DataComp: In search of the next generation of multimodal datasets

Samir Yitzhak Gadre

Columbia University
IN THE CITY OF NEW YORK

DataComp: In search of the next generation of multimodal datasets

Samir Yitzhak Gadre*[2] Gabriel Ilharco*[1] Alex Fang*[1] Jonathan Hayase[1] Georgios Smyrnis[5]
Thao Nguyen[1] Ryan Marten[7,9] Mitchell Wortsman[1] Dhruba Ghosh[1] Jieyu Zhang[1]
Eyal Orgad[3] Rahim Entezari[10] Giannis Daras[5] Sarah Pratt[1] Vivek Ramanujan[1]
Yonatan Bitton[11] Kalyani Marathe[1] Stephen Mussmann[1] Richard Vencu[6]
Mehdi Cherti[6,8] Ranjay Krishna[1] Pang Wei Koh[1,12] Olga Saukh[10] Alexander Ratner[1,13]
Shuran Song[2] Hannaneh Hajishirzi[1,7] Ali Farhadi[1] Romain Beaumont[6]
Sewoong Oh[1] Alexandros G. Dimakis[5] Jenia Jitsev[6,8]
Yair Carmon[3] Vaishaal Shankar[4] Ludwig Schmidt[1,6,7]

Abstract

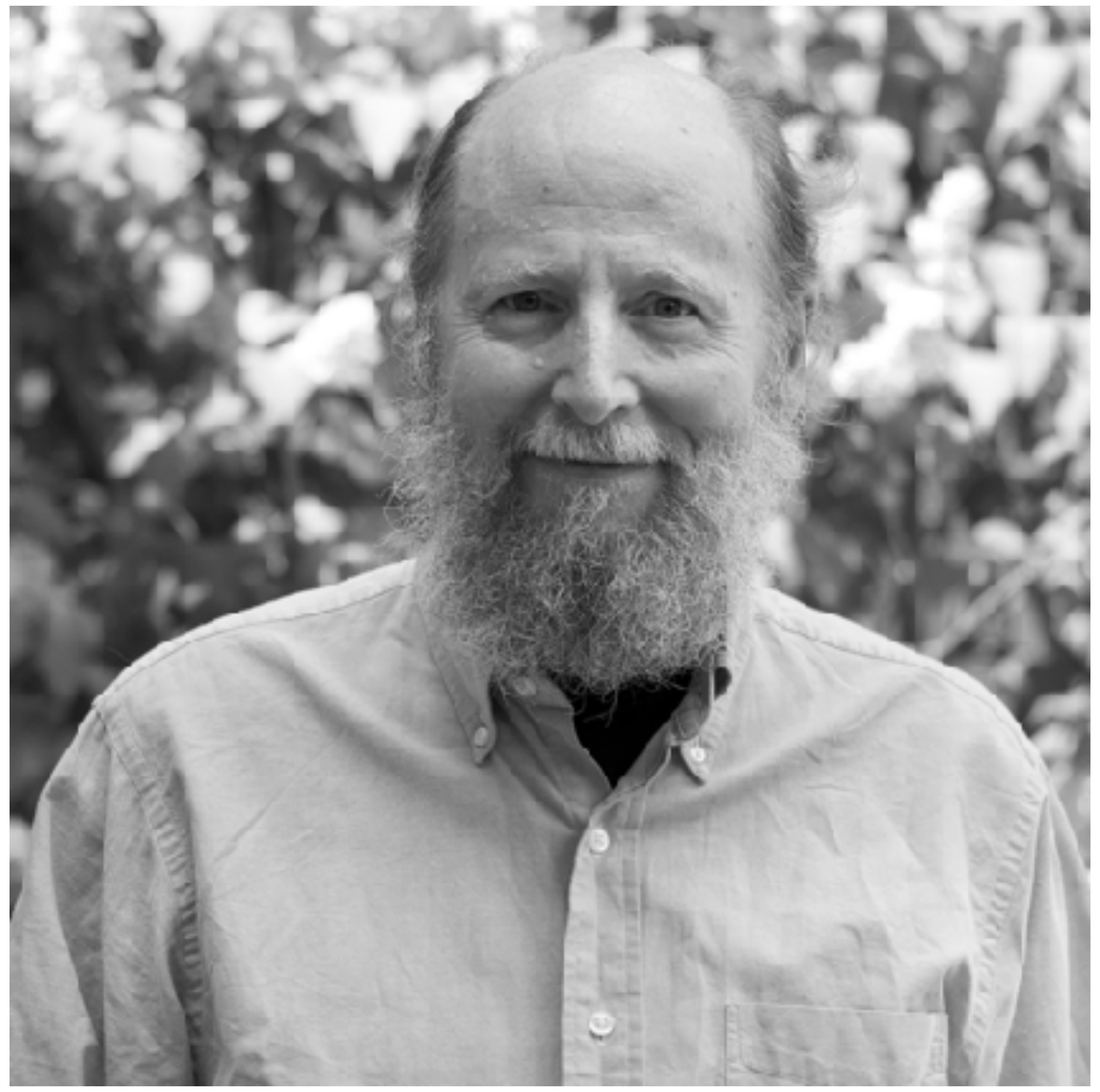
Multimodal datasets are a critical component in recent breakthroughs such as Stable Diffusion and GPT-4, yet their design does not receive the same research attention as model architectures or training algorithms. To address this shortcoming in the ML ecosystem, we introduce DataComp, a testbed for dataset experiments centered around a new candidate pool of 12.8 billion image-text pairs from Common Crawl. Participants in our benchmark design new filtering techniques or curate new data sources and then evaluate their new dataset by running our standardized CLIP training code and testing the resulting model on 38 downstream test sets. Our benchmark consists of multiple compute scales spanning four orders of magnitude, which enables the study of scaling trends and makes the benchmark accessible to researchers with varying resources. Our baseline experiments show that the DataComp workflow leads to better training sets. In particular, our best baseline, DataComp-1B, enables training a CLIP ViT-L/14 from scratch to 79.2% zero-shot accuracy on ImageNet, outperforming OpenAI's CLIP ViT-L/14 by 3.7 percentage points while using the same training procedure and compute. We release DataComp and all accompanying code at www.datacomp.ai.

Gadre et al. DataComp: In search of the next generation of multimodal datasets. NeurIPS D&B 2023.

The bitter lesson

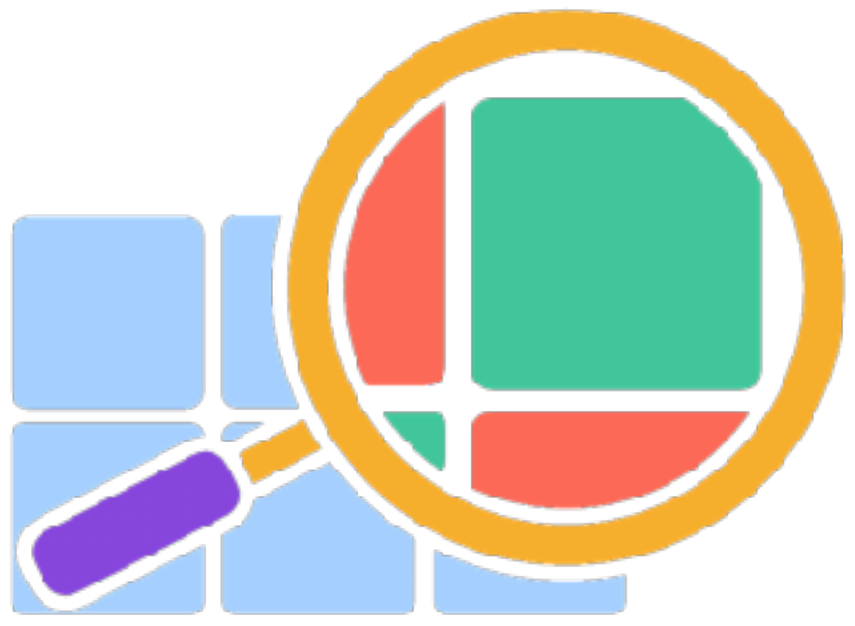
"The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin."

Rich Sutton. The bitter lesson. 2019.

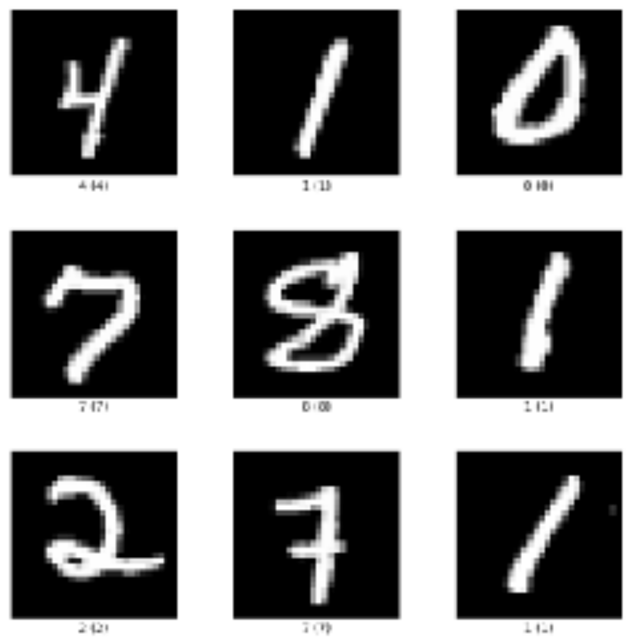


The data lesson addendum?

"The biggest lesson that can be read from … AI research is that general methods that leverage computation are ultimately the most effective," **especially when applied in conjunction with rigorous dataset construction.**

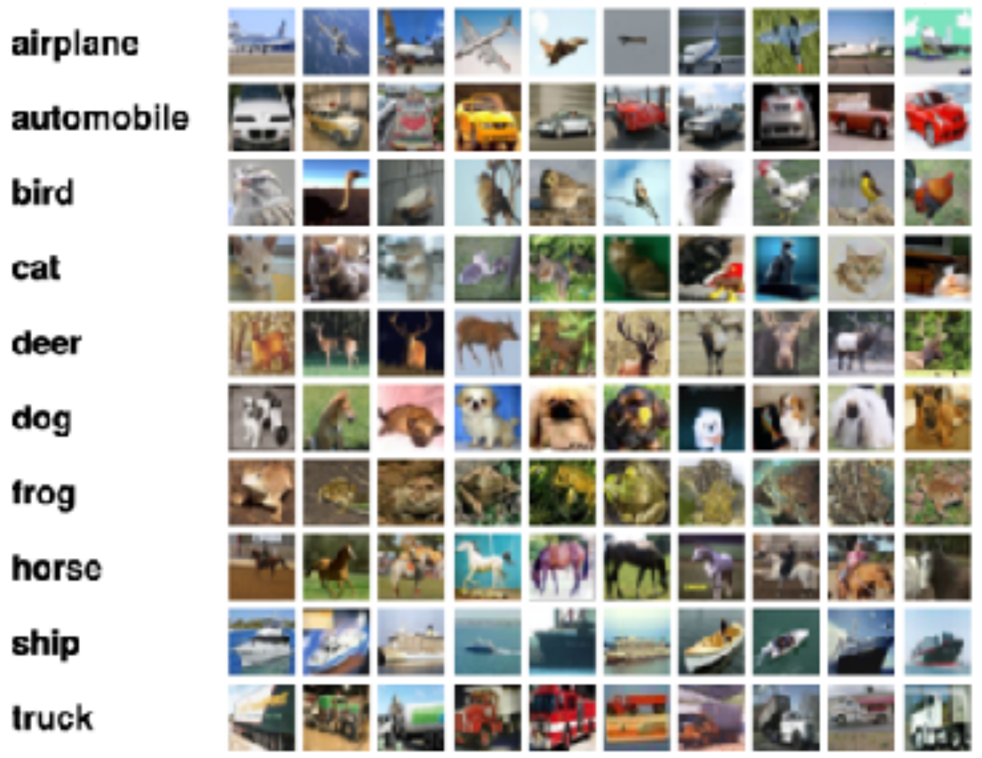


Datasets are the foundation of progress in ML



MNIST (1994)

Convolutional neural networks



CIFAR-10 (2009)

Training on GPUs

ImageNet (2012)

Deep learning resurgence, ResNets, transfer learning, etc.

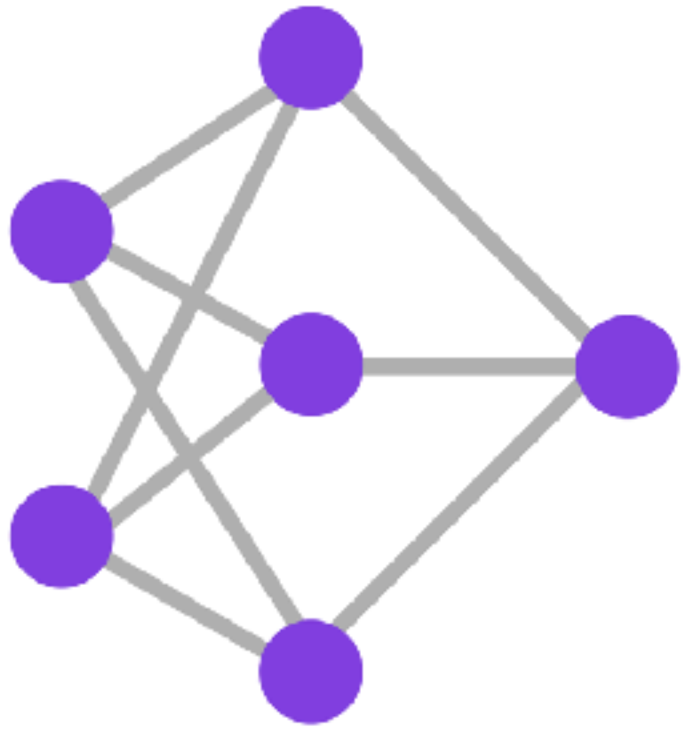
OpenAI

WebImageText (2021)

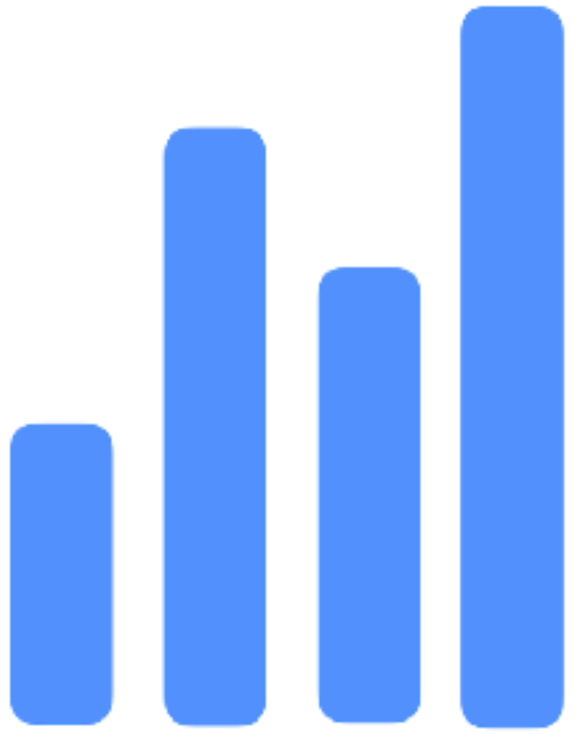
Zero-shot classification (CLIP), text-guided image generation (DALL-E)

The standard ML research pipeline

A. Select datasets



B. Train



C. Evaluate

This pipeline has produced better models

- Architectures
- Optimizers
- Normalization
- Tuned hyperparameters
- Activation functions
- Weight initialization schemes
- Stable training tricks

But how much performance are we leaving on the table by fixing datasets?

Dataset iteration for better models?

