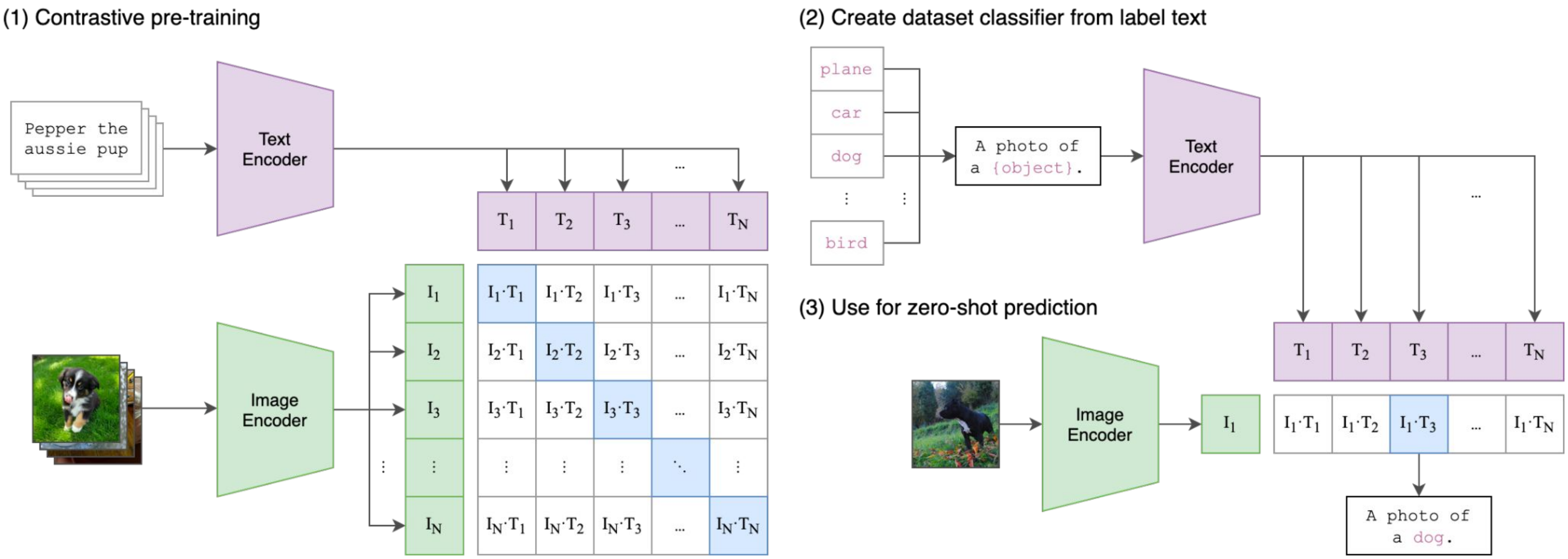
A study of CLIP on model robustness and data imbalance

Xin Wen, HKU

16 Aug, 2024

Background: CLIP



Radford et al., *Learning Transferable Visual Models From Natural Language Supervision*, ICML 2021

Strong zero-shot cls. thanks to scaling law

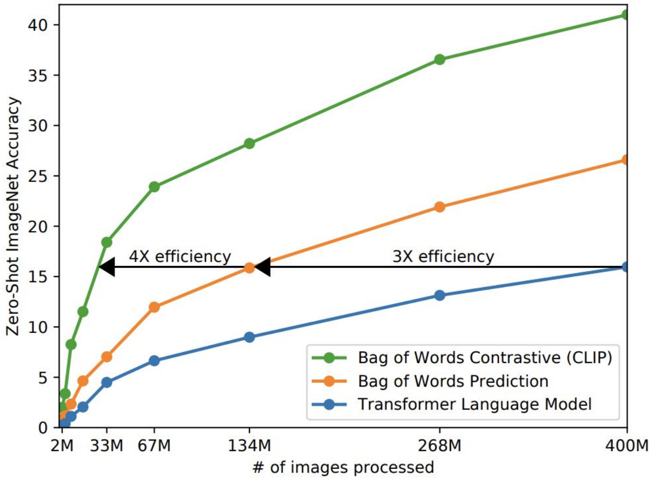


Figure 2. **CLIP is much more efficient at zero-shot transfer than our image caption baseline.** Although highly expressive, we found that transformer-based language models are relatively weak at zero-shot ImageNet classification. Here, we see that it learns 3x slower than a baseline which predicts a bag-of-words (BoW) encoding of the text (Joulin et al., 2016). Swapping the prediction objective for the contrastive objective of CLIP further improves efficiency another 4x.

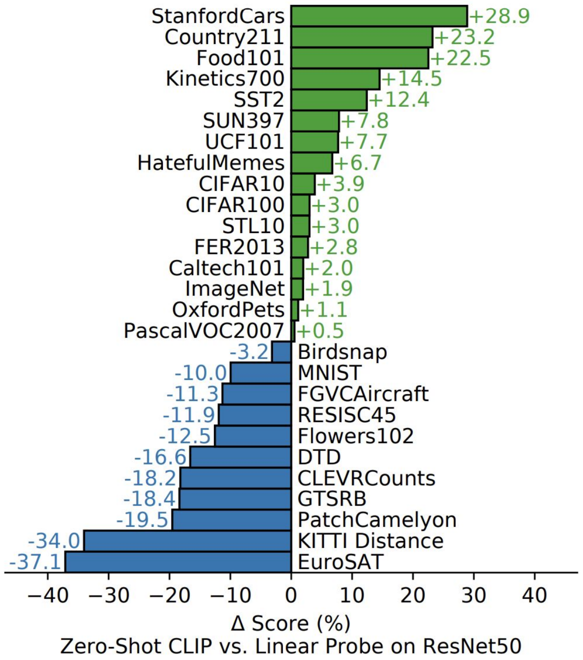
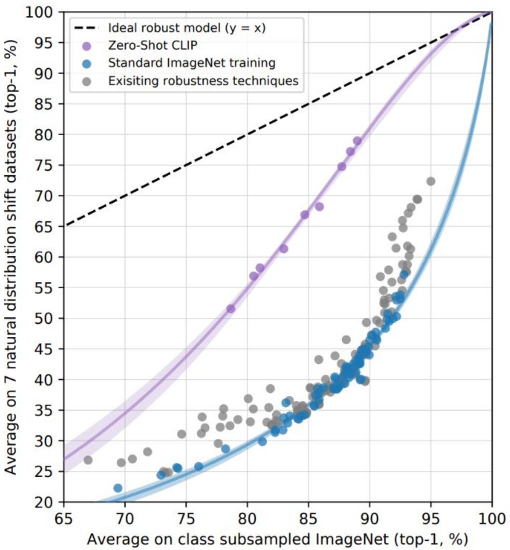


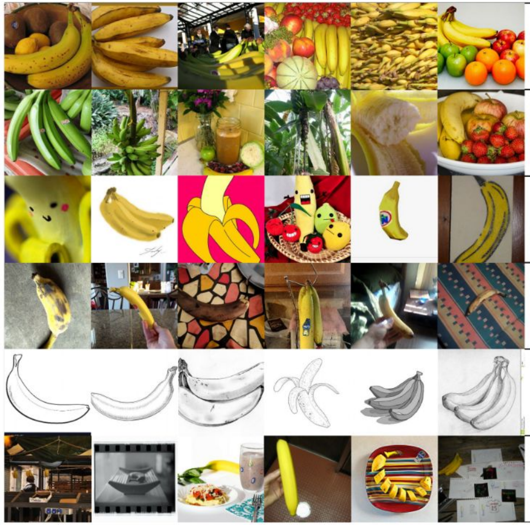
Figure 4. **Zero-shot CLIP is competitive with a fully supervised baseline.** Across a 27 dataset eval suite, a zero-shot CLIP classifier outperforms a fully supervised linear classifier fitted on ResNet50 features on 16 datasets, including ImageNet.

Radford et al., *Learning Transferable Visual Models From Natural Language Supervision*, ICML 2021

Strong robustness to OOD test data



Dataset Examples



Dataset	ImageNet ResNet101	Zero-Shot CLIP	Δ Score
ImageNet	**76.2**	**76.2**	0%
ImageNetV2	64.3	**70.1**	+5.8%
ImageNet-R	37.7	**88.9**	+51.2%
ObjectNet	32.6	**72.3**	+39.7%
ImageNet Sketch	25.2	**60.2**	+35.0%
ImageNet-A	2.7	**77.1**	+74.4%

Figure 7. **Zero-shot CLIP is much more robust to distribution shift than standard ImageNet models.** (Left) An ideal robust model (dashed line) performs equally well on the ImageNet distribution and on other natural image distributions. Zero-shot CLIP models shrink this "robustness gap" by up to 75%. Linear fits on logit transformed values are shown with bootstrap estimated 95% confidence intervals. (Right) Visualizing distribution shift for bananas, a class shared across 5 of the 7 natural distribution shift datasets. The performance of the best zero-shot CLIP model is compared with a model that has the same performance on the ImageNet validation set, ResNet101.

