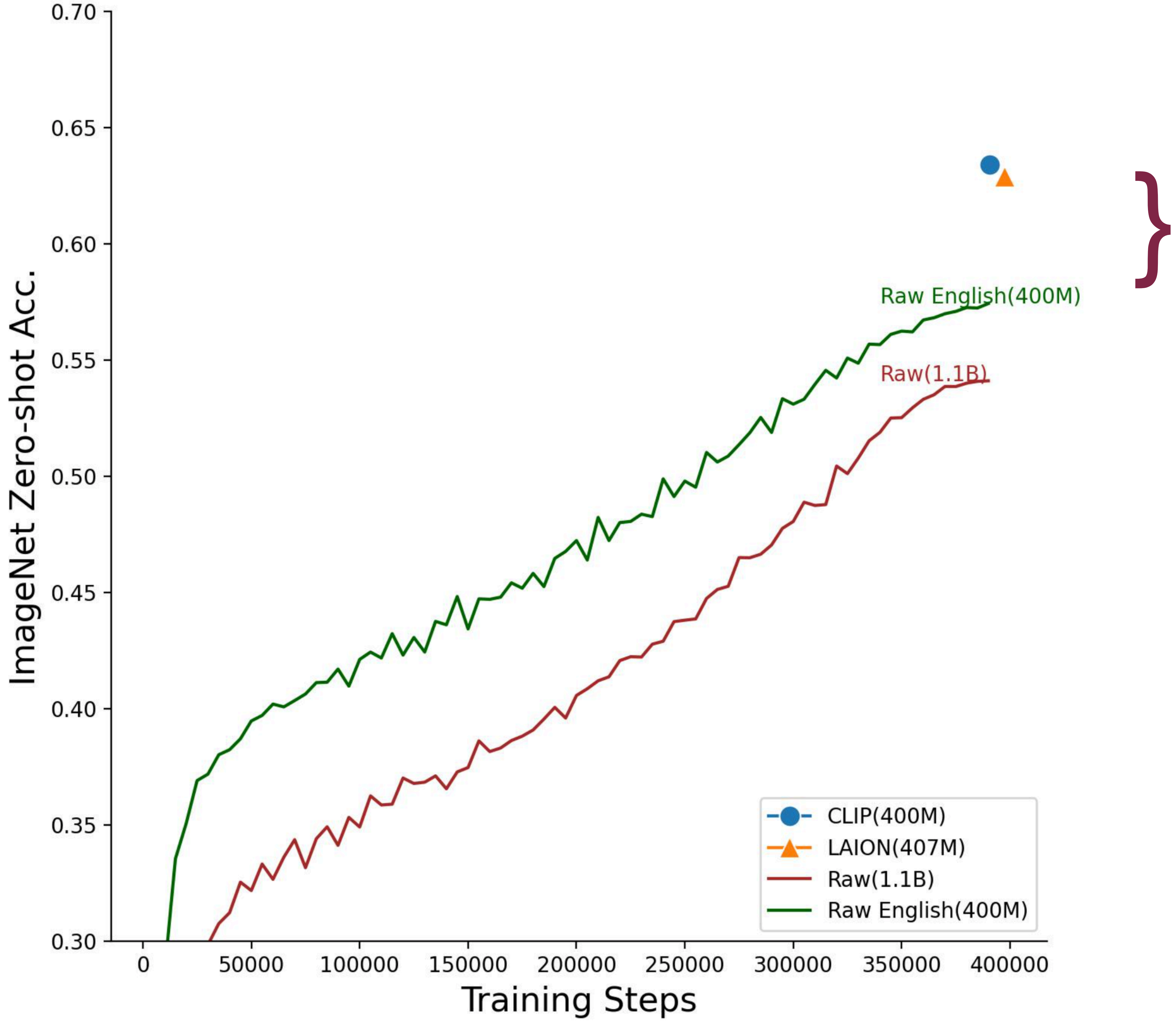
that can run in both a data pipeline and a data loader;

(Check the paper for details)