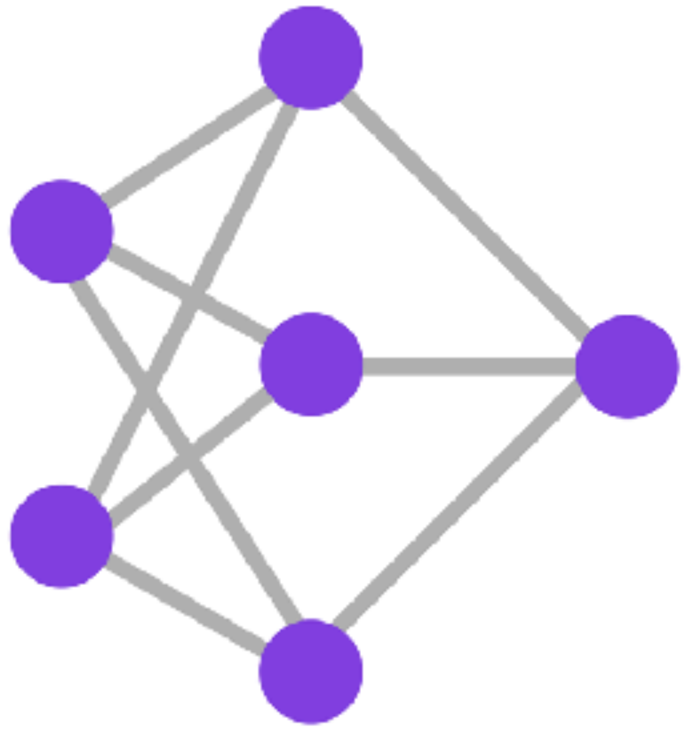
- Diversity?
- Hard vs. easy examples?
- Class distributions?
- Label quality?
- Scale?

Clearly this is a platypus says ImageNet

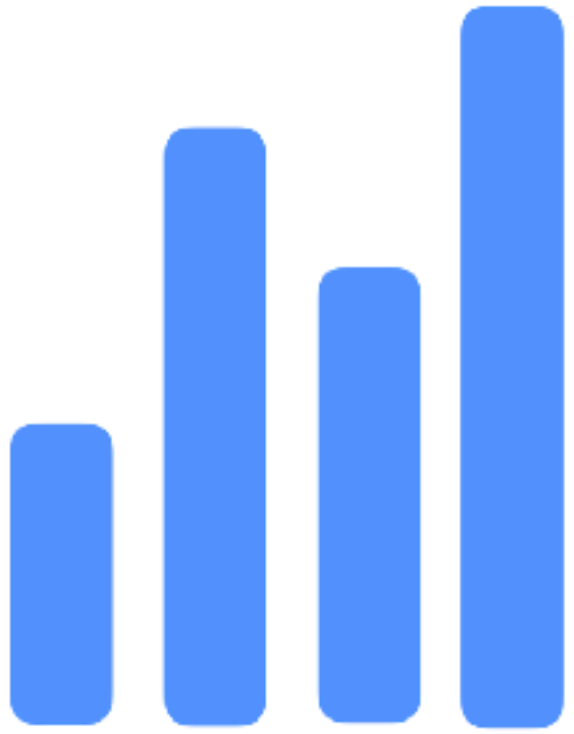
DataComp is a benchmark for dataset development

Enter DataComp

A. Select datasets

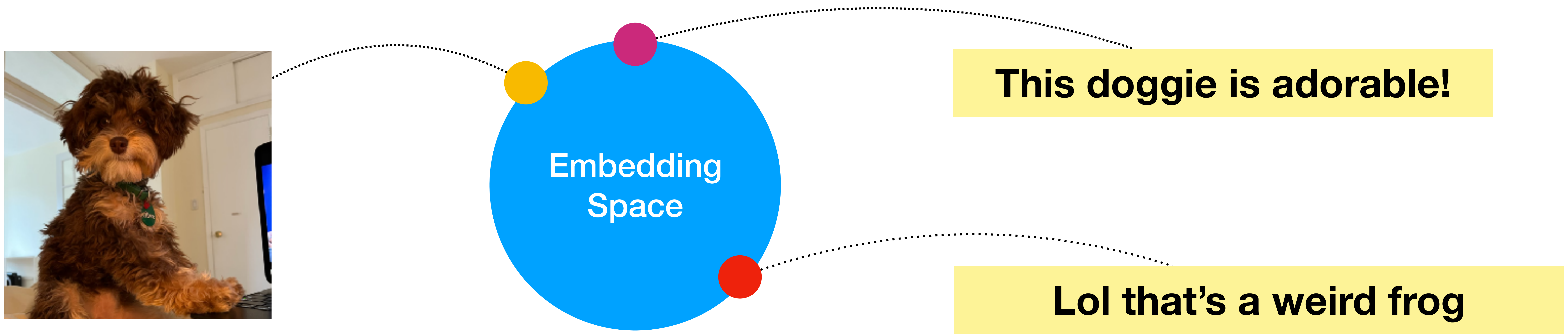


B. Train



C. Evaluate

Interlude: DataComp targets CLIP training



Radford et al. Learning Transferable Vision Models From Natural Language Supervision. ICML 2021.

Interlude: CLIP for zero-shot inference

- With vision-language features we can create *arbitrary* image classifiers.

Input image

Prompts to create classifier

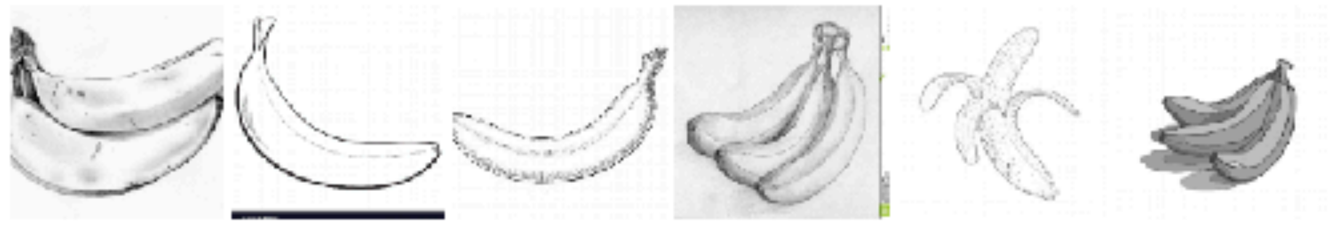
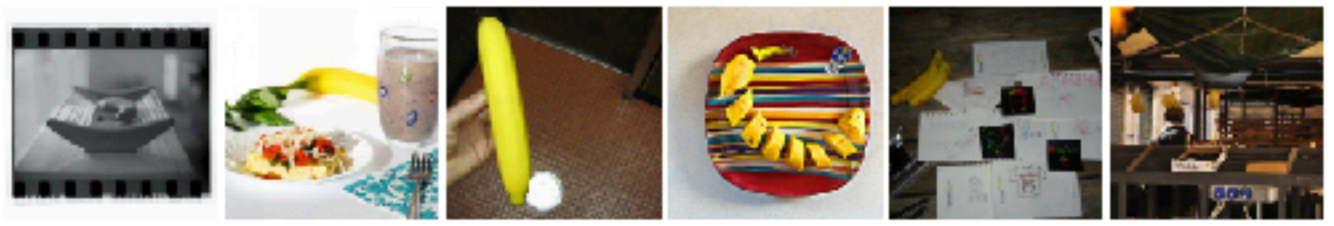
Similarity scores give class label

A photo of a dog.

A photo of a frog.

Radford et al. Learning Transferable Vision Models From Natural Language Supervision. ICML 2021.

Interlude: Why CLIP?

Dataset	ImageNet ResNet101	CLIP ViT-L
 ImageNet	76.2%	76.2%
 ImageNet V2	64.3%	70.1%
 ImageNet Rendition	37.7%	88.9%
 ObjectNet	32.6%	72.3%
 ImageNet Sketch	25.2%	60.2%
 ImageNet Adversarial	2.7%	77.1%

https://openai.com/research/clip

Interlude: Why CLIP?

- Many model vision models utilize CLIP backbones for V&L tasks, segmentation, detection, image generation, embodied tasks, etc.

- Reasonable signal that improving CLIP models also leads to downstream model gains

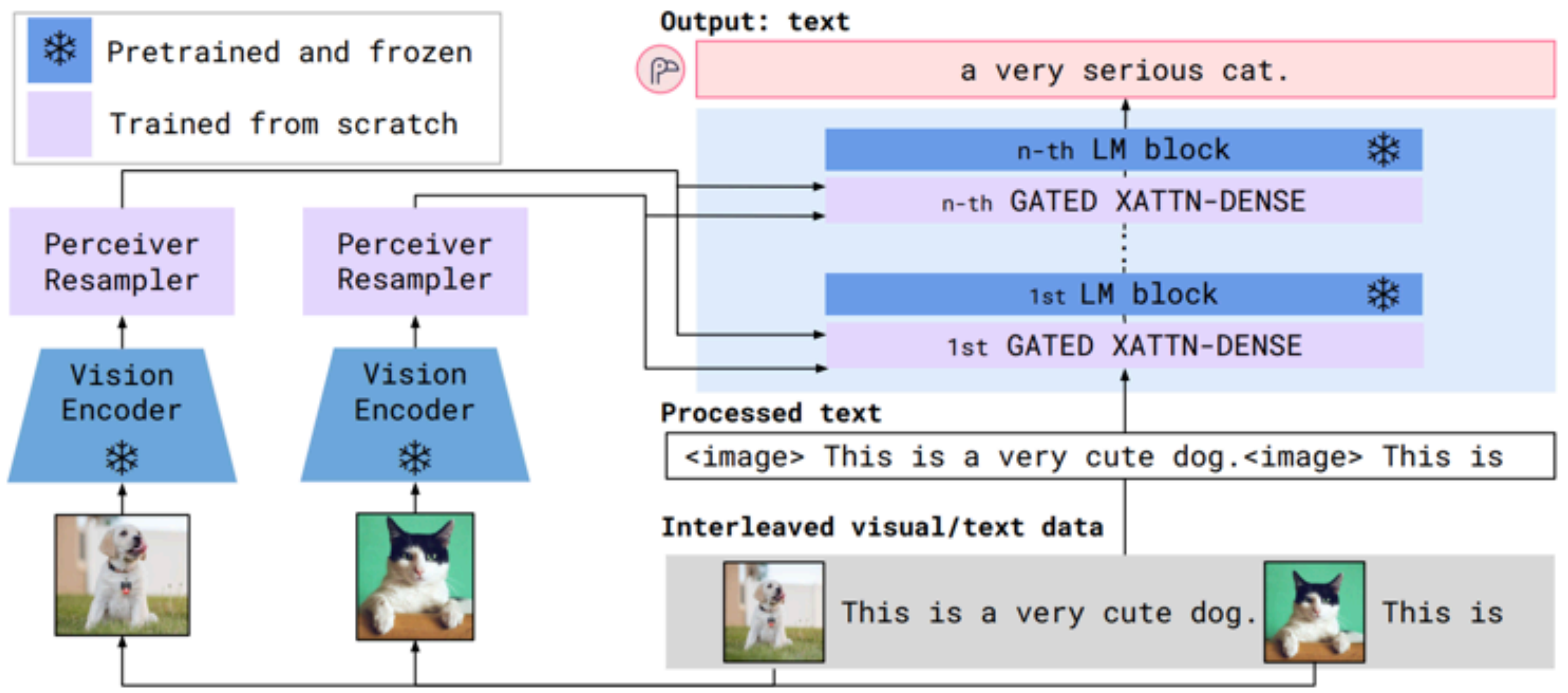
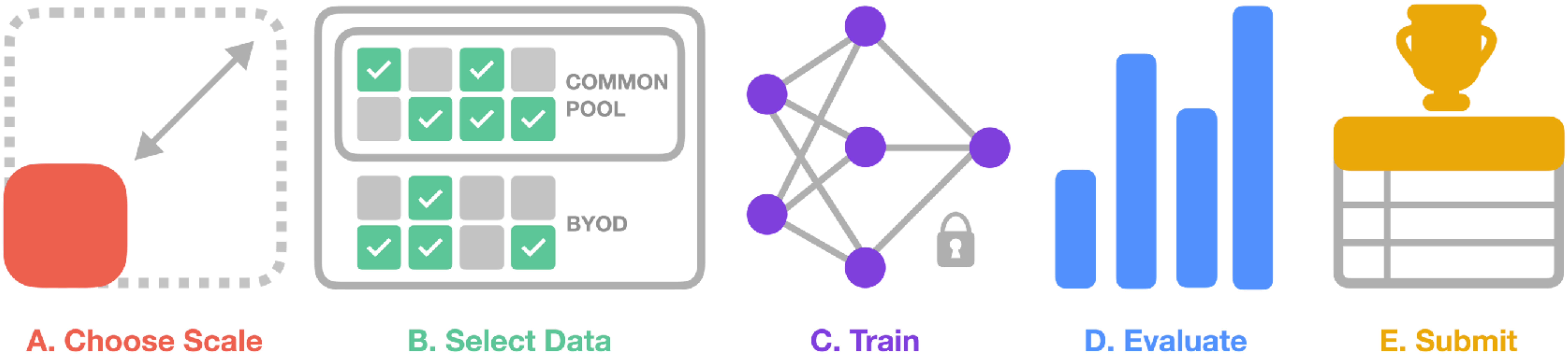


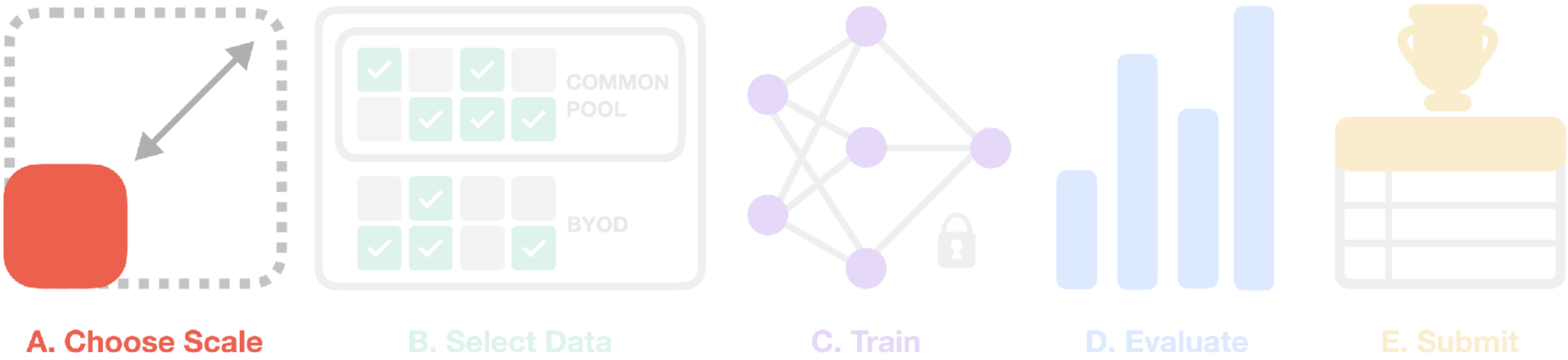
Figure 3: **Flamingo architecture overview.** Flamingo is a family of visual language models (VLMs) that take as input visual data interleaved with text and produce free-form text as output.

Alayrac et al. Flamingo: A visual language model for few-shot learning. NeurIPS 2022.