Radford et al., *Learning Transferable Visual Models From Natural Language Supervision*, ICML 2021

Chapter 1 (can be skipped):

Where does CLIP's
OOD robustness come from?

How to measure OOD robustness: effective robustness

- Let's first look at effective robustness
  - Def. as the slope between ID & OOD acc.
  - Expected to be like *y=x*
- OOD robustness is one important property of CLIP
- Then how to study it?
  - Mainly from a data-centric perceptive
  - Ie, to ablate the data to train CLIP on

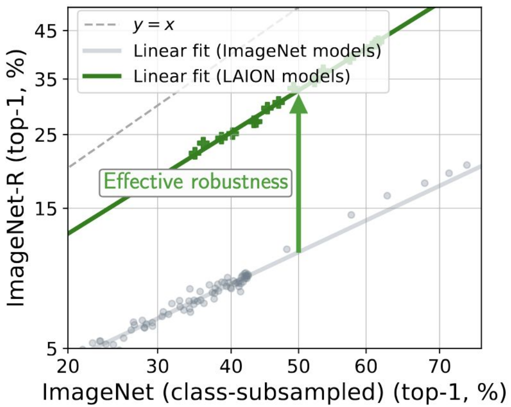


Figure 1: Models pre-trained on LAION exhibit *effective robustness* [75] compared to standard models trained on ImageNet. Effective robustness is defined as movement towards a classifier which is robust to distribution shift. A classifier is more robust the closer it is to the $y = x$ line. A classifier on the $y = x$ line is not affected by the distribution shift.

Ludwig Schmidt
AP at ~~UW~~ Stanford

Member of LAION and Allen AI

Author of ImageNetV2 and Effective Robustness

Lead of OpenCLIP, OpenFlamingo, and LAION-5B

Ablate on different web data: YFCC, LAION, WIT, RedCaps,etc. no much difference between

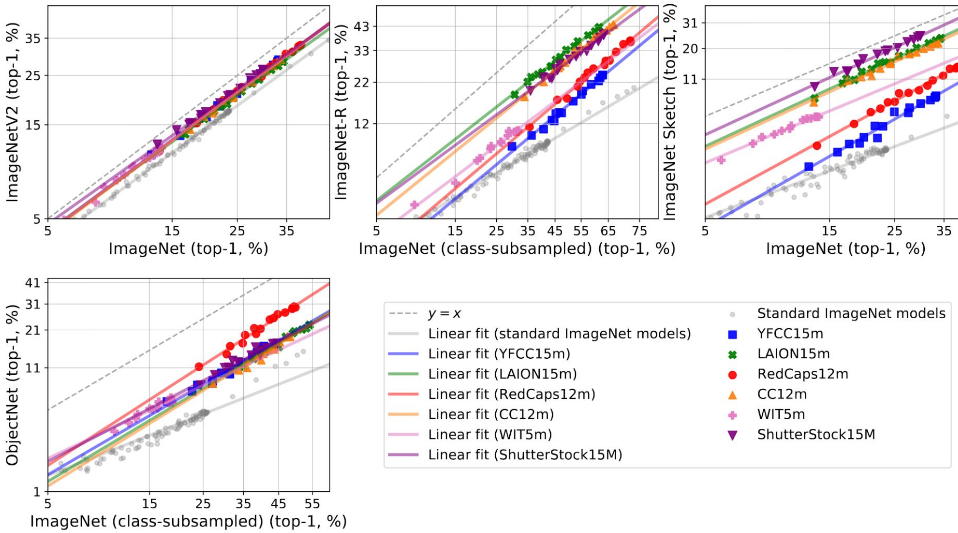


Figure 2: **Performance of the six pre-training data sources under various distribution shifts.** We find that the behavior—both in terms of accuracy and the slope of the linear trend—of the pre-training data varies substantially across distribution shifts, with no single data source dominating. Most shifts help highlight the strengths and weaknesses of different data sources, except for ImageNet-V2, where the linear trends produced by individual sources are highly correlated with one another.

- YFCC: We experiment with the 15M subset of the YFCC100M dataset [76] that the original CLIP paper [61] used for dataset ablation studies. The images and captions are collected from Flickr.
- LAION [68]: The images and corresponding alt-texts come from web pages collected by Common Crawl [1] between 2014 and 2021. We randomly select a subset of 15M samples to experiment with, and ensure that the accompanying NSFW tags of all chosen images are 'UNLIKELY'.
- Conceptual Captions [13]: We use CC-12M for our experiments, which consists of images and HTML alt-text from an unspecified set of web pages.
- RedCaps [20]: This dataset contains 12M examples, obtained from 350 manually curated subreddits between 2008 and 2020. The subreddits are selected to contain a large number of image posts that are mostly photographs and not images of people.
- Shutterstock: 15M images and captions were crawled from the Shutterstock website in 2021.
- WIT [71]: Image-text pairs come from Wikipedia pages. We use reference description as the source of text data and obtain 5M examples in total after filtering to include only English language examples.

Appendix A.1 contains an analysis of image and text statistics, as well as randomly selected data samples from each source.

Evaluation. Similar to Taori et al. [75] and Radford et al. [61], we choose ImageNet as the reference distribution and evaluate CLIP on four natural distribution shifts derived from ImageNet:

- ImageNet-V2 [65]: A reproduction of the ImageNet validation set closely following the original dataset creation process.
- ImageNet-R [36]: Renditions (e.g., sculptures, paintings, etc.) for 200 ImageNet classes.
- ImageNet-Sketch [81]: Sketches of ImageNet class objects.
- ObjectNet [6]: A test set of objects in novel backgrounds, rotations, and viewpoints with 113 classes overlapping with ImageNet

Nguyen et al., *Quality Not Quantity: On the Interaction between Dataset Design and Robustness of CLIP,* NeurIPS 2022

Mixing web datasets do not introduce extra robustness, but rather a decline

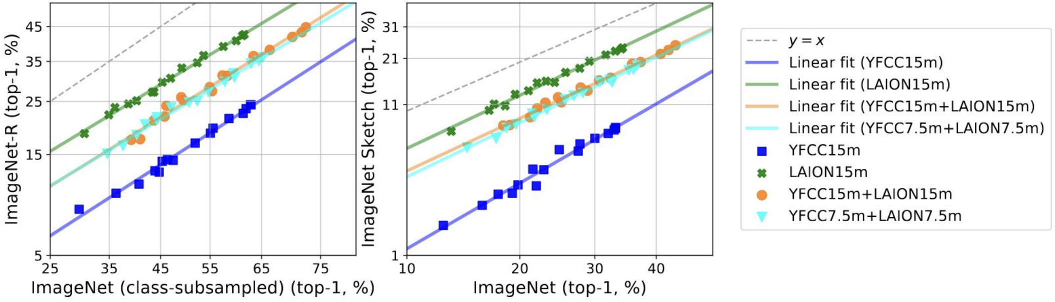


Figure 3: **Combining YFCC and LAION training data in equal ratios produces models with intermediate robustness.** Given a fixed data budget of 15M samples, the linear trend produced by training CLIP on a YFCC-LAION data mixture, with 7.5M datapoints from each source (cyan line), lies between that of training CLIP on YFCC (blue line) and LAION (green line) entirely. Even when we increase the total training set size (30M) and use all data available from both sources (orange line), the same pattern persists.

- YFCC: We experiment with the 15M subset of the YFCC100M dataset [76] that the original CLIP paper [61] used for dataset ablation studies. The images and captions are collected from Flickr.
- LAION [68]: The images and corresponding alt-texts come from web pages collected by Common Crawl [1] between 2014 and 2021. We randomly select a subset of 15M samples to experiment with, and ensure that the accompanying NSFW tags of all chosen images are 'UNLIKELY'.
- Conceptual Captions [13]: We use CC-12M for our experiments, which consists of images and HTML alt-text from an unspecified set of web pages.
- RedCaps [20]: This dataset contains 12M examples, obtained from 350 manually curated subreddits between 2008 and 2020. The subreddits are selected to contain a large number of image posts that are mostly photographs and not images of people.
- Shutterstock: 15M images and captions were crawled from the Shutterstock website in 2021.
- WIT [71]: Image-text pairs come from Wikipedia pages. We use reference description as the source of text data and obtain 5M examples in total after filtering to include only English language examples.

Appendix A.1 contains an analysis of image and text statistics, as well as randomly selected data samples from each source.