Naive Scaling to the Internet (CommonCrawl) doesn't work

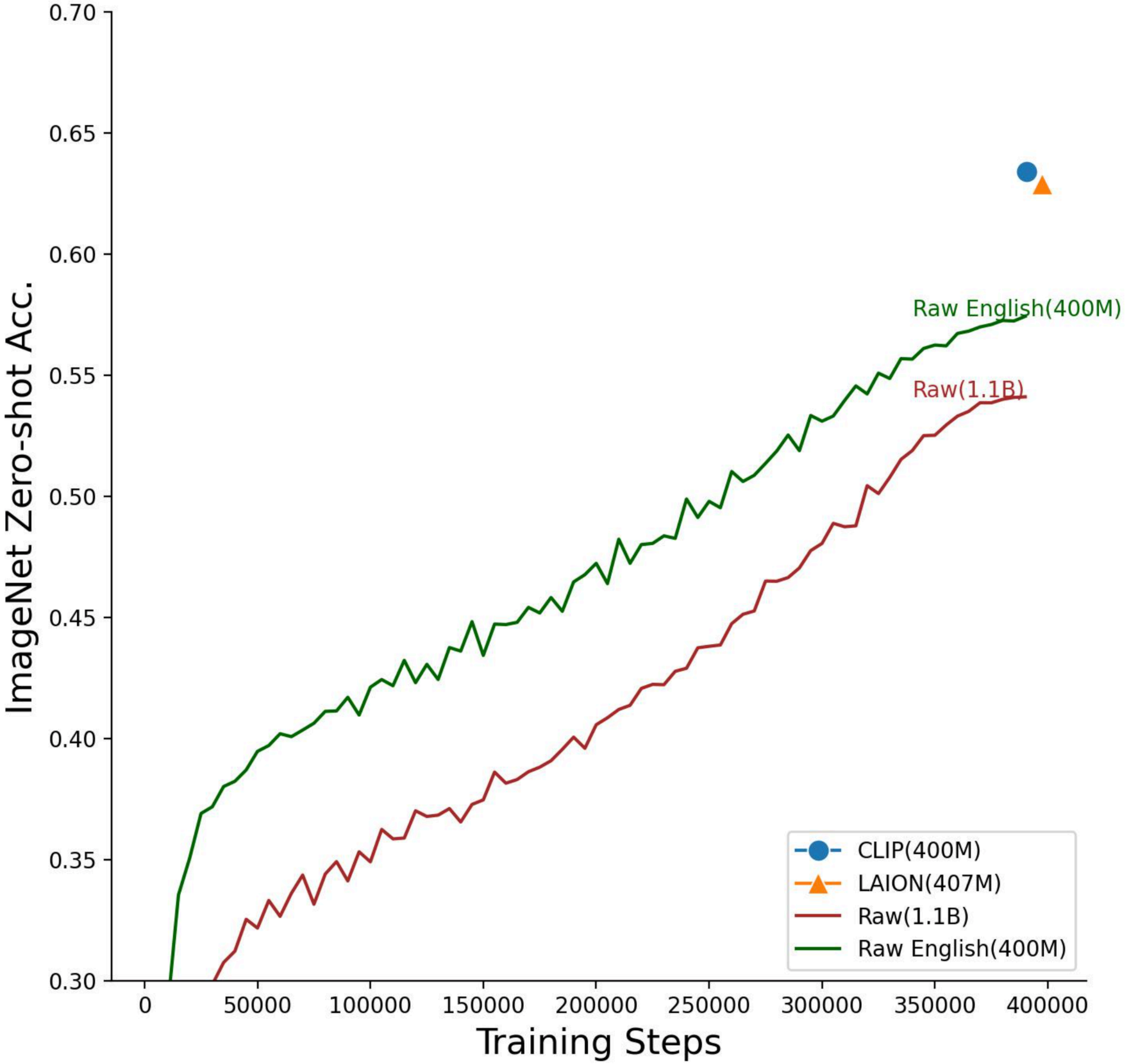


Naive Scaling to the Internet (CommonCrawl) doesn't work



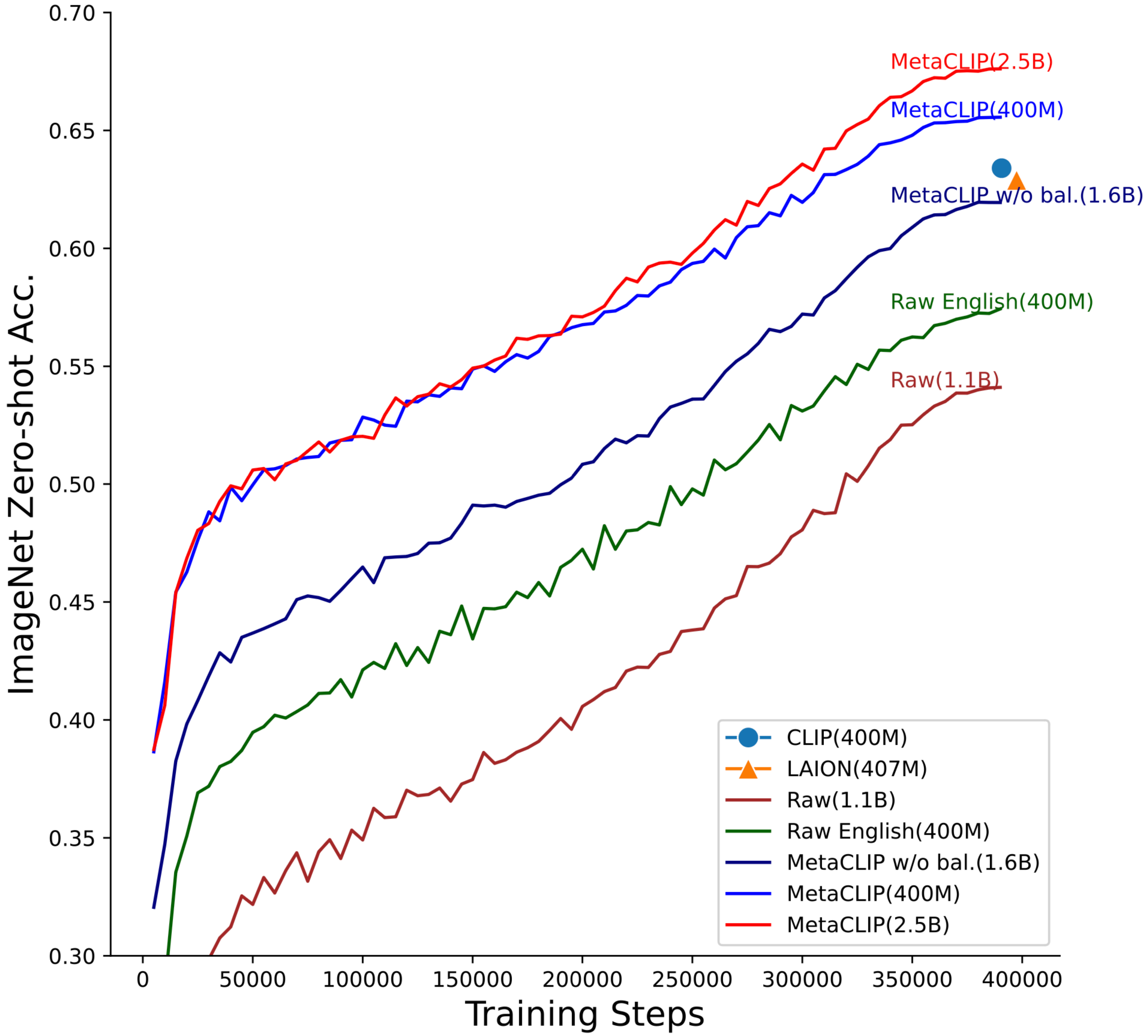
Meta AI

Naive Scaling to the Internet (CommonCrawl) doesn't work



Meta AI

MetaCLIP 400M and 2.5B



Meta AI

Metadata, Code, Model and Demo

- https://github.com/facebookresearch/MetaCLIP
- Also available on Hugging Face Transformers and OpenCLIP.
    - `AutoModel.from_pretrained("facebook/metaclip-h14-fullcc2.5b")`