Re-enter DataComp

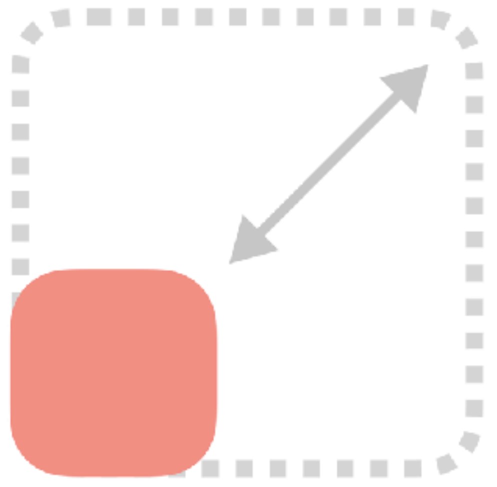


Choosing a scale



DataComp is compute accessible

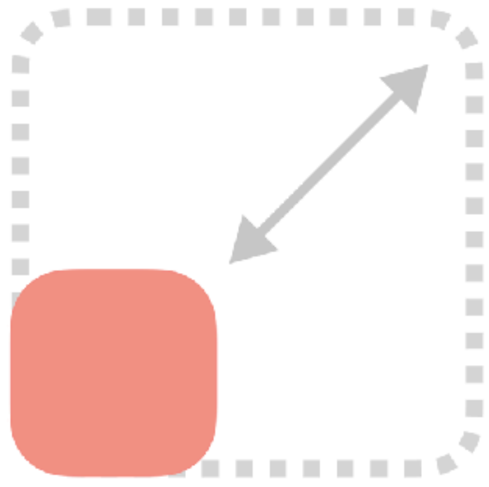
- Academics may have less resources and usually can’t train many FLOPs
- Industry labs may not want to participate unless DataComp can produce SOTA models
- Solution: different compute scales for participants



A. Choose Scale

Scale configurations

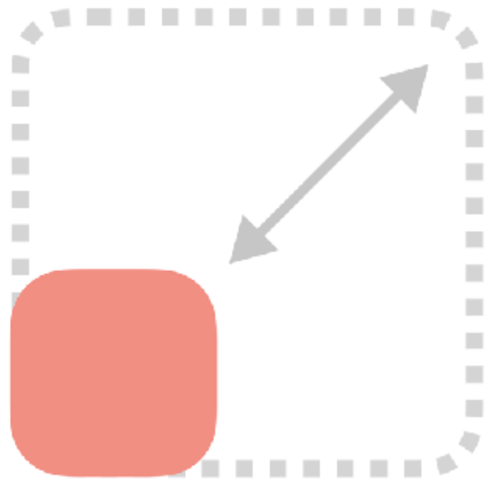
	small	medium	large	xlarge
samples seen				
model				
training A100 hours				
compute analogy				



A. Choose Scale

Scale configurations

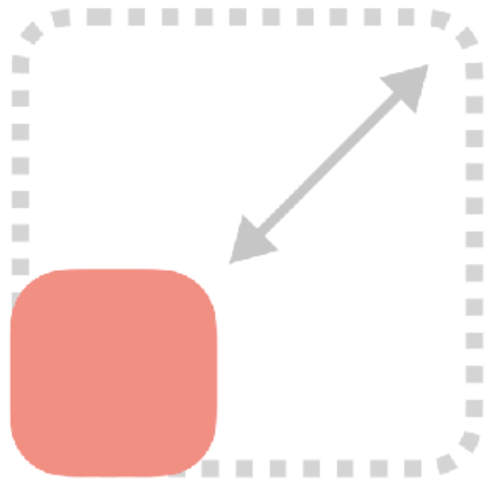
	small	medium	large	xlarge
samples seen	12.8M			
model	ViT-B/32			
training A100 hours	8			
compute analogy	fine-tune IN-1k			



A. Choose Scale

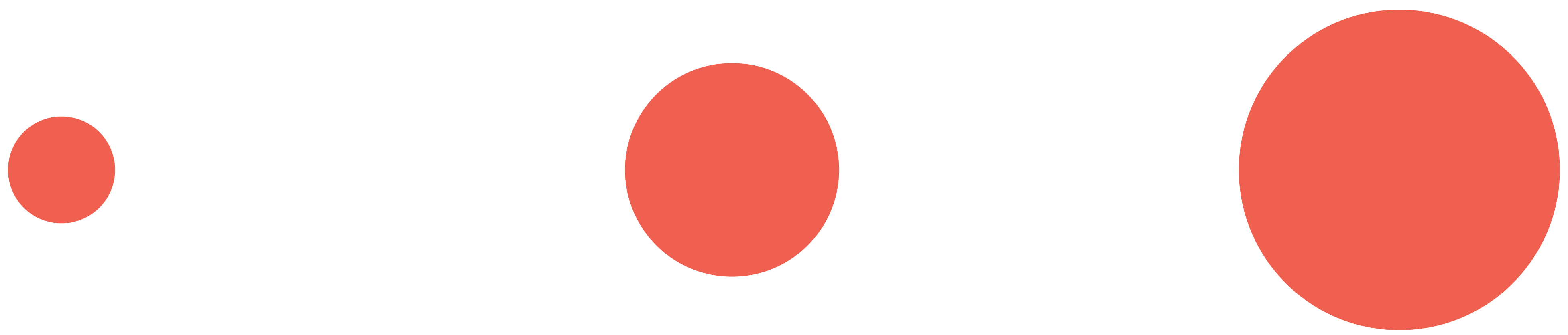
Scale configurations

	small	medium	large	xlarge
samples seen	12.8M	128M		
model	ViT-B/32	ViT-B/32		
training A100 hours	8	80		
compute analogy	fine-tune IN-1k	training IN-1k		

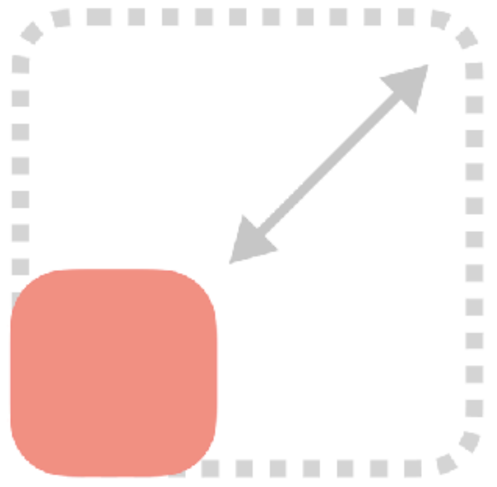


A. Choose Scale

Scale configurations

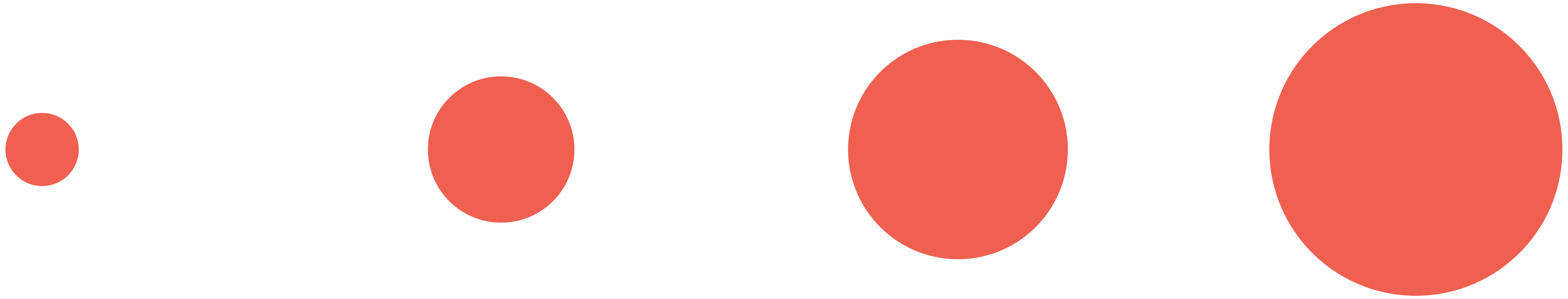


	small	medium	large	xlarge
samples seen	12.8M	128M	1.28B	
model	ViT-B/32	ViT-B/32	ViT-B/16	
training A100 hours	8	80	1,000	
compute analogy	fine-tune IN-1k	training IN-1k	training IN-21k	

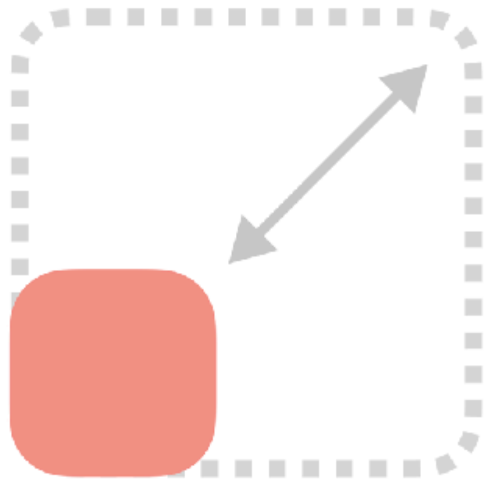


A. Choose Scale

Scale configurations

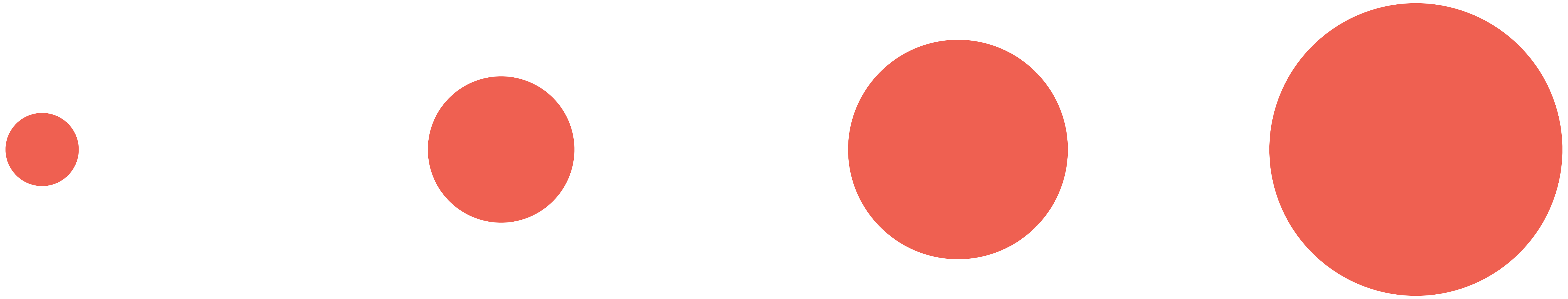


	small	medium	large	xlarge
samples seen	12.8M	128M	1.28B	12.8B
model	ViT-B/32	ViT-B/32	ViT-B/16	ViT-L/14
training A100 hours	8	80	1,000	40,000
compute analogy	fine-tune IN-1k	training IN-1k	training IN-21k	training OAI CLIP



A. Choose Scale

Scale configurations



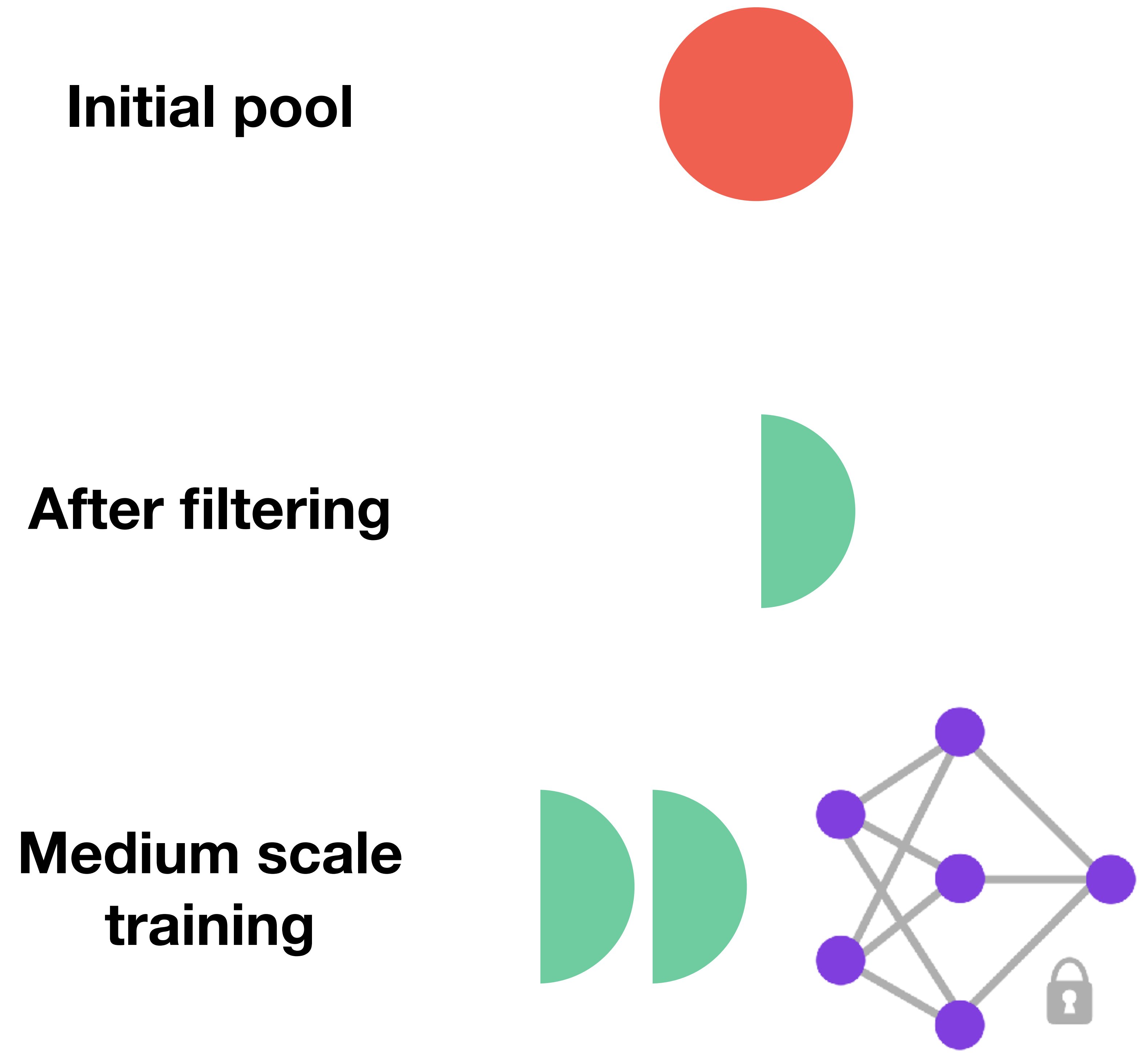
	small	medium	large	xlarge
samples seen	12.8M	128M	1.28B	12.8B
model	ViT-B/32	ViT-B/32	ViT-B/16	ViT-L/14
training A100 hours	8	80	1,000	40,000
compute analogy	fine-tune IN-1k	training IN-1k	training IN-21k	training OAI CLIP

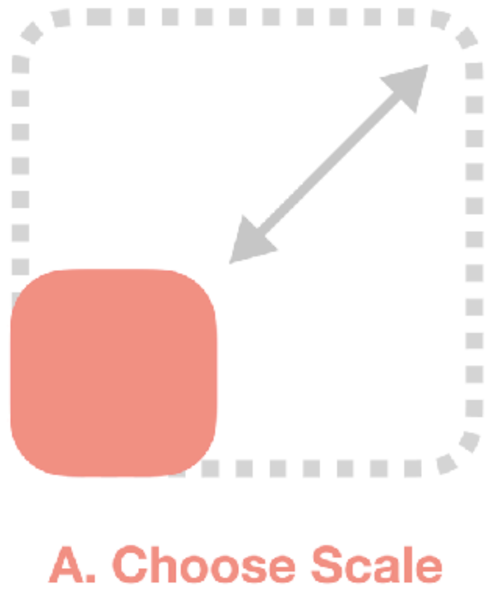
No constraint on dataset size! Real constraints are pool size and compute.