Evaluation. Similar to Taori et al. [75] and Radford et al. [61], we choose ImageNet as the reference distribution and evaluate CLIP on four natural distribution shifts derived from ImageNet:

- ImageNet-V2 [65]: A reproduction of the ImageNet validation set closely following the original dataset creation process.
- ImageNet-R [36]: Renditions (e.g., sculptures, paintings, etc.) for 200 ImageNet classes.
- ImageNet-Sketch [81]: Sketches of ImageNet class objects.
- ObjectNet [6]: A test set of objects in novel backgrounds, rotations, and viewpoints with 113 classes overlapping with ImageNet

Nguyen et al., *Quality Not Quantity: On the Interaction between Dataset Design and Robustness of CLIP*, NeurIPS 2022

They then looked into the distribution of web data: i.e., difference between ImageNet and LAION

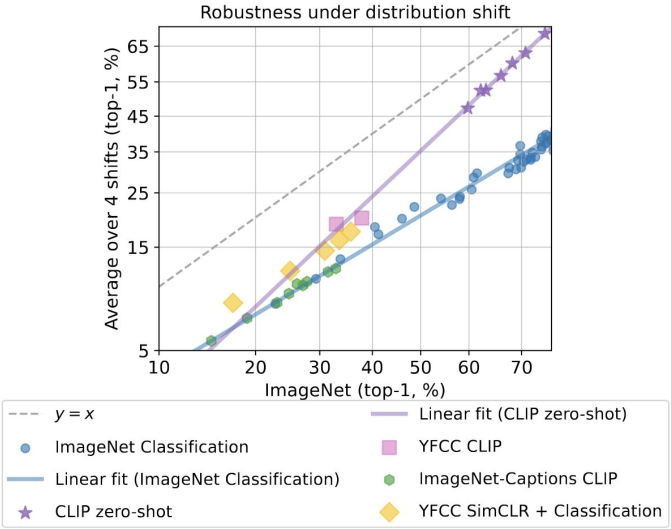


Figure 1: We compare models trained using different methods and on different datasets, measuring their robustness on a range of natural distribution shifts (ImageNetV2, ImageNet-R, ImageNet-Sketch, and ObjectNet). The CLIP models stand out with their consistent performance in the presence of distribution shift. We find that large gains in effective robustness (improvement over ImageNet models) only come from varying the training distribution. Language supervision alone does not cause robustness.

Nevertheless, there is still a long list of possible causes for CLIP's robustness:

- The large training set size (400 million images)
- The training distribution
- Language supervision at training time
- Language supervision at test time via prompts
- The contrastive loss function

Fang et al., *Data Determines Distributional Robustness in Contrastive Language Image Pre-training (CLIP),* ICML 2023

For controlled study, they need a captioned version of ImageNet, and a classification version of YFCC, how?

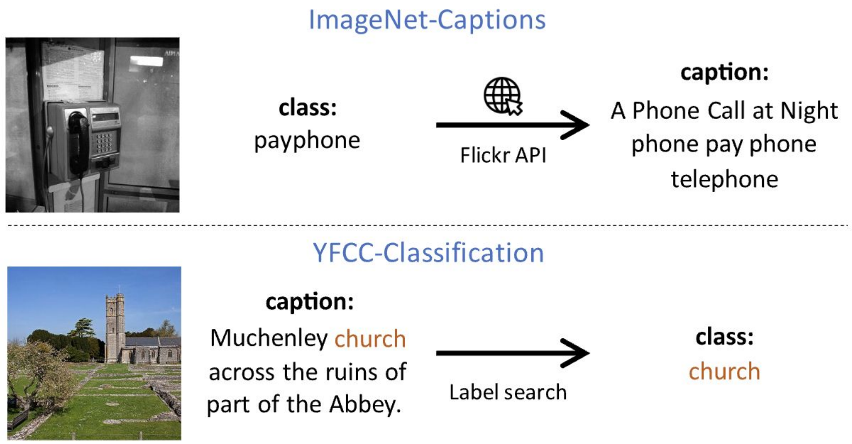


Figure 2: Overview of the two main training sets in our experiments. (Top) We introduce the ImageNet-Captions dataset, where we *augment* a subset of the ImageNet 2012 training set images with the corresponding original captions collected from Flickr. (Bottom) We convert the YFCC image-caption dataset into YFCC-Classification by searching for class labels in the YFCC captions and then *removing* the text annotations. These two datasets allows us to evaluate the impact of language-image training on robustness because we can compare language-image training with standard classification training *on the same set of images*.

1. Going back to where ImageNet was collected, and retrieve corresponding metadata, including *original captions*.
2. Use substring matching on YFCC captions to get their labels
   - There can be better ways, but this was good enough

Figure 3: Three sample images from ImageNet-Captions. Their respective ImageNet labels are: drake, rocking chair, payphone.

Fang et al., *Data Determines Distributional Robustness in Contrastive Language Image Pre-training (CLIP),* ICML 2023

For OOD robustness, it doesn't matter if language is used

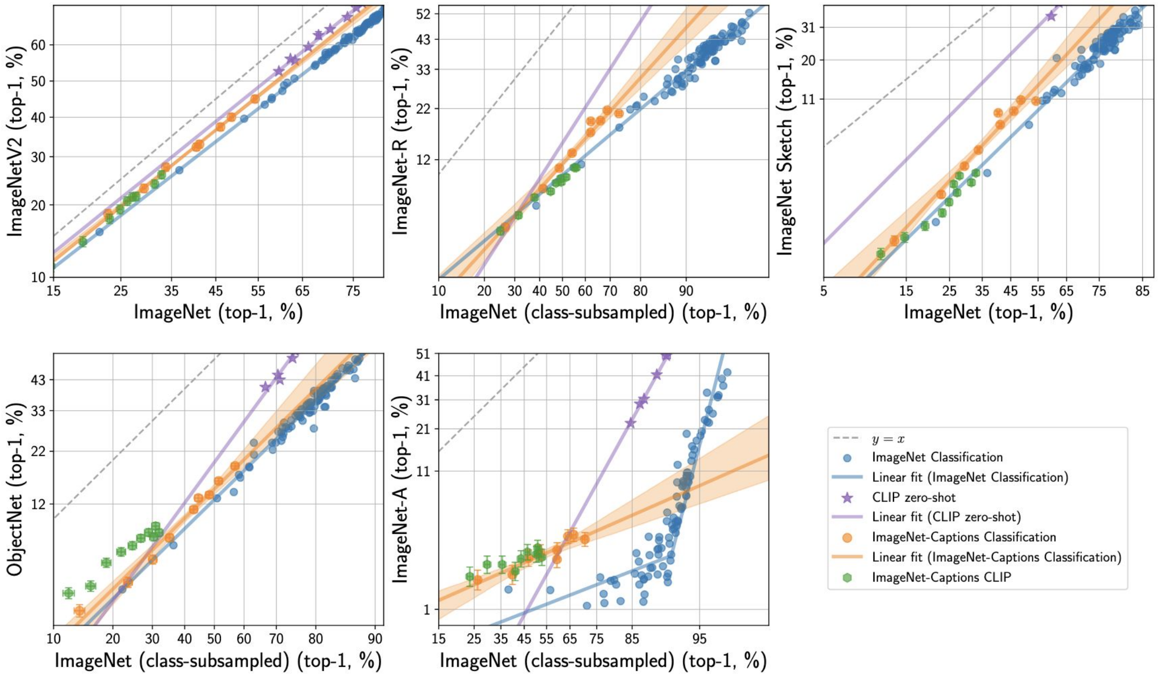


Figure 4: On most natural distribution shifts, models trained with language information from ImageNet-Captions follow the same trend as models trained without it. Neither comes close to achieving the robustness of OpenAI's CLIP models.

Fang et al., *Data Determines Distributional Robustness in Contrastive Language Image Pre-training (CLIP),* ICML 2023

A classification model can be as robust when trained on web data (YFCC in this case)

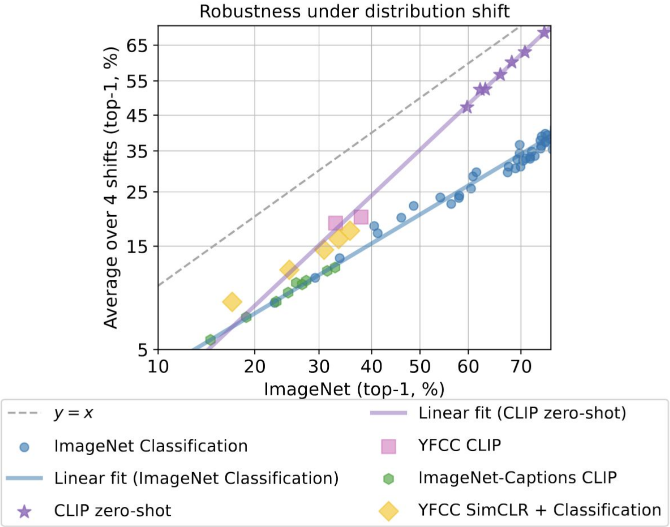


Figure 1: We compare models trained using different methods and on different datasets, measuring their robustness on a range of natural distribution shifts (ImageNetV2, ImageNet-R, ImageNet-Sketch, and ObjectNet). The CLIP models stand out with their consistent performance in the presence of distribution shift. We find that large gains in effective robustness (improvement over ImageNet models) only come from varying the training distribution. Language supervision alone does not cause robustness.