Meta AI

Effects of Balancing on Data Distribution

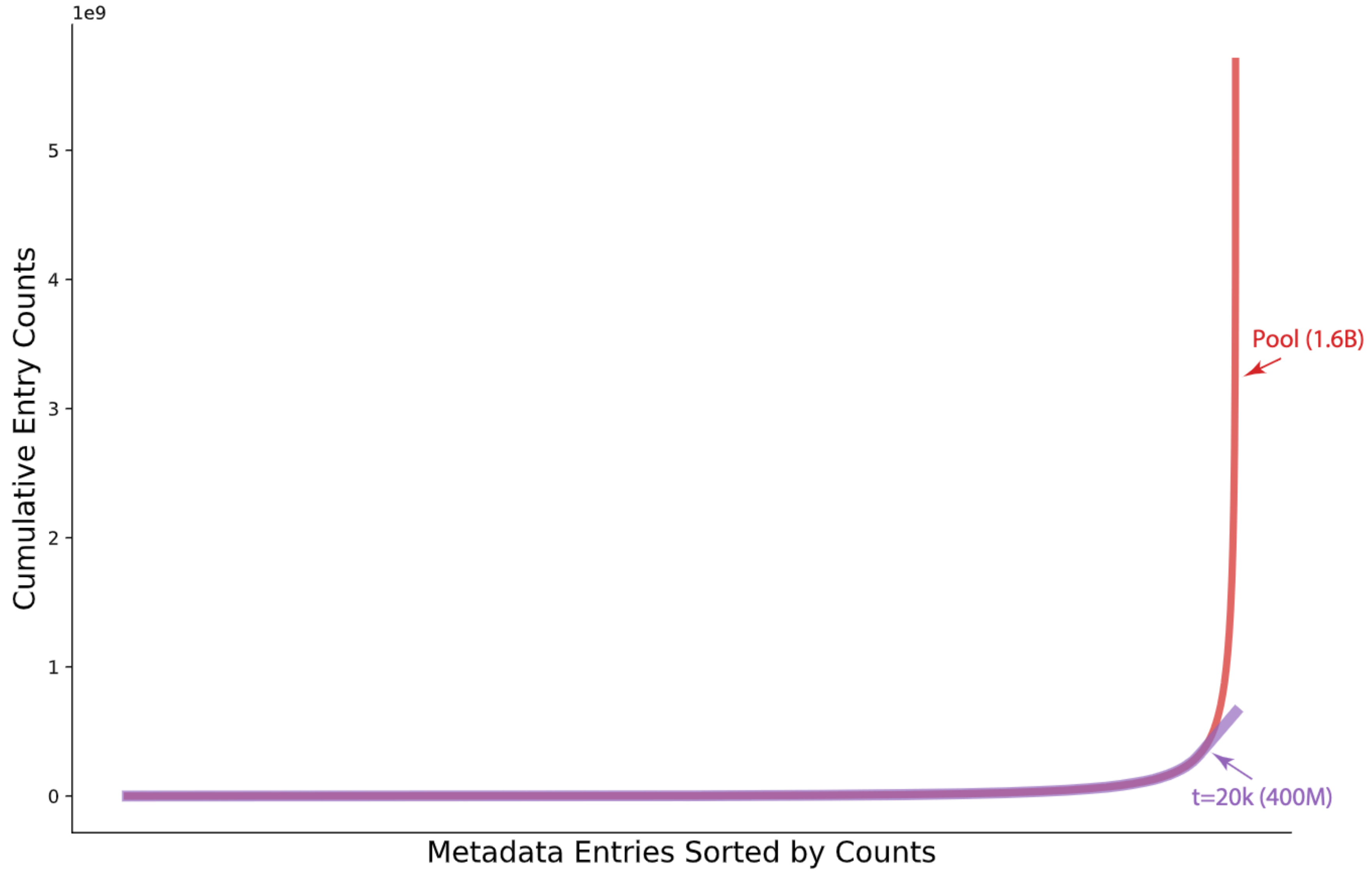


Figure 2: Cumulative sum of counts on entries from *tail to head* on a data pool with 1.6B image-text pairs (5.6B match counts). (1) raw/unbalanced cumulative counts, $t = \infty$; (2) balanced cumulative counts after applying $t = 20$k. The limit t defines the transition of tail/head entries.

1 Metadata

Source	# of Entries	Desc. of Threshold	Threshold
WordNet synsets	86,654	N/A	[ALL] (follow CLIP)
Wiki uni-gram	251,465	Count	100 (follow CLIP)
Wiki bi-gram	100,646	Pointwise Mutual Info.(PMI)	30 (estimated)
Wiki titles	61,235	View Frequency	70 (estimated)

Table 1: Composition of CLIP Metadata.

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

The Algorithm (t=20000 for CLIP)

Algorithm 1: Pseudo-code of Curation Algorithm in Python/NumPy style.

```
# D: raw image-text pairs;
# M: metadata;
# t: max matches per entry in metadata;
# D_star: curated image-text pairs;

D_star = []
# Part 1: sub-string matching: store entry indexes in text.matched_entry_ids and
    output counts per entry in entry_count.
entry_count = substr_matching(D, M)
# Part 2: balancing via indepenent sampling
entry_count[entry_count < t] = t
entry_prob = t / entry_count
for image, text in D:
   for entry_id in text.matched_entry_ids:
      if random.random() < entry_prob[entry_id]:
         D_star.append((image, text))
         break
```

- Our contribution: we turn search queries into independent sampling each data point as curation to scale in a data collection pipeline.