A. Choose Scale

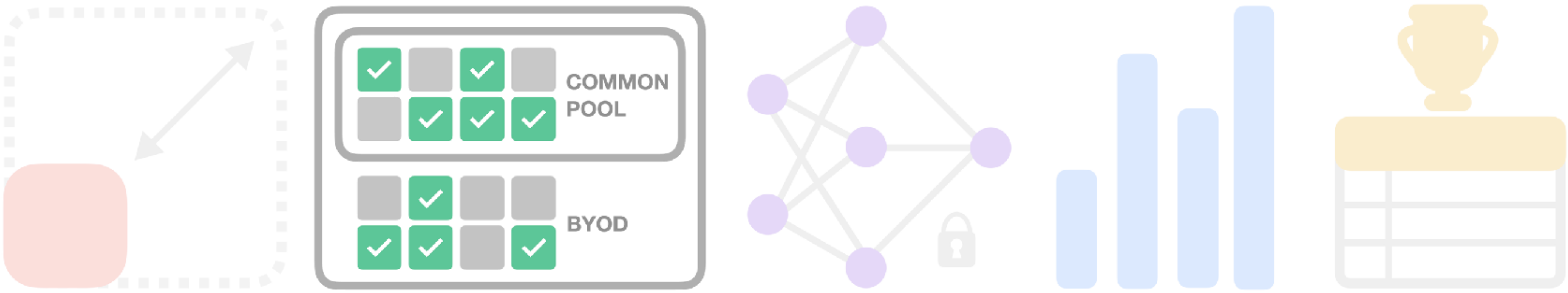
Example of samples seen

- Participating at the medium scale (128M samples seen)
- Filter a dataset to 64M samples
- Each sample will then get seen twice (in expectation) during training





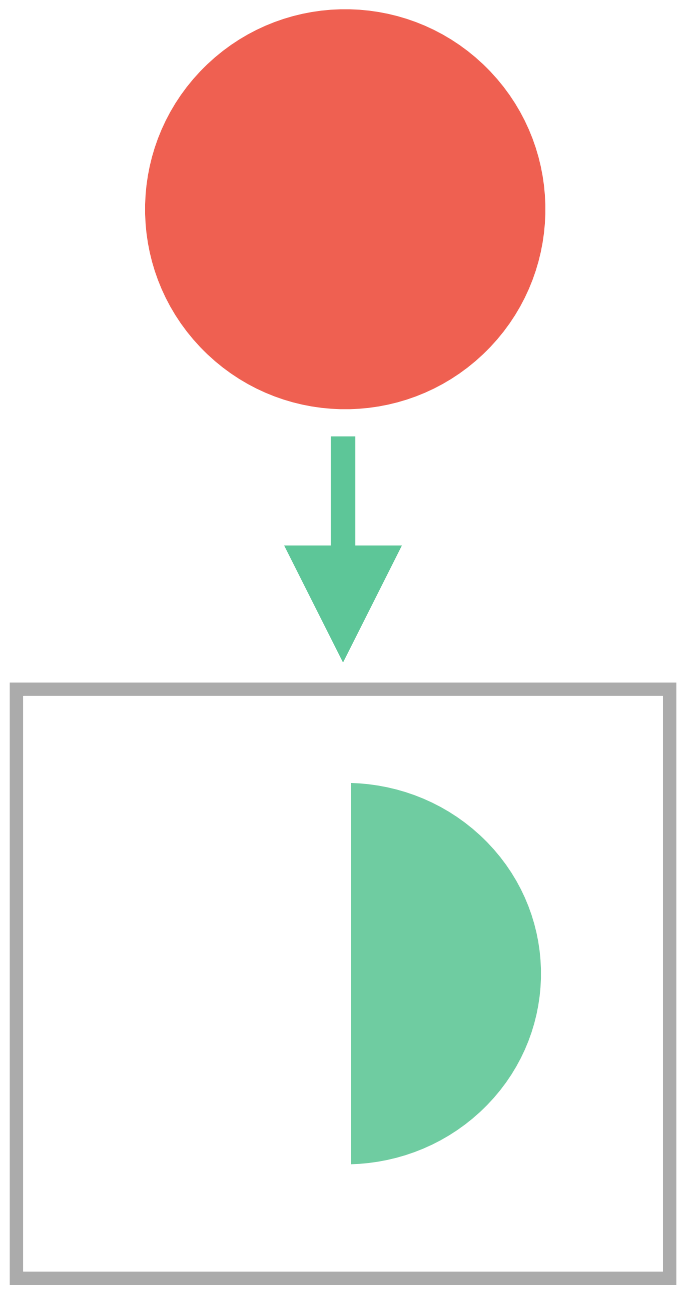
Selecting data



A. Choose Scale B. Select Data C. Train D. Evaluate E. Submit

Two tracks: Filtering and BYOD

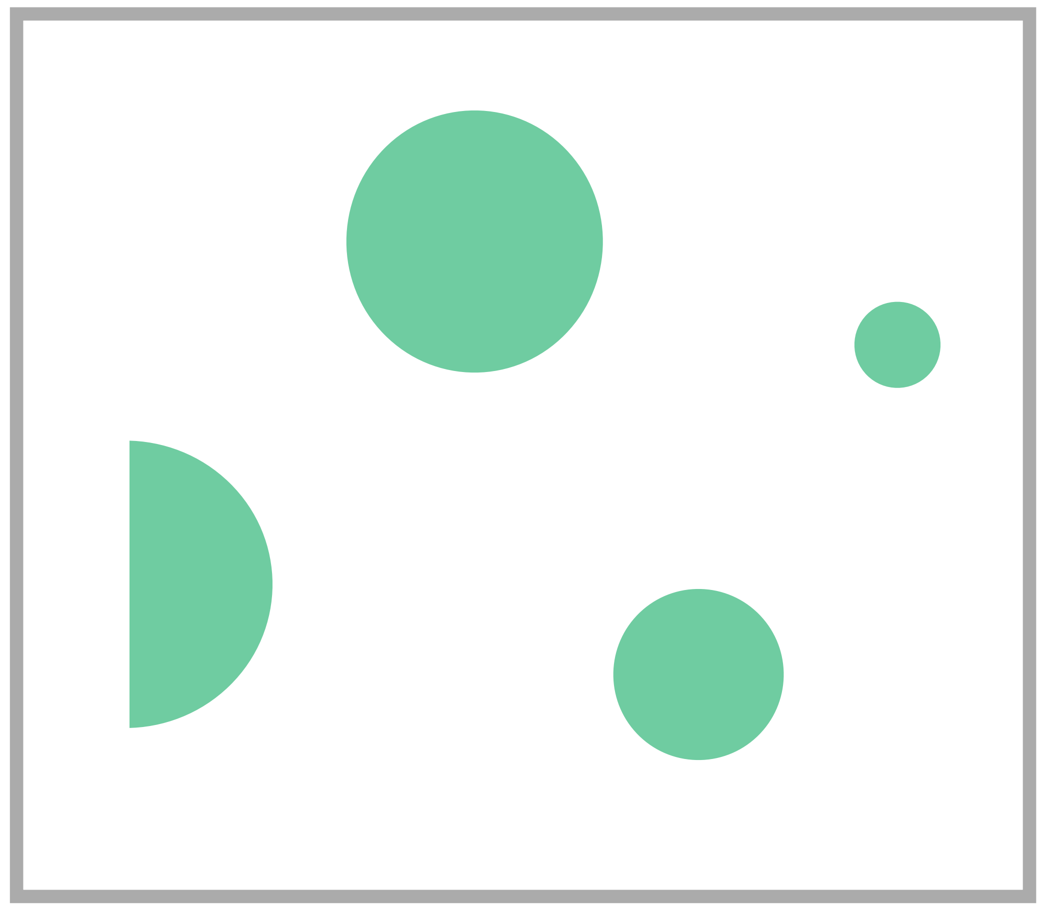
Filtering

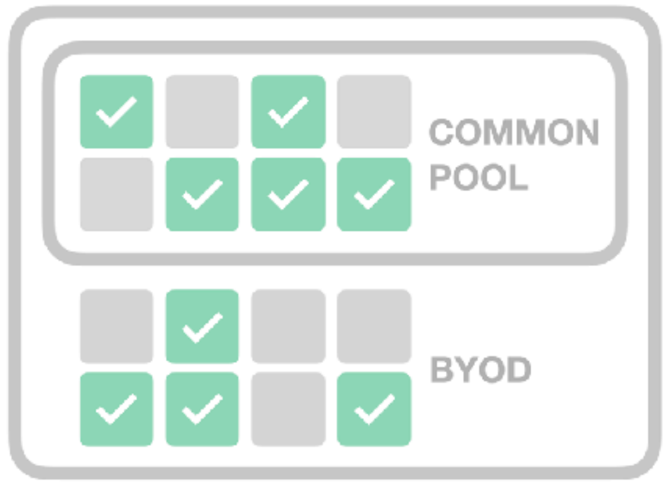


Pare down a noisy source

Bring your own data (BYOD)

Cobble together many sources

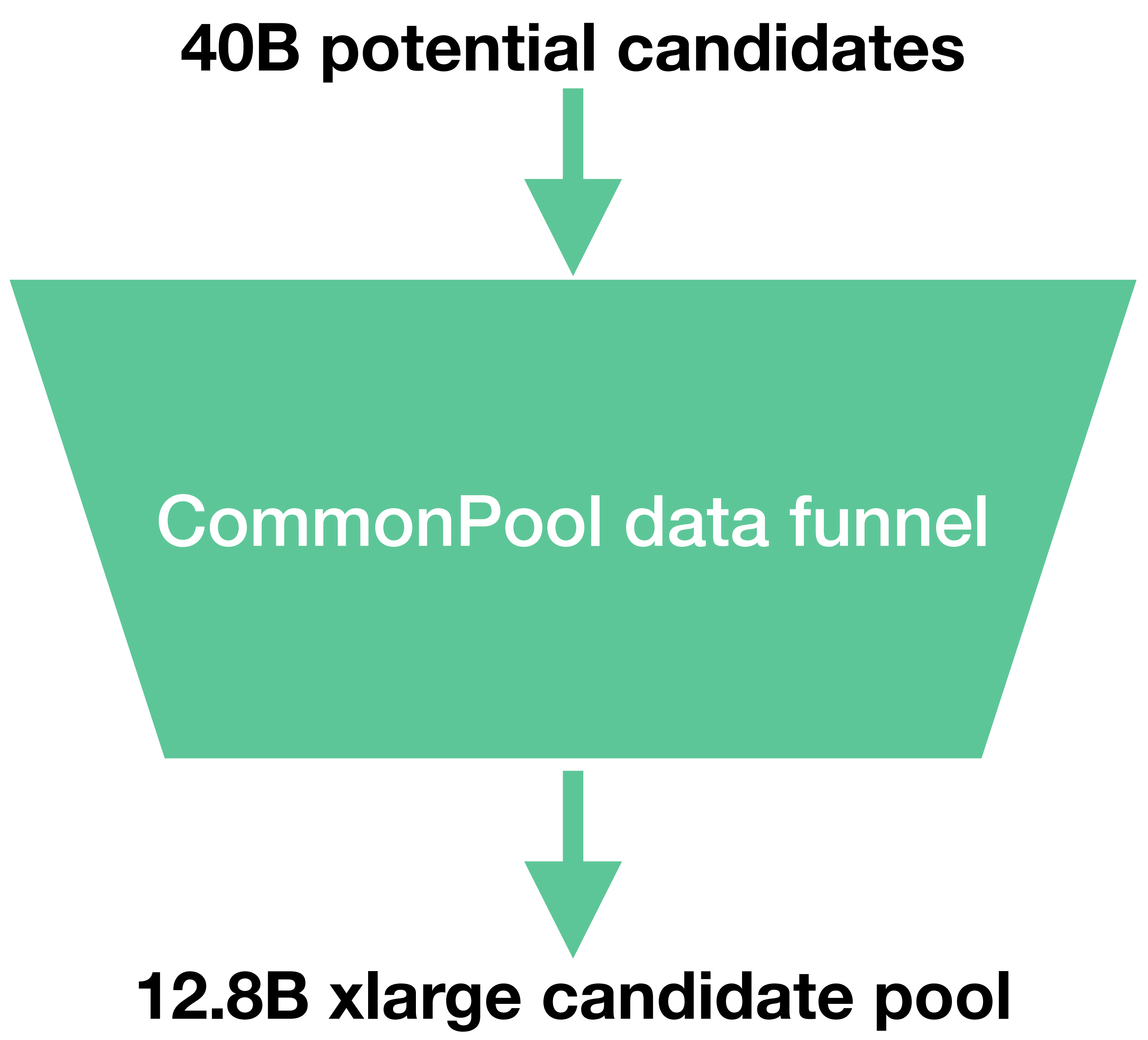


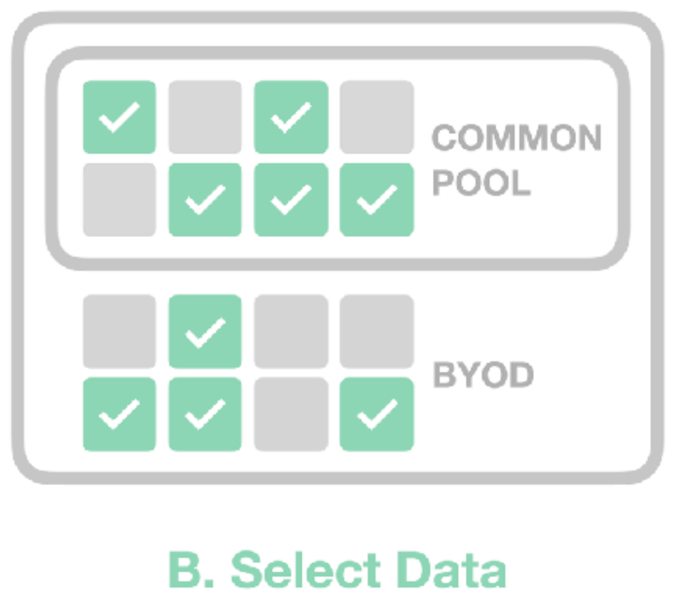


B. Select Data

CommonPool to facilitate the filtering track

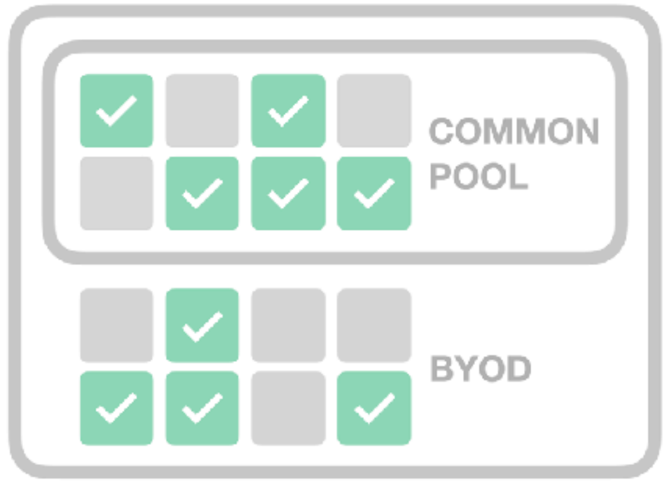
- 88B url-(alt)text pairs from CommonCrawl
- 40B attempted image downloads
- 16.8B successfully downloaded
- 13.1B retained after pre-processing
- 12.8B sampled for the xlarge pool





Pre-filtering for safety and eval decontamination

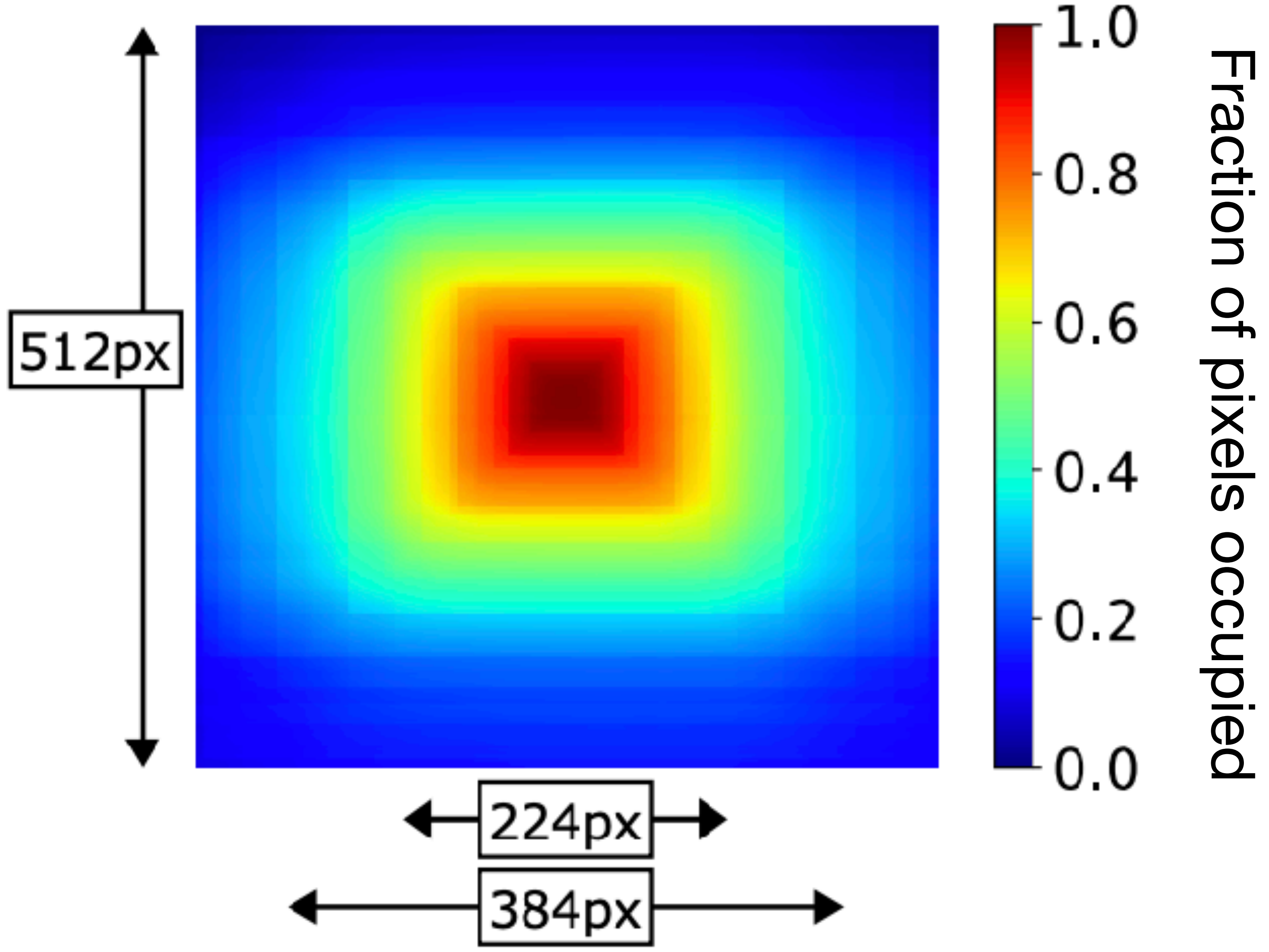
- Near deduplication against downstream evaluation images
- NSFW image removal
- NSFW text removal
- Face blurring automatically in download tooling
- Notably, not pre-processing for “quality”
- Dataset safety is an active area of research!