Table 3: Comparing CLIP training with (language model free) classification models on YFCC-15M. All experiments use a ViT-B/16 backbone. The CLIP results are from Mu et al. [24]. Image-only contrastive learning followed by a simple text matching stage for classification nearly matches the performance of CLIP with a full language model.

Training style	ImageNet	Avg OOD
CLIP	37.9	19.9
SimCLR → Classification	35.7	18.8

Fang et al., *Data Determines Distributional Robustness in Contrastive Language Image Pre-training (CLIP),* ICML 2023

Another German team then asks: does CLIP's generalization come from info leak (test ⊆ train)?

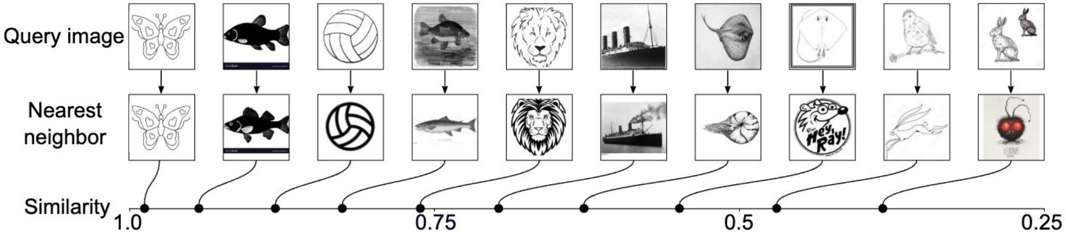


Figure 2: **Relation between *perceptual similarity* and visual closeness of nearest neighbors**. Query

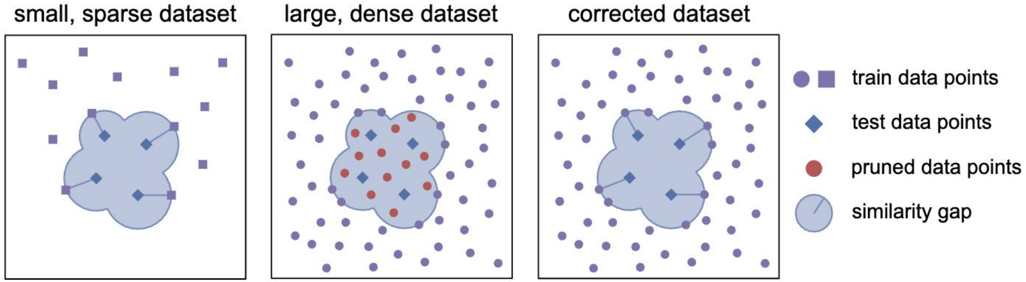
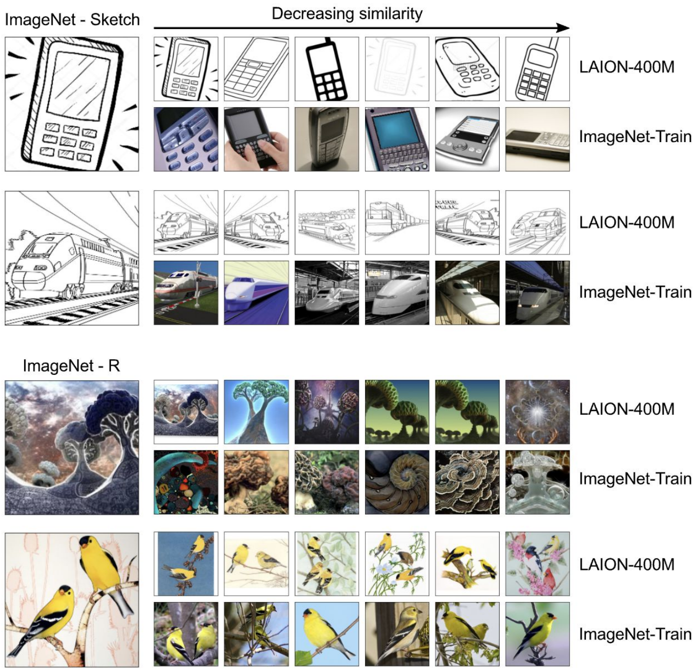


Figure 5: **Aligning the similarity gap of two datasets**. A larger, denser, more diverse dataset likely contains samples more similar to given test points than a smaller, sparser one. To control for this, we compute the nearest-neighbor similarity of each test point to the smaller dataset (left) and prune points from the larger dataset that lie within this hull (center). We end up with a corrected large dataset replicating the *similarity gap* of the small one (right).



Mayilvahanan et al., *Does CLIP's generalization performance mainly stem from high train-test similarity?,* ICLR 2024

Another German team then asks: does CLIP's generalization come from info leak (test ⊆ train)? TL;DR: No.

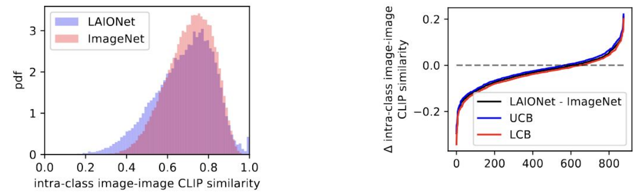
Dataset	Size	Top-1 Accuracy					
		Val	Sketch	A	R	V2	ON
OpenAI (Radford et al., 2021)	400 000 000	63.38	42.32	31.44	69.24	55.96	44.14
L-400M (Schuhmann et al., 2021)	413 000 000	62.94	49.39	21.64	73.48	55.14	43.94
L-200M	199 824 274	62.12	48.61	21.68	72.63	54.16	44.80
L-200M + IN-Train	200 966 589	68.66	50.21	23.33	72.9	59.7	43.99
├─ val-pruned	*−377 340*	68.62	49.58	23.47	72.74	59.47	45.08
├─ sketch-pruned	*−8 342 783*	68.34	44.78	22.7	69.35	59.52	44.12
├─ a-pruned	*−138 852*	68.85	50.25	22.99	72.44	60.05	44.43
├─ r-pruned	*−5 735 749*	68.71	46.92	23.44	69.48	59.6	45.08
├─ v2-pruned	*−274 325*	68.79	50.45	23.19	72.58	59.84	45.33
├─ objectnet-pruned	*−266 025*	68.75	50.14	22.70	72.82	59.37	43.73
└─ combined-pruned	*−12 352 759*	68.05	44.12	22.15	67.88	58.61	44.39

Mayilvahanan et al., *Does CLIP's generalization performance mainly stem from high train-test similarity?*, ICLR 2024

Chapter 2:
How does imbalance in pre-training data interact with model performance?

What makes ImageNet look unlike LAION?

TL;DR: highly imbalanced class distribution, and lower intra-class similarity



(a) Dist. of aggregated similarities

(b) Comparison across classes

Figure 6: Comparing the intra-class similarity of LAIONet and ImageNet. (a) In each class, pairwise similarities of LAIONet images are sampled to match ImageNet in number. All the classes combined, the distribution of intra-class similarity is depicted. (b) For each class, the average intra-class similarity of ImageNet images was subtracted from the same value in LAIONet. The blue and red curves show upper and lower 95% confidence intervals. All values are sorted ascendingly.