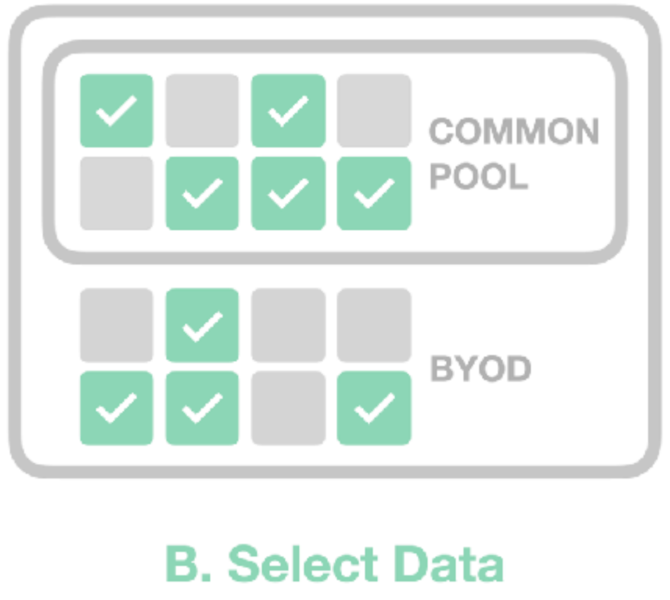


B. Select Data

Metadata

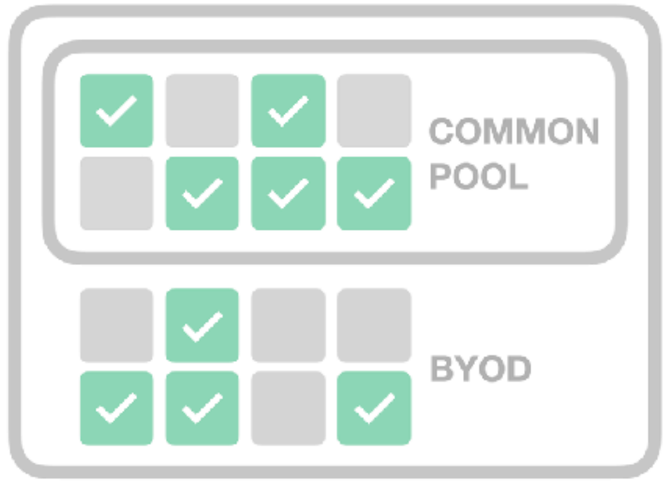
- Original width/height
- Caption
- Image sha256
- CLIP features (B/32 and L/14)
- CLIP scores
- Face bounding boxes (for automatic blurring)





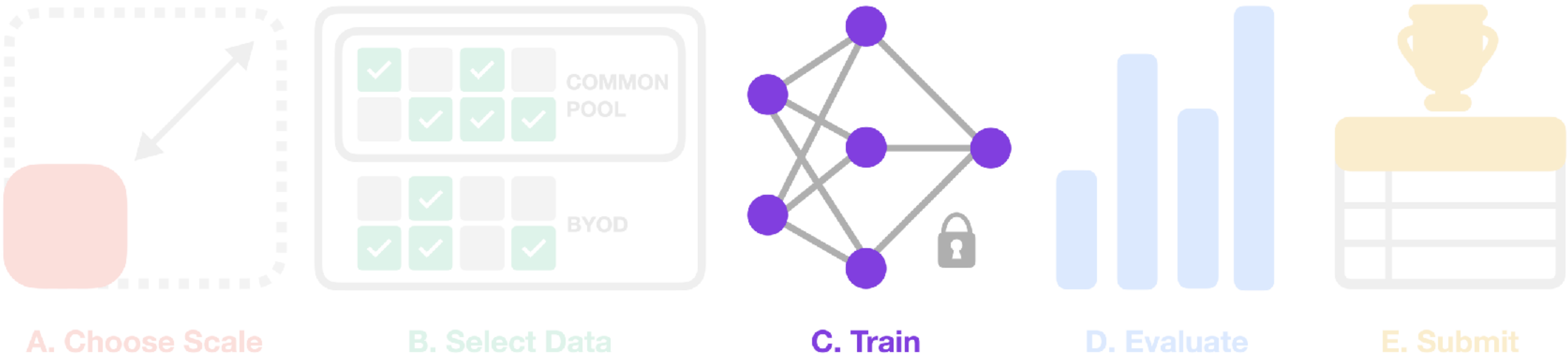
Bring your own data (BYOD)

- Filtering is only one way to curate datasets
- Combine other data sources (e.g., YFCC-15M, CC12M, RedCaps, etc.)
- CommonPool filtering ++
- The BYOD track allows this flexibility



B. Select Data

Train



COMMON POOL
BYOD
A. Choose Scale
B. Select Data
C. Train
D. Evaluate
E. Submit

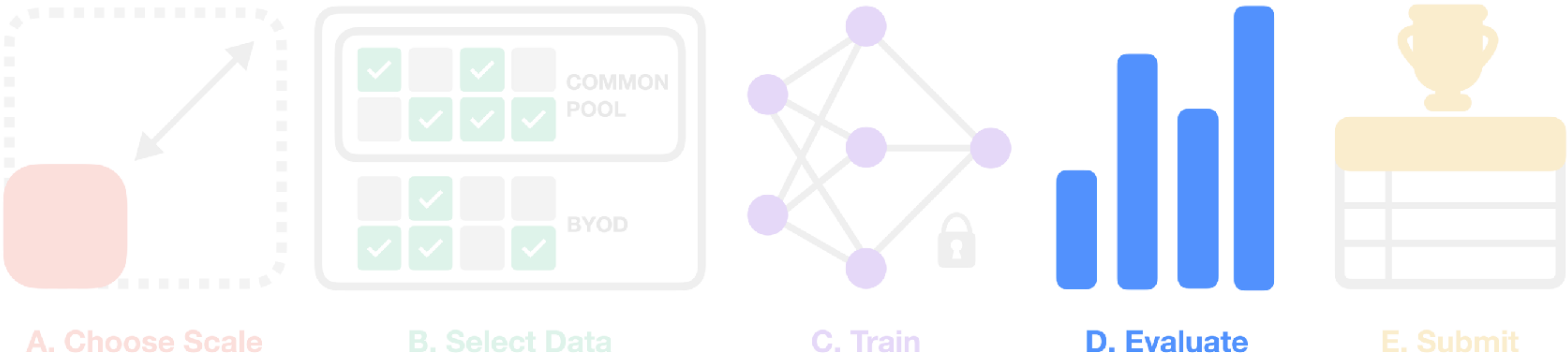
Fixed training configurations

- Hyperparameters based on OpenAI, LAION (open_clip) runs
- Fixed, so participants cannot modify
- Ablations on architecture, batch size, etc. show relatively consistent trends, suggesting dataset and modeling choices can be considered independently

```
"medium": {
    "batch_size": 4096,
    "learning_rate": 5e-4,
    "train_num_samples": 128_000_000,
    "warmup": 500,
    "model": "ViT-B-32",
    "beta2": None,
},
```

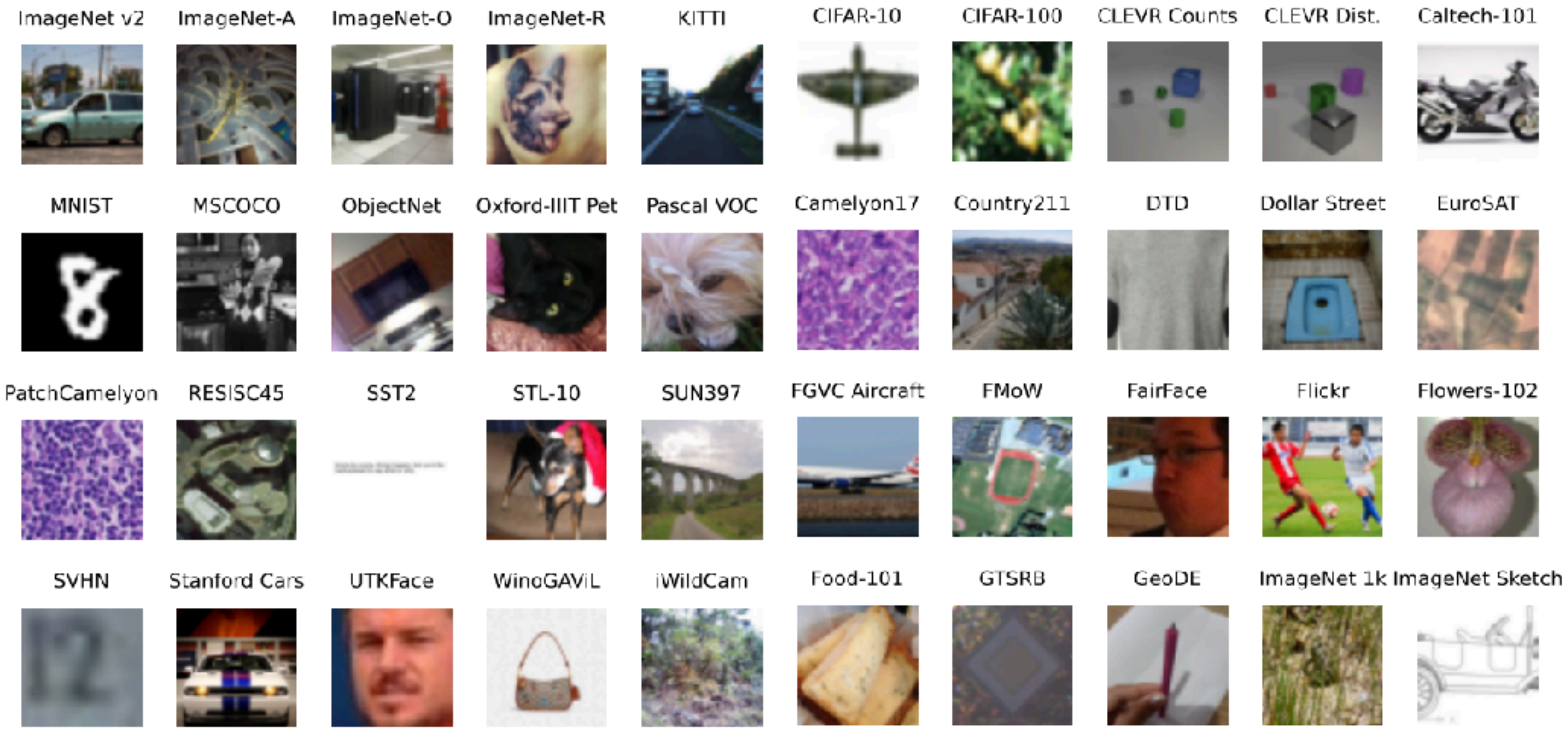
C. Train

Evaluate


COMMON
POOL
BYOD
A. Choose Scale
B. Select Data
C. Train
D. Evaluate
E. Submit

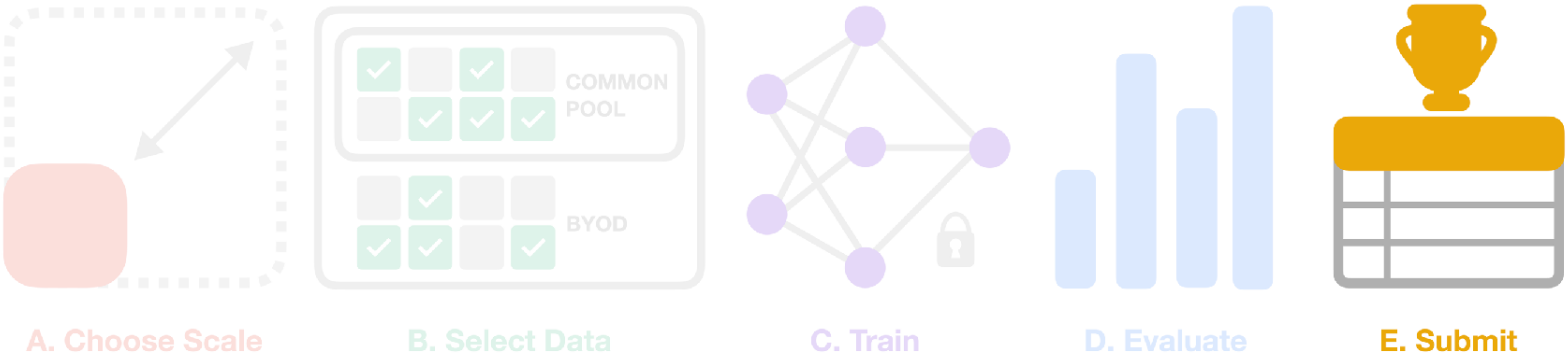
Downstream eval sets

- 38 core classification and retrieval tasks
- Evaluations are zero-shot (no fine-tuning)
- We look at both ImageNet and average acc.



D. Evaluate

Submit


COMMON
POOL
BYOD
A. Choose Scale
B. Select Data
C. Train
D. Evaluate
E. Submit

datacomp.ai

DataComp

Home Participate Tracks FAQs Workshop Team

Leaderboard

DataComp

Welcome to DataComp, the machine learning benchmark where the models are fixed and the challenge is to find the best possible data!

E. Submit

A unified leaderboard

Select the track and scale

Filtering track		BYOD track	
small	medium	large	xlarge

Leaderboard

Rank	Created	Submission	ImageNet acc.	Average perf.
1	10-02-2023	Data Filtering Networks	0.371	0.373
2	09-08-2023	The Devil Is in the Details	0.320	0.371
3	08-17-2023	T-MARS: Improving Visual Representations by Circumventing Text Feature Learning	0.330	0.361

E. Submit

DataComp leads to better models

Teaser results

Dataset	Dataset size	Samples seen	Architecture	ImageNet-1k accuracy
OpenAI WIT	0.4B	13B	ViT-L/14	75.5

Teaser results

Dataset	Dataset size	Samples seen	Architecture	ImageNet-1k accuracy
OpenAI WIT	0.4B	13B	ViT-L/14	75.5
LAION-400M	0.4B	13B	ViT-L/14	72.8

Teaser results

Dataset	Dataset size	Samples seen	Architecture	ImageNet-1k accuracy
OpenAI WIT	0.4B	13B	ViT-L/14	75.5
LAION-400M	0.4B	13B	ViT-L/14	72.8
LAION-2B	2.3B	13B	ViT-L/14	73.1
LAION-2B	2.3B	34B	ViT-H/14	78.0
LAION-2B	2.3B	34B	ViT-g/14	78.5

Teaser results

Dataset	Dataset size	Samples seen	Architecture	ImageNet-1k accuracy
OpenAI WIT	0.4B	13B	ViT-L/14	75.5
LAION-400M	0.4B	13B	ViT-L/14	72.8
LAION-2B	2.3B	13B	ViT-L/14	73.1
LAION-2B	2.3B	34B	ViT-H/14	78.0
LAION-2B	2.3B	34B	ViT-g/14	78.5
DataComp-1B	1.4B	13B	ViT-L/14	**79.2**

Teaser results

Dataset	Dataset size	Samples seen	Architecture	ImageNet-1k accuracy
OpenAI WIT	0.4B	13B	ViT-L/14	75.5
LAION-400M	0.4B	13B	ViT-L/14	72.8
LAION-2B	2.3B	13B	ViT-L/14	73.1
LAION-2B	2.3B	34B	ViT-H/14	78.0
LAION-2B	2.3B	34B	ViT-g/14	78.5
DataComp-1B	1.4B	13B	ViT-L/14	**79.2**

+3.7pp

Teaser results

Dataset	Dataset size	Samples seen	Architecture	ImageNet-1k accuracy
OpenAI WIT	0.4B	13B	ViT-L/14	75.5
LAION-400M	0.4B	13B	ViT-L/14	72.8
LAION-2B	2.3B	13B	ViT-L/14	73.1
LAION-2B	2.3B	34B	ViT-H/14	78.0
LAION-2B	2.3B	34B	ViT-g/14	78.5
DataComp-1B	1.4B	13B	ViT-L/14	**79.2**

9x compute savings

How did we get there? Baselines!

- No filtering
- CLIP-score filtering
- Basic: filtering based on aspect ratio, caption length, etc.
- Image-based filtering: clustering against ImageNet-1k train
- Text-based filtering: looking for ImageNet-1k synsets

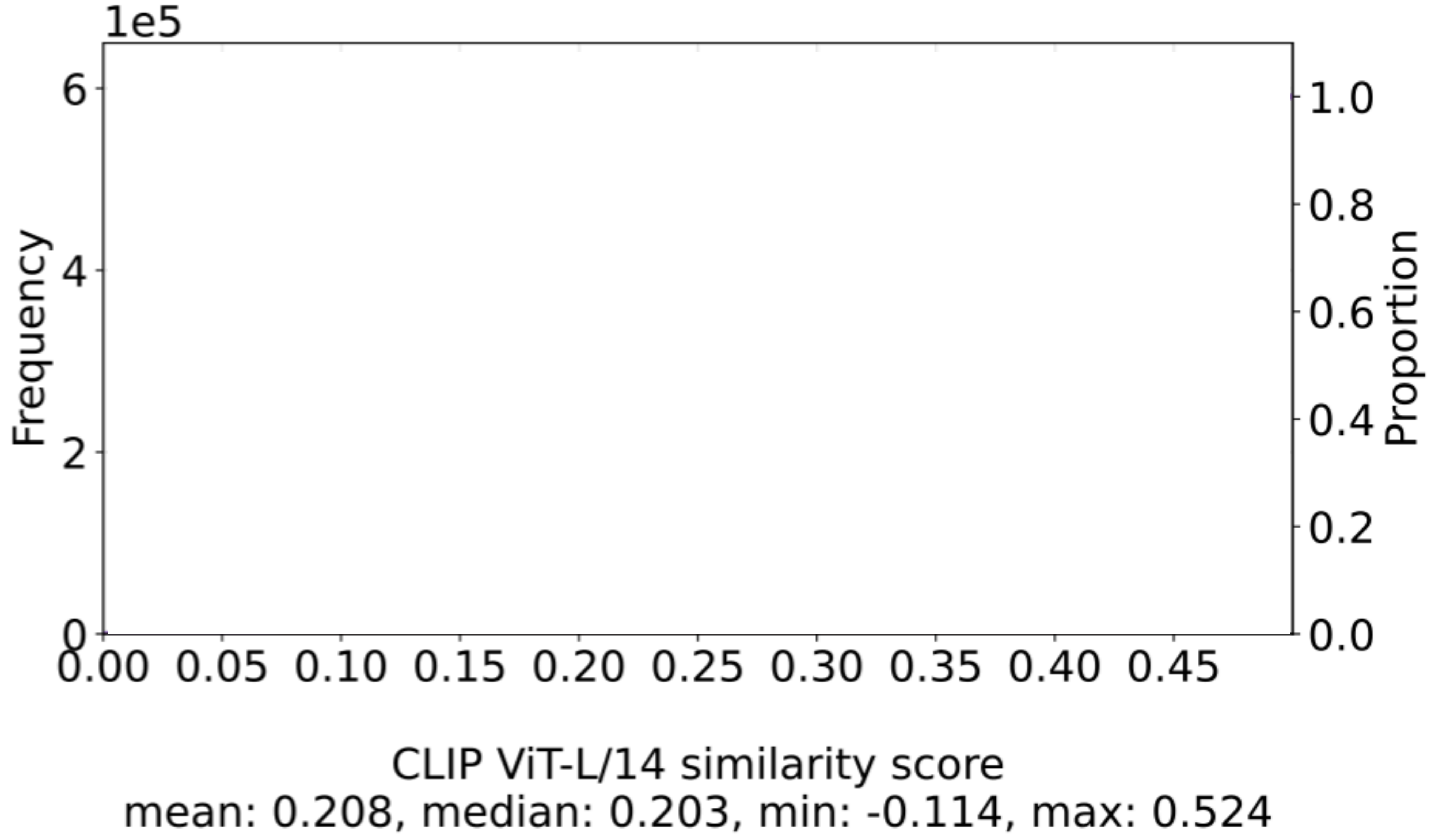
IMG_2187.jpg

No filtering

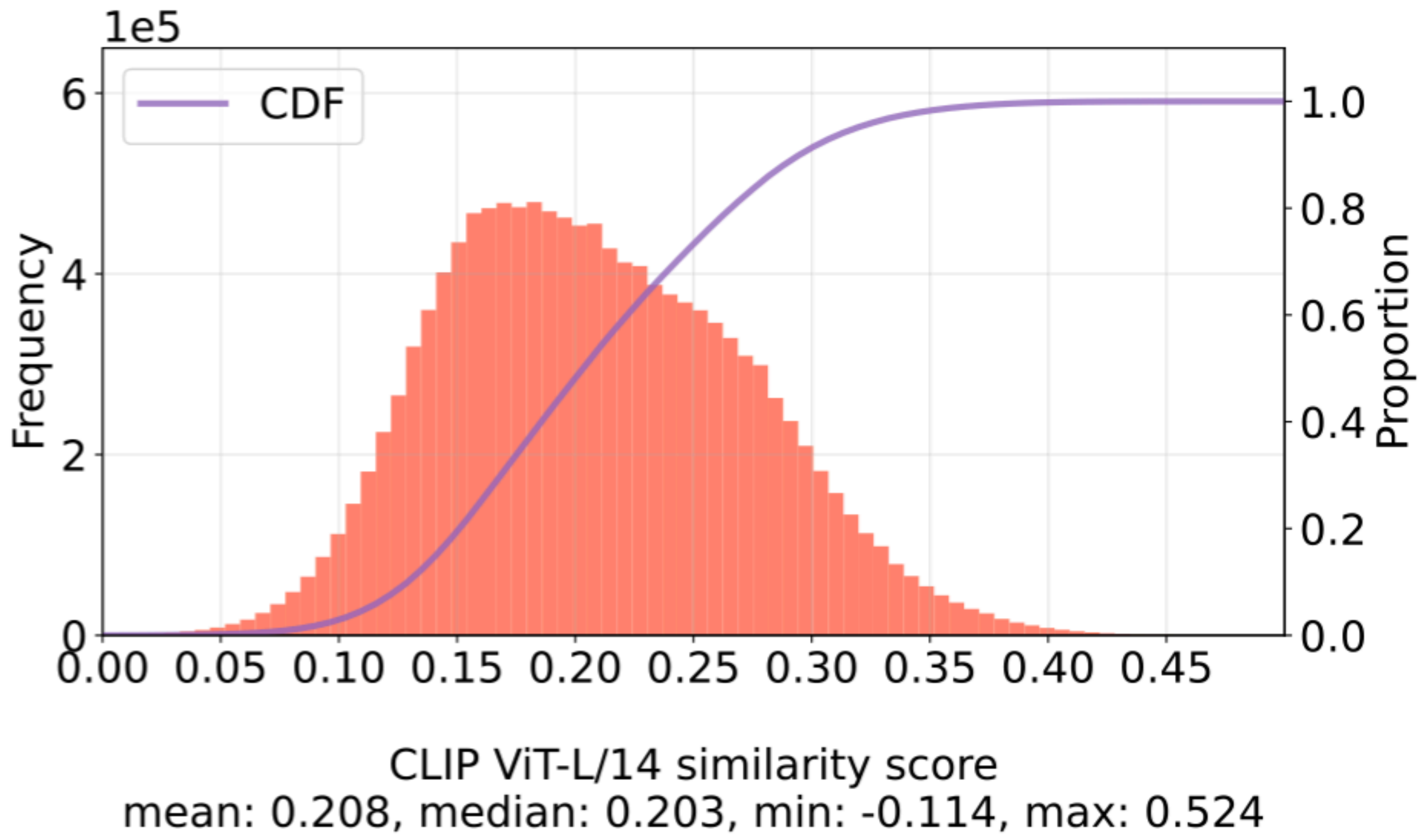
Porsche Cayman S

CLIP filtering (pool top 30%)

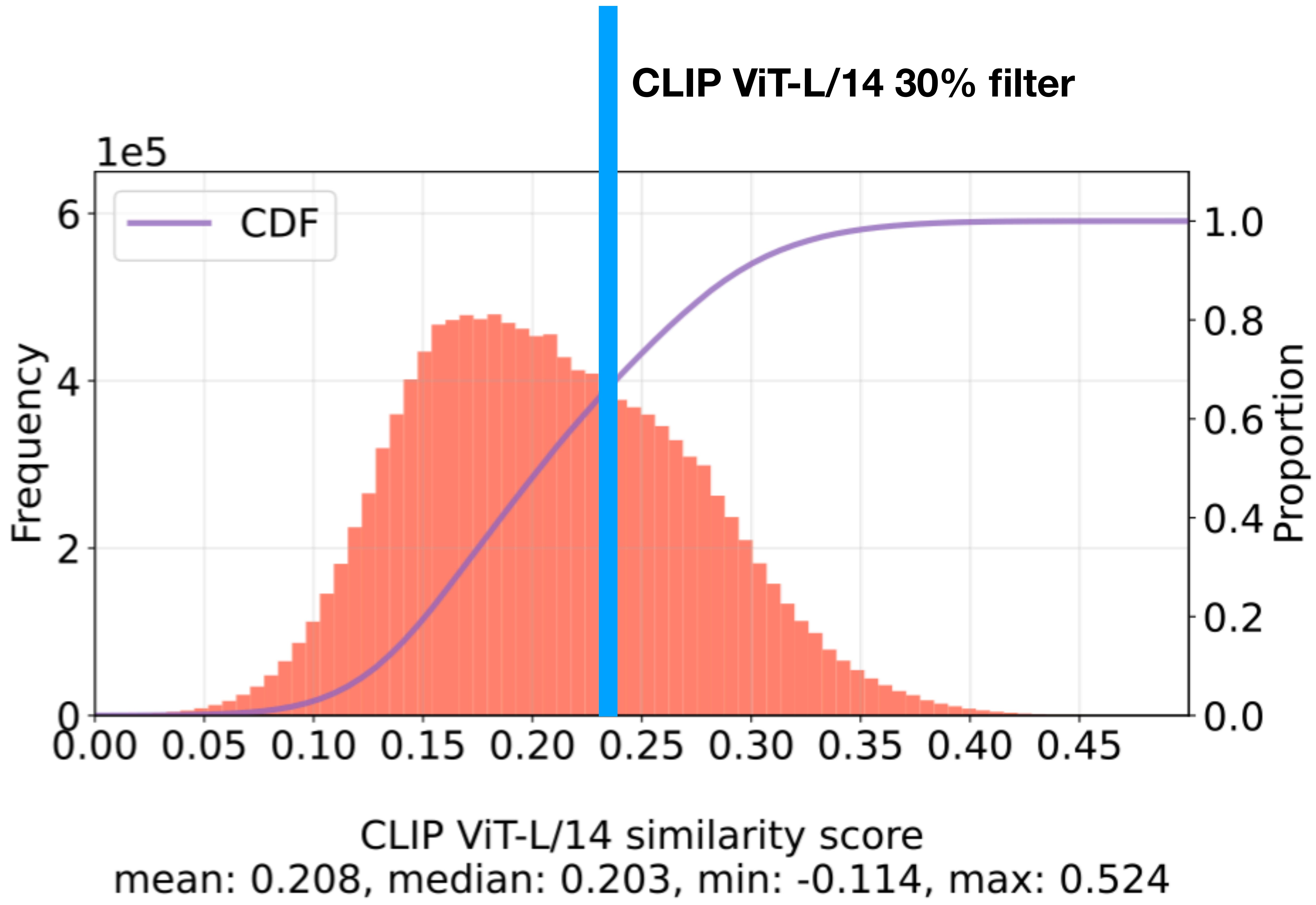
How did we get there? Baselines!



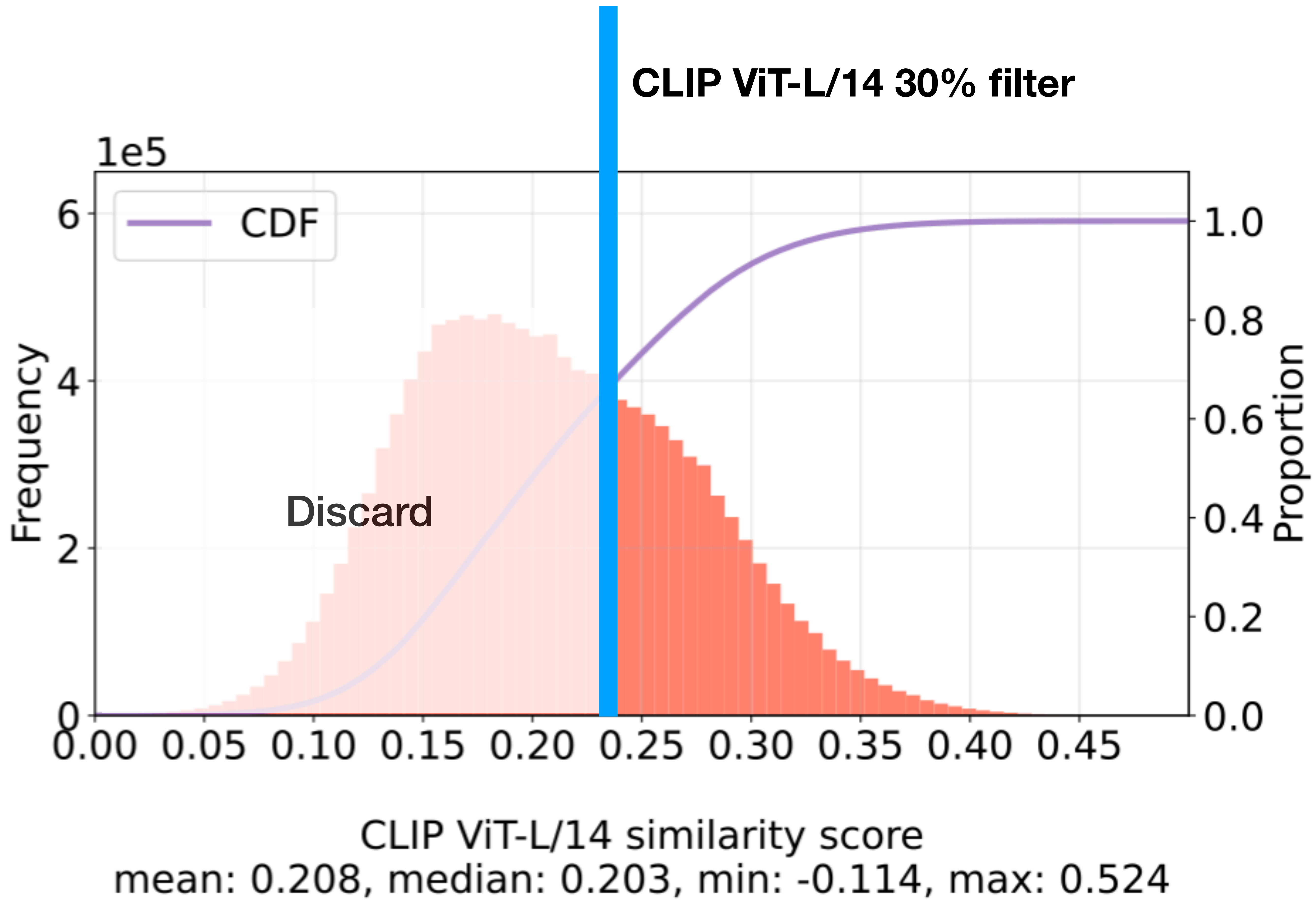
How did we get there? Baselines!