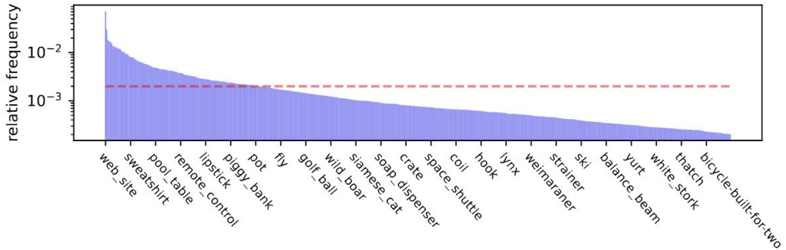


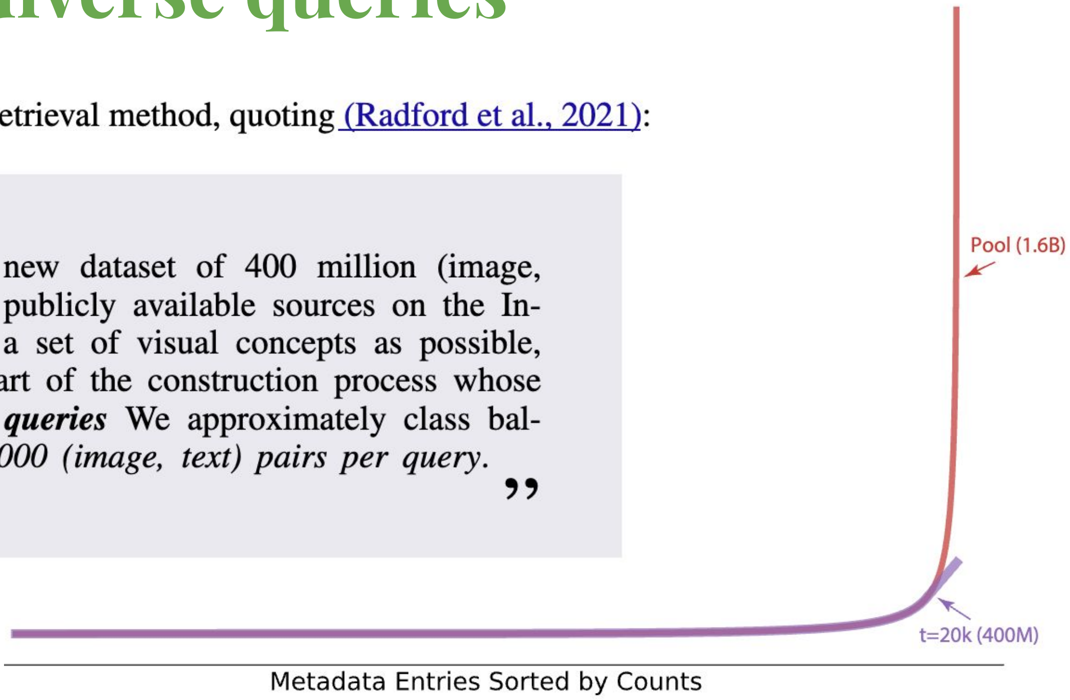
Figure 4: Relative frequencies of different classes in LAIONet sorted in descending order for the 500 most frequent classes. Some class names shown. Red line shows uniform weight.

Shirali et al., *What Makes ImageNet Look Unlike LAION*, arXiv 2023

Community efforts in rebalancing data curation: MetaCLIP re-implements OpenAI's WIT-400M by retrieving 500k diverse queries

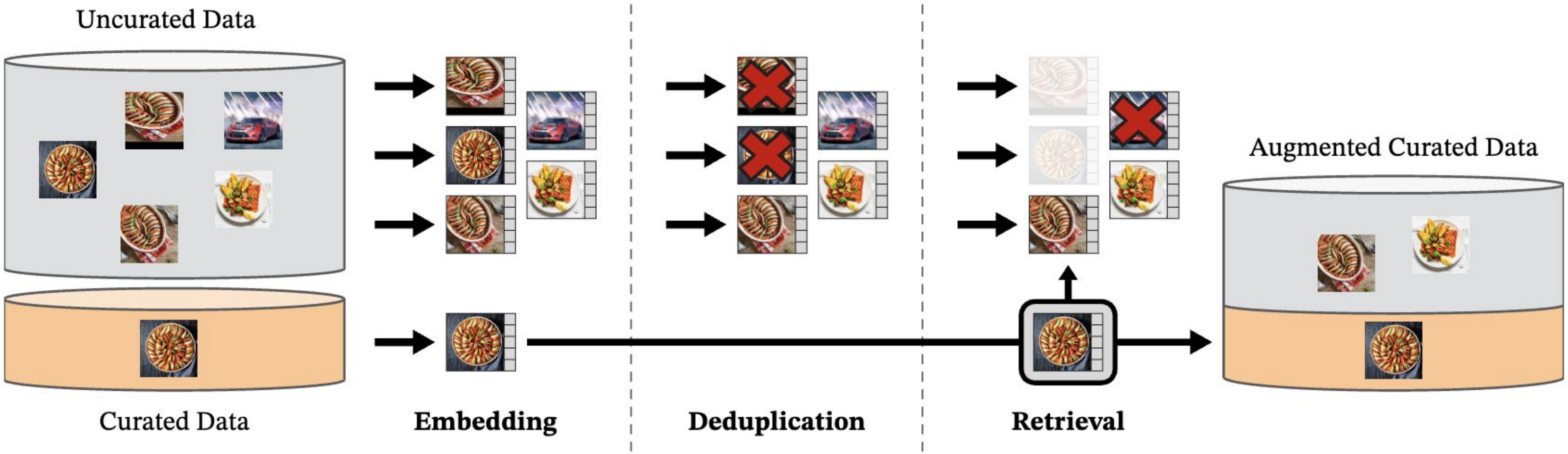
CLIP's WIT400M is curated with an information retrieval method, quoting (Radford et al., 2021):

> "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*."



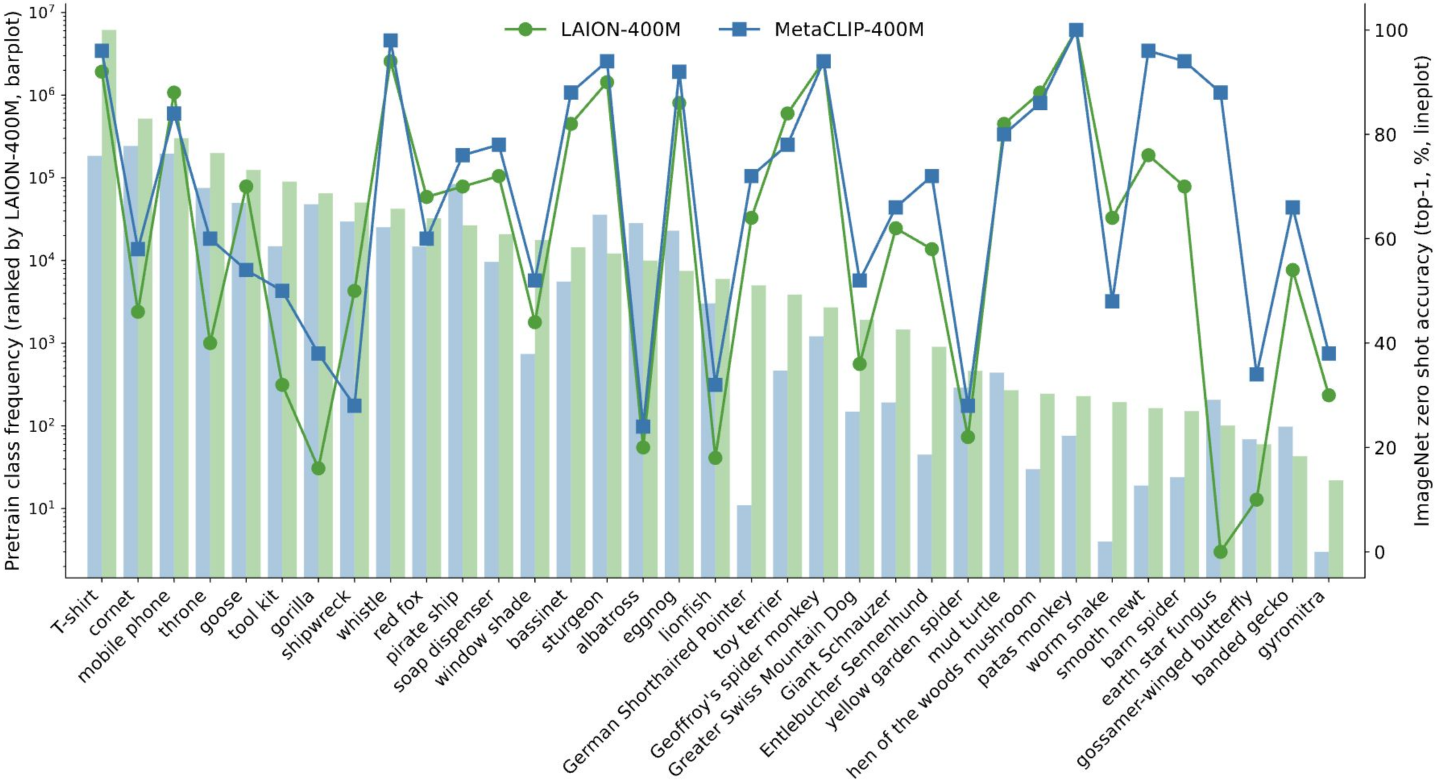
Xu et al., *Demystifying CLIP Data (MetaCLIP)*, ICLR 2024

Community efforts in rebalancing data curation: Deduplication also matters for DINOv2



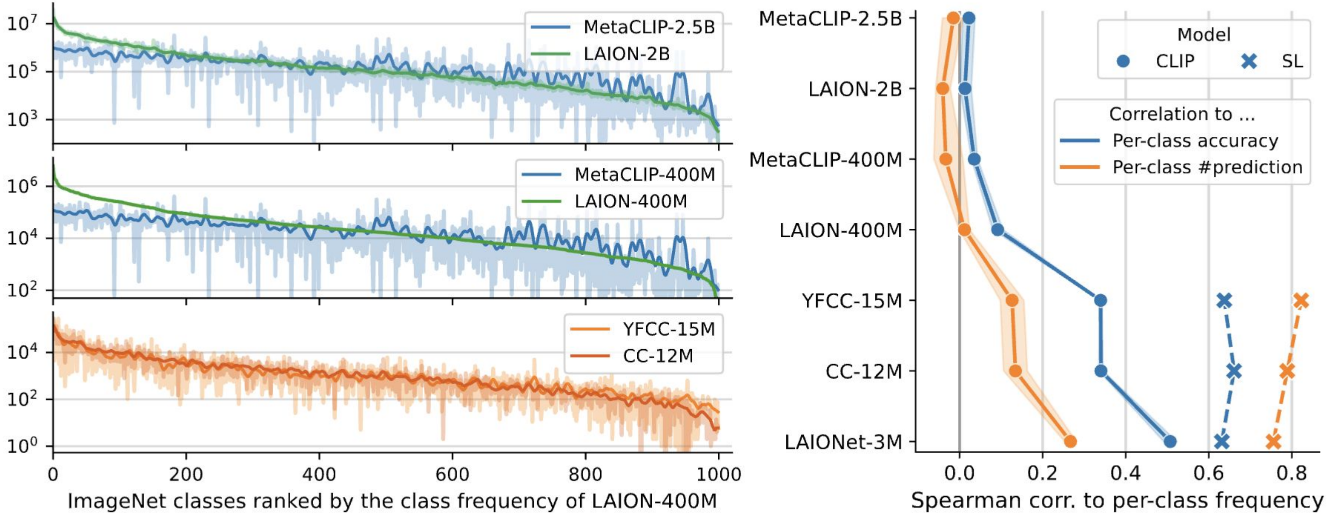
Oquab et al., *DINOv2: Learning Robust Visual Features without Supervision*, TMLR 2024

Looking into the concept distribution of web datasets



Wen et al., *What Makes CLIP More Robust to Long-Tailed Pre-Training Data? A Controlled Study for Transferable Insights,* NeurIPS 2024

A shared long-tail, and a scaling law against it



(a) Class frequencies (log scale) ranked by LAION-400M. (b) Correlation between class-wise statistics.

Figure 1: Per-class statistics of image-text datasets and models trained on top. (a) A highly imbalanced class distribution is *shared* across datasets.[1] (b) Compared to supervised learning (✖ SL), CLIP's performance (measured by ● accuracy) is *less biased* by data frequency, and the classifier is notably uncorrelated (measured by model's number of ● prediction per class). Besides, the correlation narrows as data scales up. Both aspects indicate implicit re-balancing mechanisms exist in CLIP.

Wen et al., *What Makes CLIP More Robust to Long-Tailed Pre-Training Data? A Controlled Study for Transferable Insights,* NeurIPS 2024

A data-centric toolbox for controlled ablations

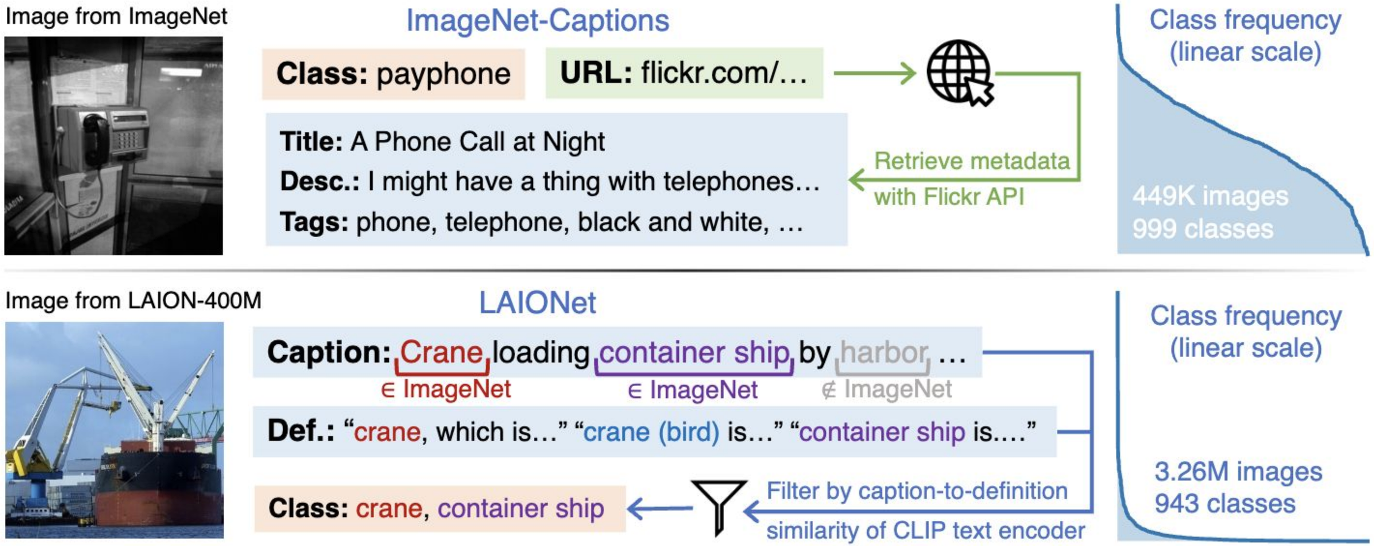


Figure 2: Curation process and distribution of datasets used in our controlled study. Top: IN-Caps [26] augments train images of ImageNet with texts by querying Flickr with image URLs. The texts include title, description, and tags. Bottom: LAIONet [74] is a filtered subset of LAION-400M [70], obtained by matching ImageNet classes with captions and filtering by CLIP text encoder for disambiguation.

Wen et al., *What Makes CLIP More Robust to Long-Tailed Pre-Training Data? A Controlled Study for Transferable Insights,* NeurIPS 2024

Green dots: (Descriptive) language as supervision signal

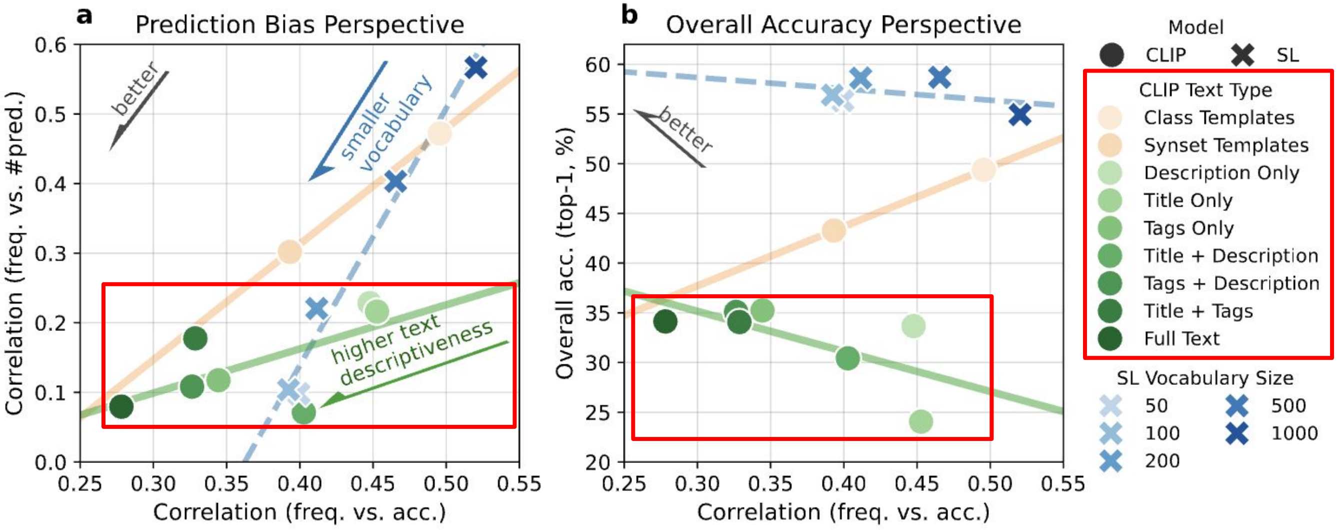


Figure 3: Results on IN-Caps about ● text descriptiveness and ✖ vocabulary size. 1) Increasing ● text descriptiveness improves both robustness (a) and discriminability (b) of CLIP, but the tendency varies if using ● less descriptive (template-based) supervision. 2) The gap between SL and CLIP (a) implies CLIP re-balances predictions, which is replicable by ✖ subsampling the vocabulary SL trains with.