How did we get there? Baselines!

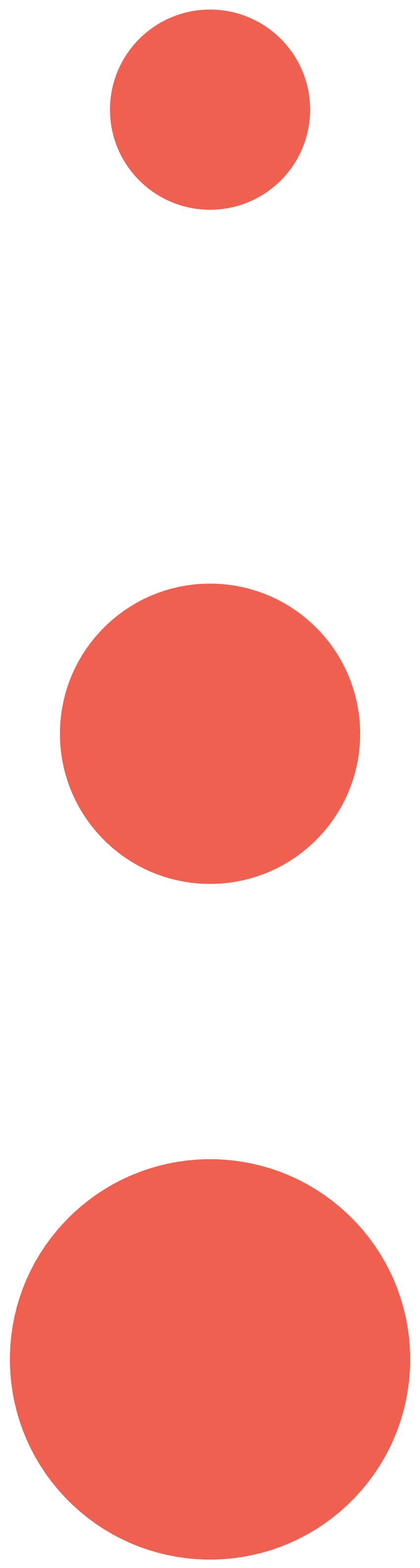


How did we get there? Baselines!



How did we get there? Workflow!

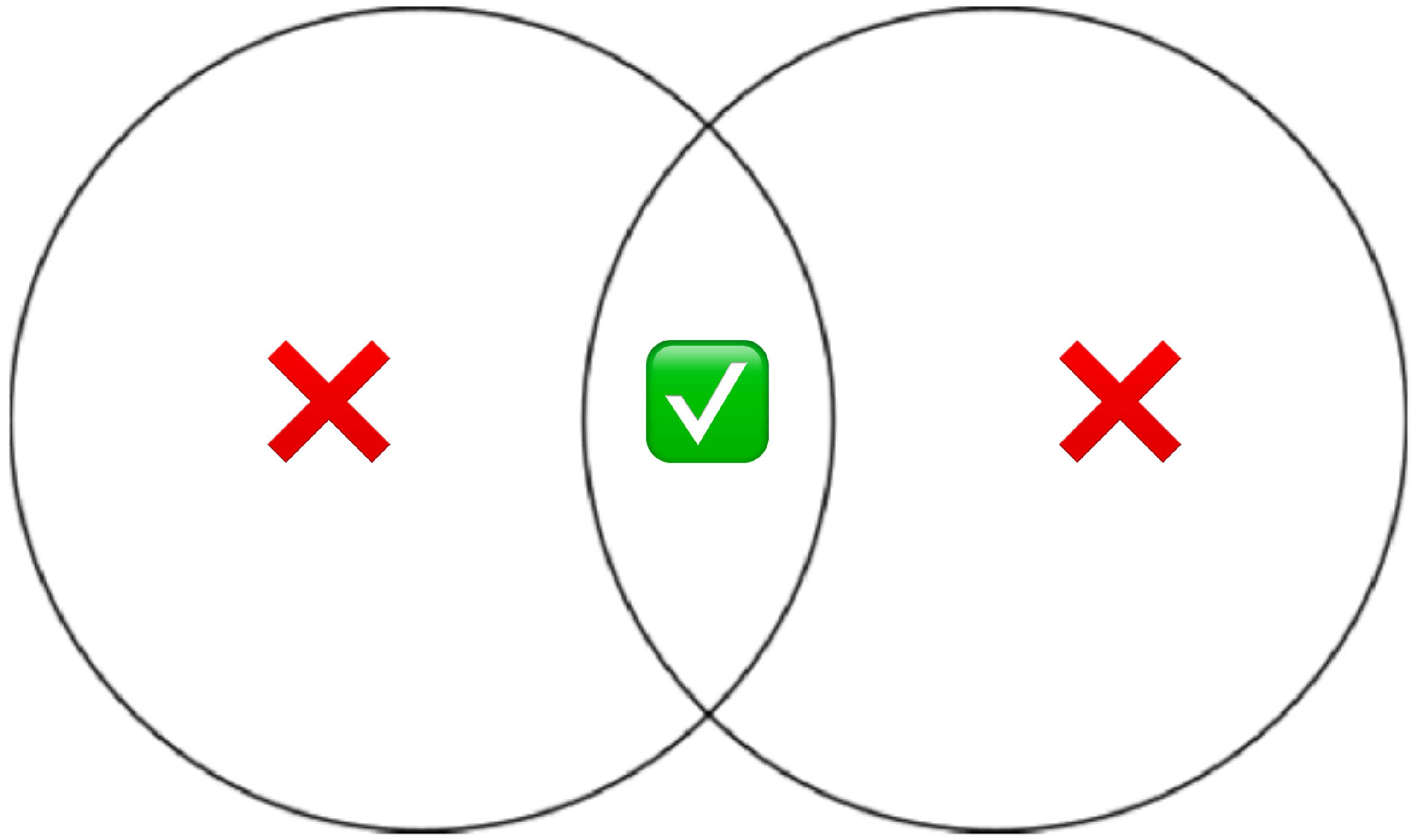
- Ran many experiments at small and medium scale
- Best methods we run at large scale
- Best at large (often same as at medium), we run at xlarge



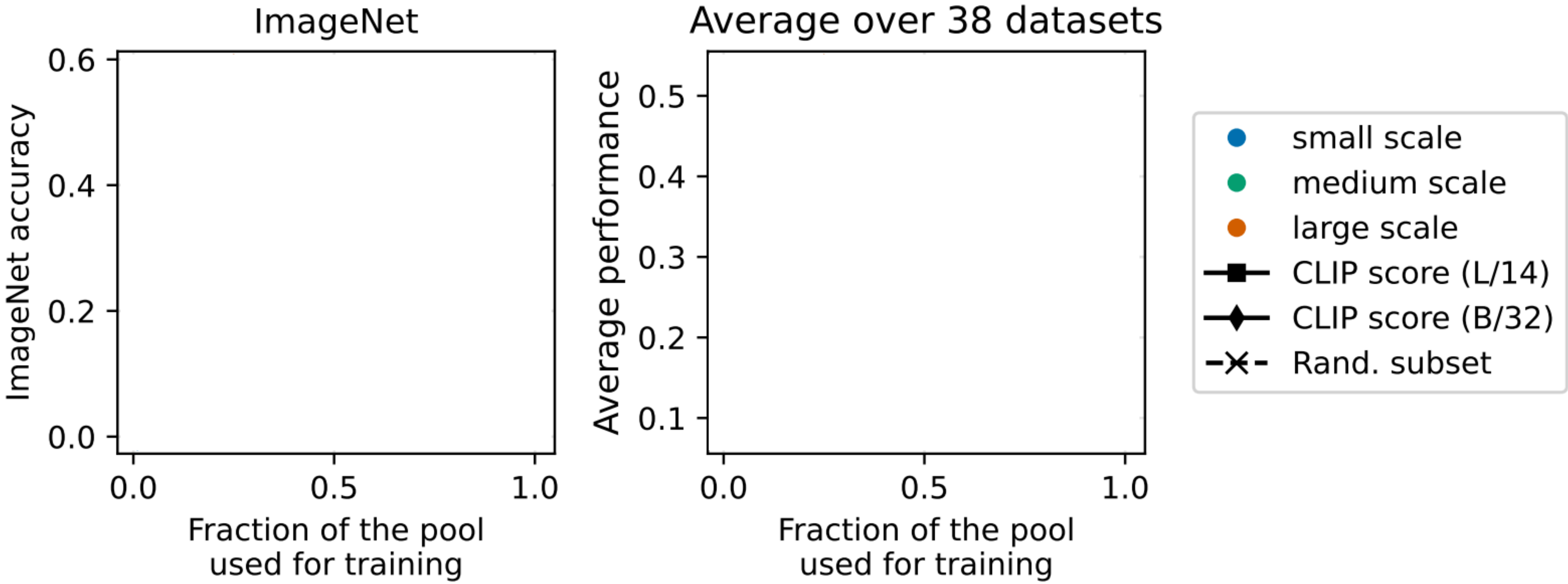
DataComp-1B

- Combination of 2 baseline strategies (CLIP-filter ∩ image-based)

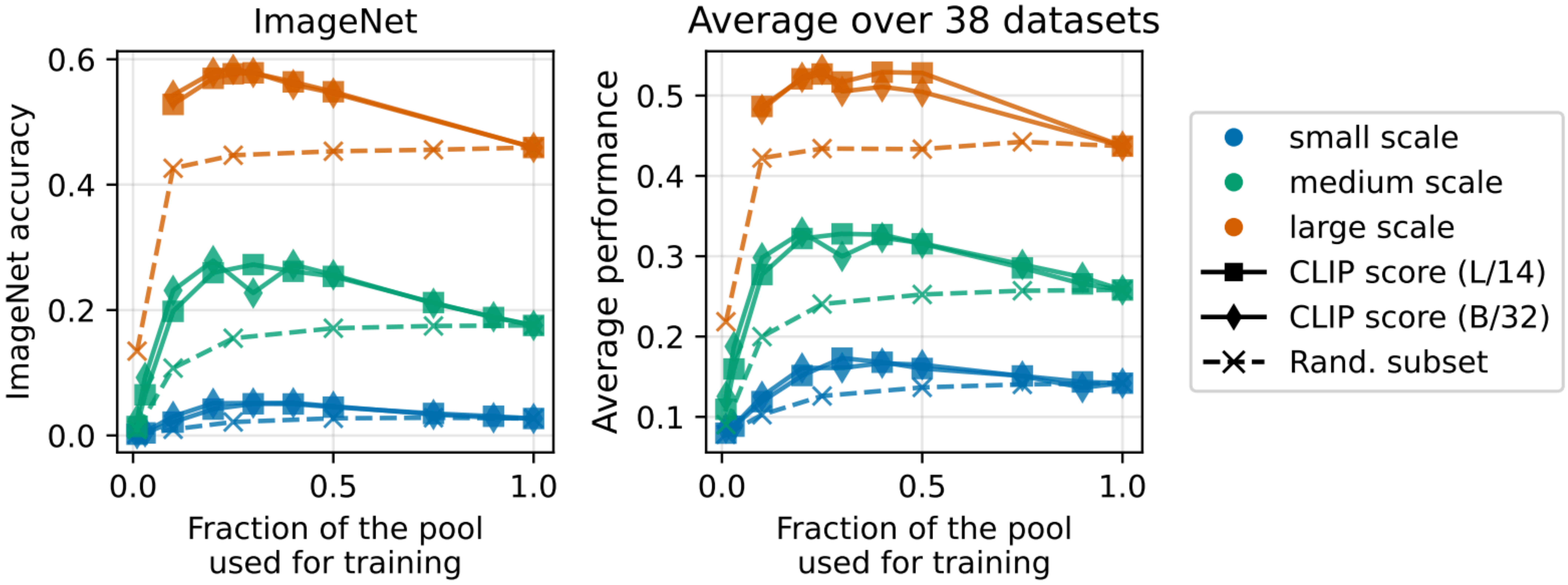
- First (public) training set that's better than OpenAI's.