Train on IN-Caps w/ texts of different level of informativeness (or descriptiveness)

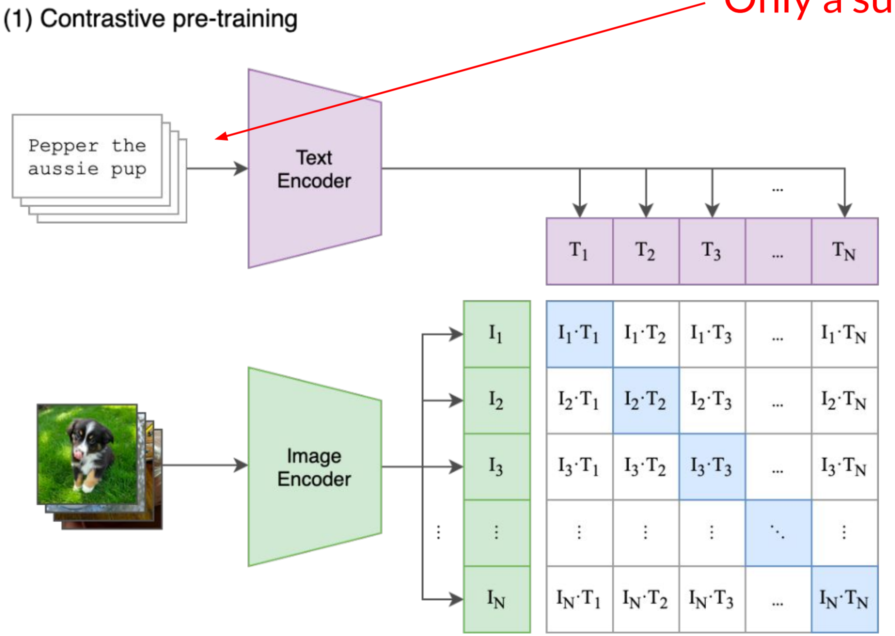
Better texts: better performance & less bias

However, CLIP w/ class templates is still more robust than SL

Wen et al., *What Makes CLIP More Robust to Long-Tailed Pre-Training Data? A Controlled Study for Transferable Insights,* NeurIPS 2024

Recall CLIP loss (InfoNCE, contrastive loss)