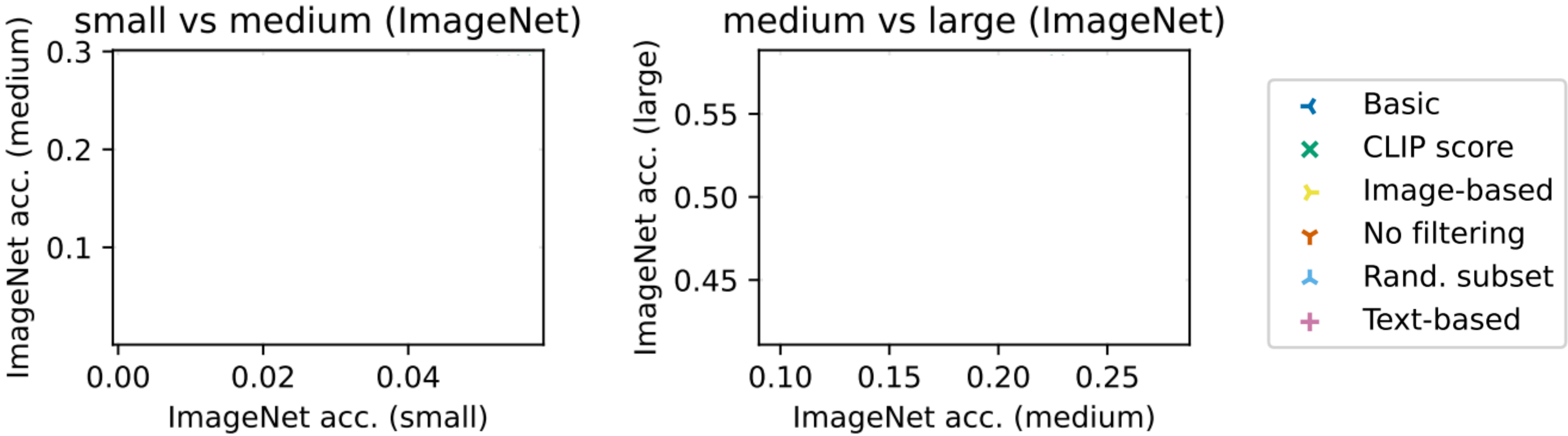
Dataset size is not the full story



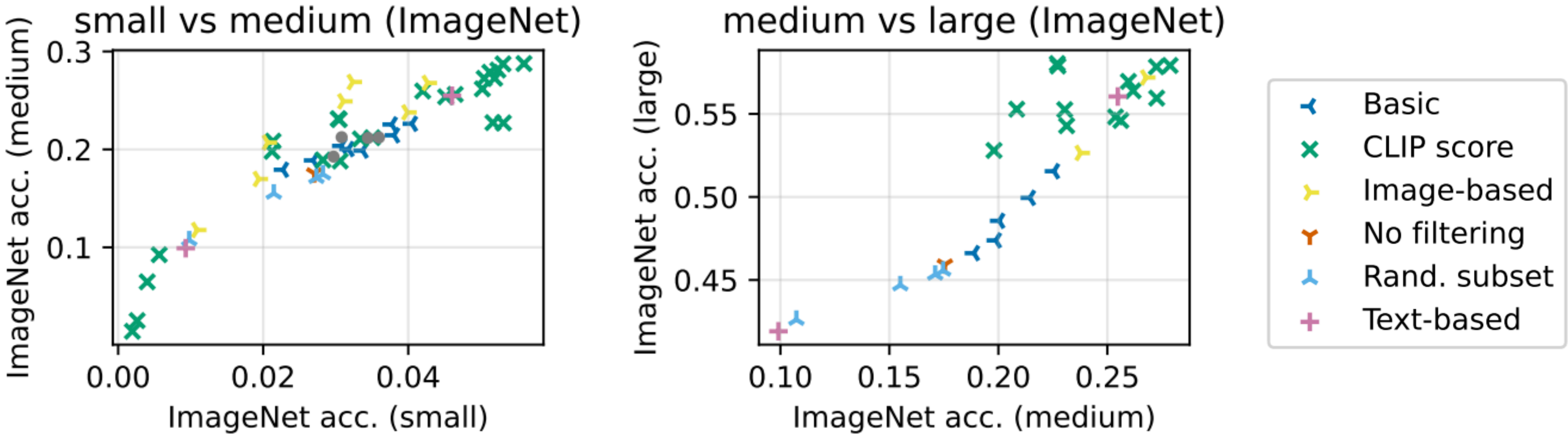
Dataset size is not the full story



Consistent ordering across scales



Consistent ordering across scales



BYOD mixing can help

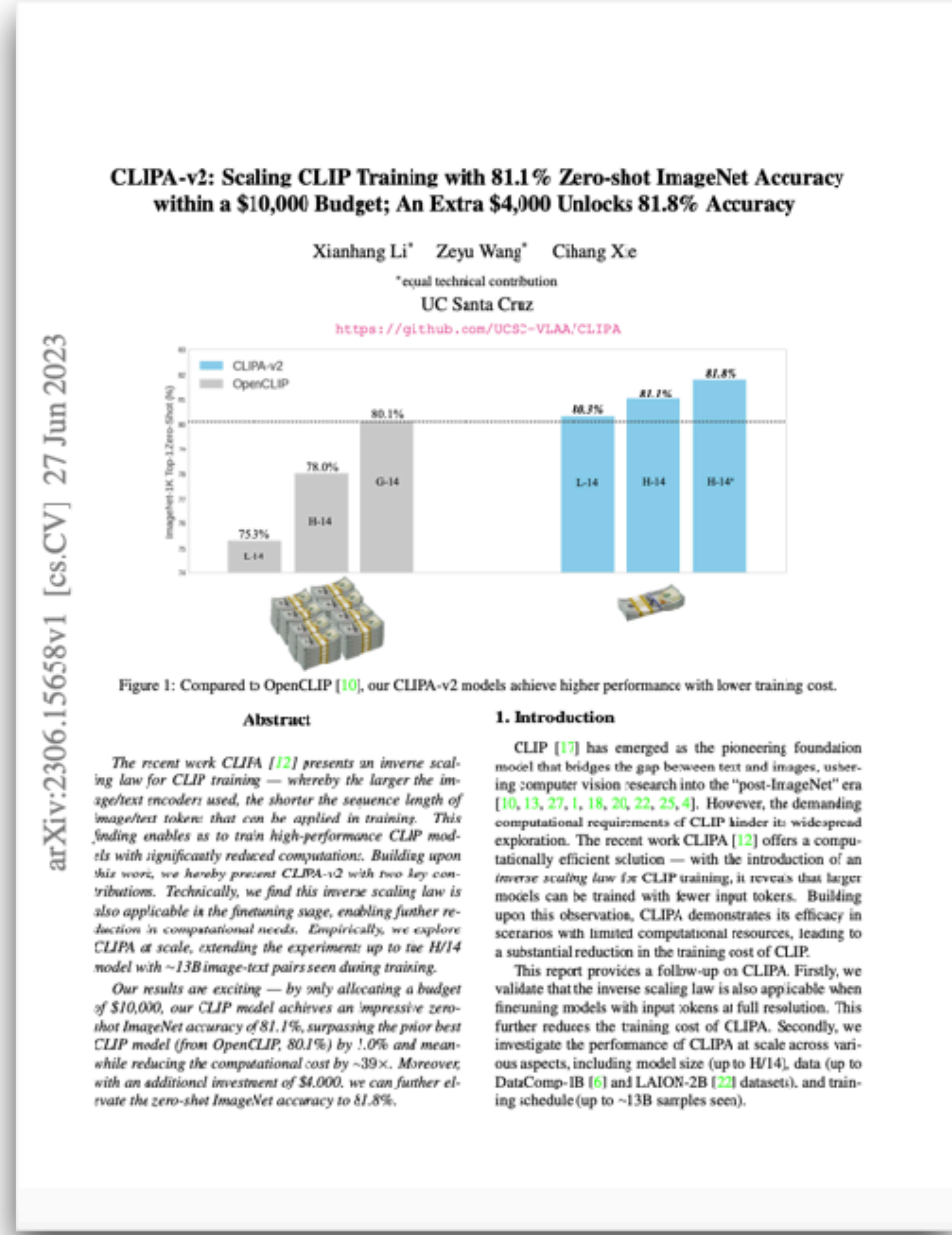
- BYOD seems to help up to large scale
- Currently diminishing returns at xlarge
- Lots left to explore here (rich, scalable sources beyond CommonCrawl?)

	large	xlarge
CLIP-filter ∩ image-based	63.1	79.2
+ 4 BYOD sources	65.6	79.2

Future directions

- Curating more dataset sources
- Improved filtering and dataset balancing methods
- Further (weak) supervision signals
- Additional modalities
- Broader evaluation on vision, vision-language, robotics tasks, and model bias

People are already iterating on DataComp



CLIPA-v2: Scaling CLIP Training with 81.1% Zero-shot ImageNet Accuracy within a $10,000 Budget; An Extra $4,000 Unlocks 81.8% Accuracy

Xianhang Li* Zeyu Wang* Cihang Xie

*equal technical contribution

UC Santa Cruz

https://github.com/UCSC-VLAA/CLIPA

arXiv:2306.15658v1 [cs.CV] 27 Jun 2023

Figure 1: Compared to OpenCLIP [10], our CLIPA-v2 models achieve higher performance with lower training cost.

Abstract

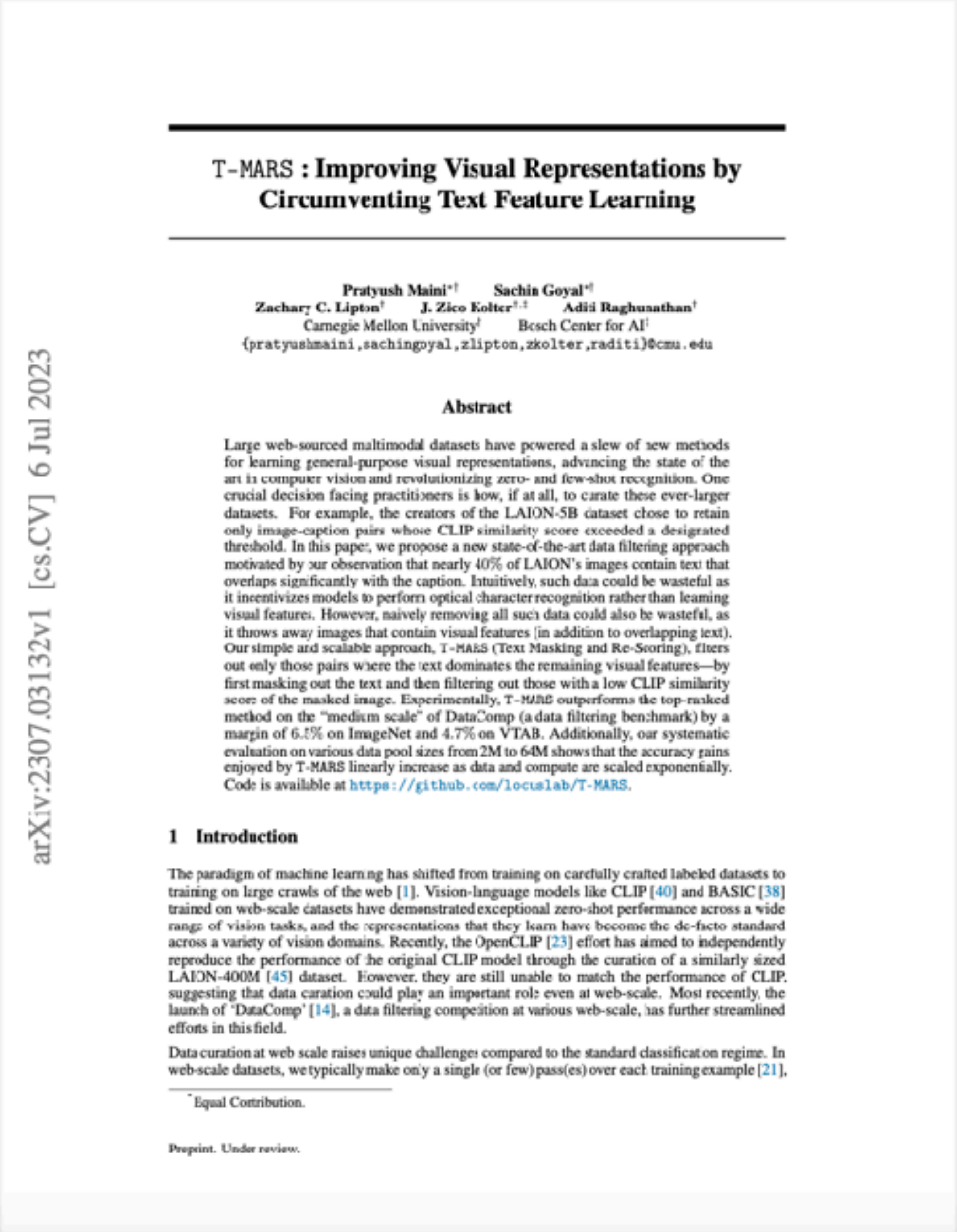
The recent work CLIPA [12] presents an inverse scaling law for CLIP training — whereby the larger the image/text encoders used, the shorter the sequence length of image/text tokens that can be applied in training. This finding enables us to train high-performance CLIP models with significantly reduced computations. Building upon this work, we hereby present CLIPA-v2 with two key contributions. Technically, we find this inverse scaling law is also applicable in the finetuning stage, enabling further reduction in computational needs. Empirically, we explore CLIPA at scale, extending the experiments up to the H/14 model with ~13B image-text pairs seen during training.

Our results are exciting — by only allocating a budget of $10,000, our CLIP model achieves an impressive zero-shot ImageNet accuracy of 81.1%, surpassing the prior best CLIP model (from OpenCLIP, 80.1%) by 1.0% and meanwhile reducing the computational cost by ~39×. Moreover, with an additional investment of $4,000, we can further elevate the zero-shot ImageNet accuracy to 81.8%.

1. Introduction

CLIP [17] has emerged as the pioneering foundation model that bridges the gap between text and images, ushering computer vision research into the "post-ImageNet" era [10, 13, 27, 1, 18, 26, 22, 25, 4]. However, the demanding computational requirements of CLIP hinder its widespread exploration. The recent work CLIPA [12] offers a computationally efficient solution — with the introduction of an inverse scaling law for CLIP training, it reveals that larger models can be trained with fewer input tokens. Building upon this observation, CLIPA demonstrates its efficacy in scenarios with limited computational resources, leading to a substantial reduction in the training cost of CLIP.

This report provides a follow-up on CLIPA. Firstly, we validate that the inverse scaling law is also applicable when finetuning models with input tokens at full resolution. This further reduces the training cost of CLIPA. Secondly, we investigate the performance of CLIPA at scale across various aspects, including model size (up to H/14), data (up to DataComp-1B [6] and LAION-2B [23] datasets), and training schedule (up to ~13B samples seen).



T-MARS : Improving Visual Representations by Circumventing Text Feature Learning

Pratyush Maini*† Sachin Goyal*† Zachary C. Lipton† J. Zico Kolter†‡ Aditi Raghunathan†
Carnegie Mellon University† Bosch Center for AI‡
{pratyushmaini,sachingoyal,zlipton,zkolter,raditi}@cmu.edu

arXiv:2307.03132v1 [cs.CV] 6 Jul 2023

Abstract

Large web-sourced multimodal datasets have powered a slew of new methods for learning general-purpose visual representations, advancing the state of the art in computer vision and revolutionizing zero- and few-shot recognition. One crucial decision facing practitioners is how, if at all, to curate these ever-larger datasets. For example, the creators of the LAION-5B dataset chose to retain only image-caption pairs whose CLIP similarity score exceeded a designated threshold. In this paper, we propose a new state-of-the-art data filtering approach motivated by our observation that nearly 40% of LAION's images contain text that overlaps significantly with the caption. Intuitively, such data could be wasteful as it incentivizes models to perform optical character recognition rather than learning visual features. However, naively removing all such data could also be wasteful, as it throws away images that contain visual features (in addition to overlapping text). Our simple and scalable approach, T-MARS (Text Masking and Re-Scoring), filters out only those pairs where the text dominates the remaining visual features—by first masking out the text and then filtering out those with a low CLIP similarity score of the masked image. Experimentally, T-MARS outperforms the top-ranked method on the "medium scale" of DataComp (a data filtering benchmark) by a margin of 6.5% on ImageNet and 4.7% on VTAB. Additionally, our systematic evaluation on various data pool sizes from 2M to 64M shows that the accuracy gains enjoyed by T-MARS linearly increase as data and compute are scaled exponentially. Code is available at https://github.com/locuslab/T-MARS.

1 Introduction

The paradigm of machine learning has shifted from training on carefully crafted labeled datasets to training on large crawls of the web [1]. Vision-language models like CLIP [40] and BASIC [38] trained on web-scale datasets have demonstrated exceptional zero-shot performance across a wide range of vision tasks, and the representations that they learn have become the de-facto standard across a variety of vision domains. Recently, the OpenCLIP [23] effort has aimed to independently reproduce the performance of the original CLIP model through the curation of a similarly sized LAION-400M [45] dataset. However, they are still unable to match the performance of CLIP, suggesting that data curation could play an important role even at web-scale. Most recently, the launch of 'DataComp' [14], a data filtering competition at various web-scale, has further streamlined efforts in this field.

Data curation at web scale raises unique challenges compared to the standard classification regime. In web-scale datasets, we typically make only a single (or few) pass(es) over each training example [21],

*Equal Contribution.

Preprint. Under review.

Improving Multimodal Datasets with Image Captioning

Thao Nguyen[1] Samir Yitzhak Gadre[2] Gabriel Ilharco[1] Sewoong Oh[1,3] Ludwig Schmidt[1,4,5]

arXiv:2307.10350v1 [cs.LG] 19 Jul 2023

Abstract

Massive web datasets play a key role in the success of large vision-language models like CLIP and Flamingo. However, the raw web data is noisy, and existing filtering methods to reduce noise often come at the expense of data diversity. Our work focuses on caption quality as one major source of noise, and studies how generated captions can increase the utility of web-scraped datapoints with nondescript text. Through exploring different mixing strategies for raw and generated captions, we outperform the best filtering method proposed by the DataComp benchmark by 2% on ImageNet and 4% on average across 38 tasks, given a candidate pool of 128M image-text pairs. Our best approach is also 2× better at Flickr and MS-COCO retrieval. We then analyze what makes synthetic captions an effective source of text supervision. In experimenting with different image captioning models, we also demonstrate that the performance of a model on standard image captioning benchmarks (e.g., NoCaps CIDEr) is not a reliable indicator of the utility of the captions it generates for multimodal training. Finally, our experiments with using generated captions at DataComp's large scale (1.28B image-text pairs) offer insights into the limitations of synthetic text, as well as the importance of image curation with increasing training data quantity.

1 Introduction

Pre-training large multimodal models on image-text pairs sourced from the web has become a standard approach to obtaining high performance on vision tasks [3, 24, 36, 39]. However, raw web data can be noisy or uninformative (Figure 1). Many existing data preprocessing efforts revolve around human-defined heuristics based on image and text content separately—e.g., caption length, presence of nouns, sentence complexity, image aspect ratio, minimum image size [8, 10, 45, 46]—or the reliability of the data source [14]. More complex filtering approaches target poorly aligned image-text pairs, by using trained CLIP models [39] to rank the cosine similarity score between image and text embeddings [45], or ensuring mentions of image objects in the captions [46]. These approaches discard between 60% to 90% of the initial data collected, regardless of whether the images themselves are suitable for training.

In this work, we seek to restore the utility of such discarded examples with the help of synthetic captions. To do so, we leverage the DataComp benchmark [18], where initial data processing is kept to a minimum, i.e. only filtering out NSFW examples and train-test overlap. This allows us to perform controlled experiments on the raw Common Crawl data and bypass subjective human-design choices that may be employed in the creation of other datasets (e.g., LAION-5B [45]). We study

[1]University of Washington. Correspondence to thaotn@cs.washington.edu [2]Columbia University [3]Google Research [4]Allen Institute for Artificial Intelligence [5]LAION

1

On the horizon

- DataComp NLP (DCNLP)
- Larger pools (100B candidate image-text pairs)
- DataComp for image generation
- Multimodal DataComp?
- Interleaved DataComp?

Everything is open source

- Central webpage: datacomp.ai
- Main repo: github.com/mlfoundations/datacomp
- CLIP training code: github.com/mlfoundations/open_clip
- Downloading billions of image-text pairs: github.com/rom1504/img2dataset
- Processing metadata for billions of image-text pairs: github.com/mlfoundations/dataset2metadata