Only a subset of all classes!




```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

Radford et al., *Learning Transferable Visual Models From Natural Language Supervision*, ICML 2021

Blue crosses: Dynamic classification (using subsampled vocabulary) as pretext task

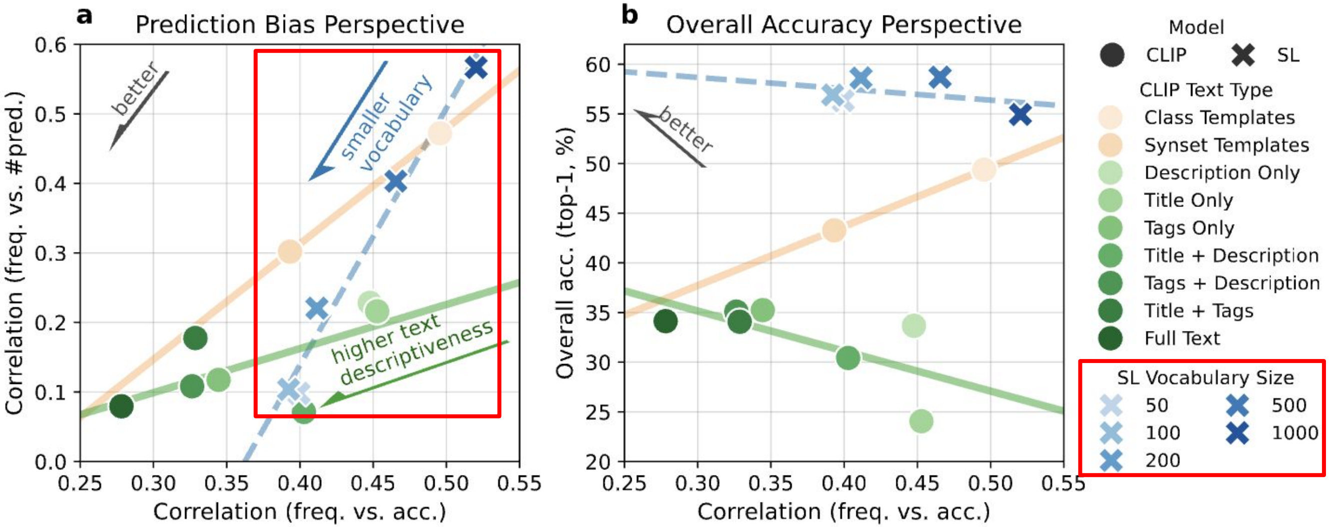


Figure 3: Results on IN-Caps about ● text descriptiveness and ✖ vocabulary size. 1) Increasing ● text descriptiveness improves both robustness (a) and discriminability (b) of CLIP, but the tendency varies if using ● less descriptive (template-based) supervision. 2) The gap between SL and CLIP (a) implies CLIP re-balances predictions, which is replicable by ✖ subsampling the vocabulary SL trains with.

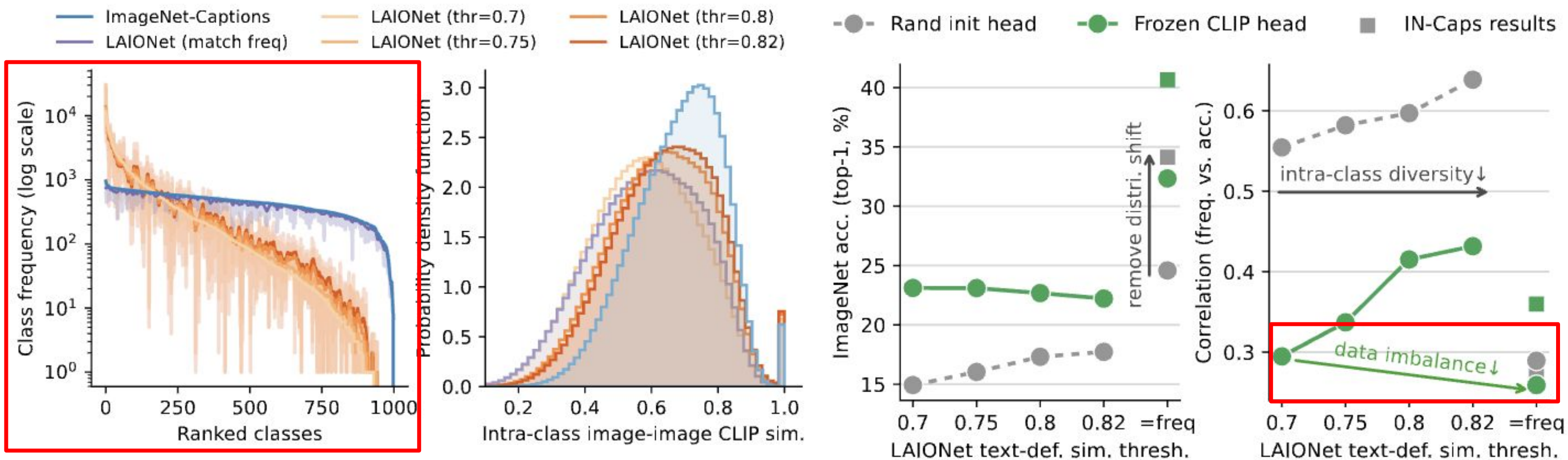
Let SL models only to discriminate from a (dynamic) subset of all classes in each forward.

This technique is also called “federated loss” in open-vocabulary object detection.

We find it can effectively re-balance learning signals.

Wen et al., *What Makes CLIP More Robust to Long-Tailed Pre-Training Data? A Controlled Study for Transferable Insights,* NeurIPS 2024

Data factors (data imbalance 👎, diversity, and distribution shift)

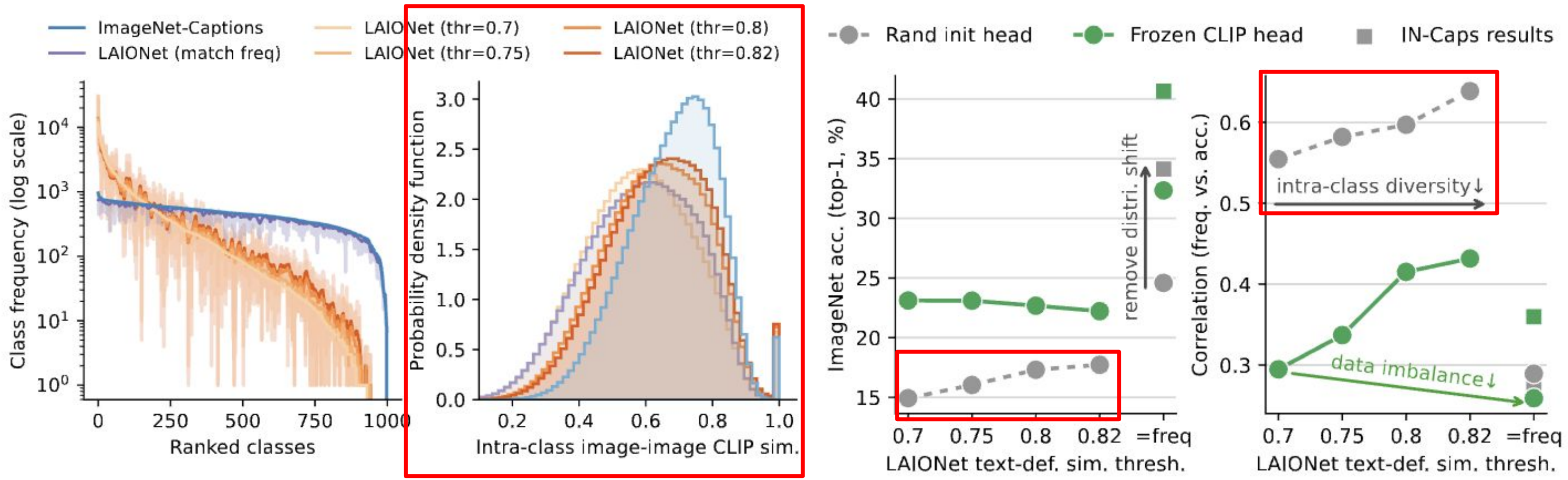


(a) Distrib. of LAIONet variants (same scale as IN-Caps). (b) Results of CLIP trained on LAIONet variants.

Figure 4: Results on LAIONet about data distribution (level of data imbalance, distribution shift, and data diversity). 1) Extreme data imbalance makes models more prone to bias (last column *vs*. others). 2) Distribution shift (●● *vs*. ■■, last column) harms discriminability but could improve robustness if pre-trained text head is used. 3) Higher data diversity (smaller threshold) also improves robustness.

Wen et al., *What Makes CLIP More Robust to Long-Tailed Pre-Training Data? A Controlled Study for Transferable Insights,* NeurIPS 2024

Data factors (data imbalance, diversity👍, and distribution shift)

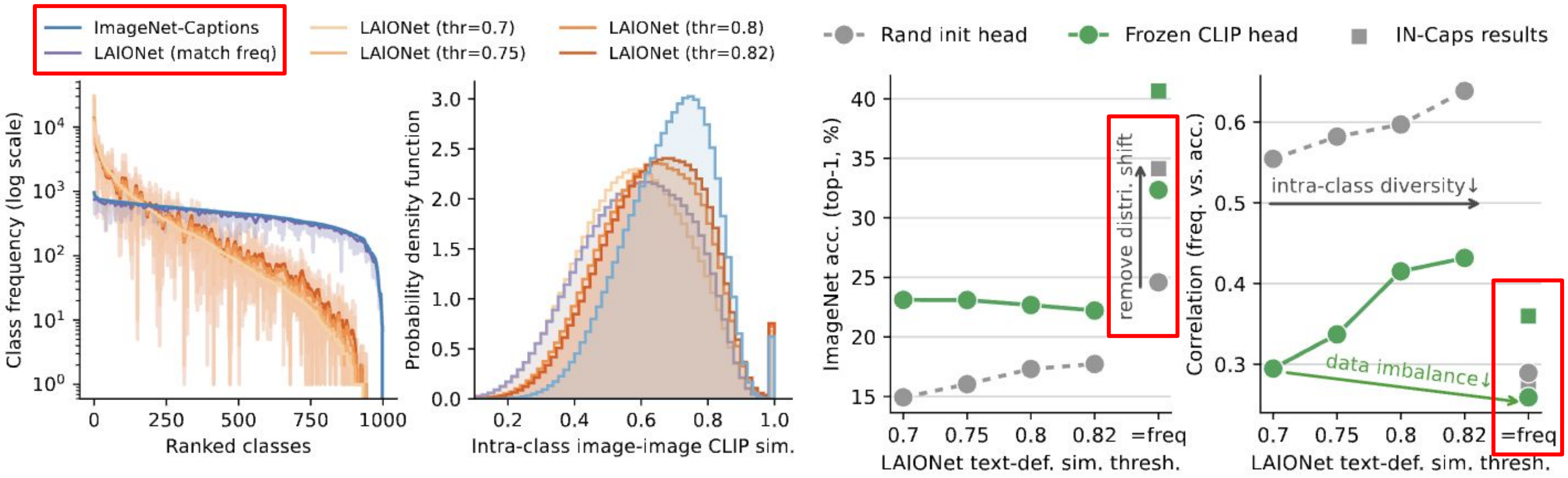


(a) Distrib. of LAIONet variants (same scale as IN-Caps). (b) Results of CLIP trained on LAIONet variants.

Figure 4: Results on LAIONet about data distribution (level of data imbalance, distribution shift, and data diversity). 1) Extreme data imbalance makes models more prone to bias (last column *vs*. others). 2) Distribution shift (●● *vs*. ■■, last column) harms discriminability but could improve robustness if pre-trained text head is used. 3) Higher data diversity (smaller threshold) also improves robustness.

Wen et al., *What Makes CLIP More Robust to Long-Tailed Pre-Training Data? A Controlled Study for Transferable Insights,* NeurIPS 2024

Data factors (data imbalance, diversity, and distribution shift👍)



(a) Distrib. of LAIONet variants (same scale as IN-Caps). (b) Results of CLIP trained on LAIONet variants.

Figure 4: Results on LAIONet about data distribution (level of data imbalance, distribution shift, and data diversity). 1) Extreme data imbalance makes models more prone to bias (last column *vs*. others). 2) Distribution shift (●● *vs*. ■■, last column) harms discriminability but could improve robustness if pre-trained text head is used. 3) Higher data diversity (smaller threshold) also improves robustness.

Wen et al., *What Makes CLIP More Robust to Long-Tailed Pre-Training Data? A Controlled Study for Transferable Insights,* NeurIPS 2024

Data scaling (also achievable via CLIP language pre-training)

Pre-trained knowledge help preserve intra-class variation while not harming inter-class margins

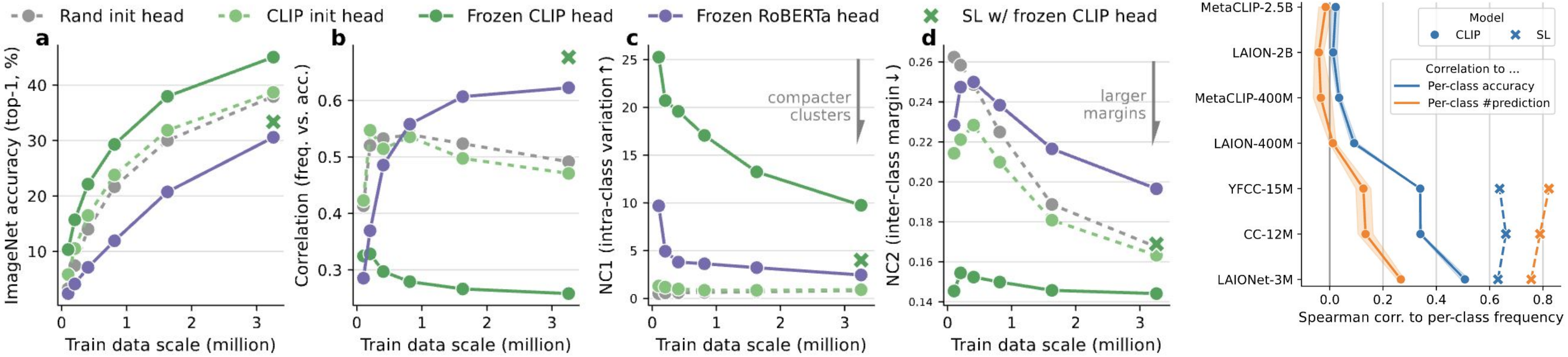


Figure 5: Results on LAIONet subsets about data scale and text encoder. 1) CLIP's discriminability (a) and robustness (b) co-improve as data scales up, and can be boosted by pre-trained heads. 2) A frozen head helps CLIP preserve intra-class variation (c) while not harming margins (d), which can be lost if fine-tuned. It is also unattainable by SL even using the same head. 3) Language pre-training using CLIP is more favorable for image-text tasks than pure language modeling (*e.g.*, RoBERTa [48]).

Wen et al., *What Makes CLIP More Robust to Long-Tailed Pre-Training Data? A Controlled Study for Transferable Insights,* NeurIPS 2024

Utilization of open-world knowledge

—

Something native for language, but hard for SL models

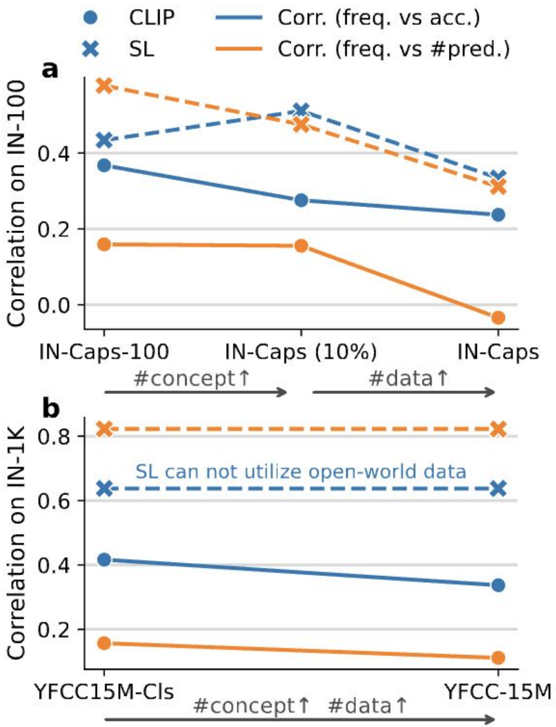
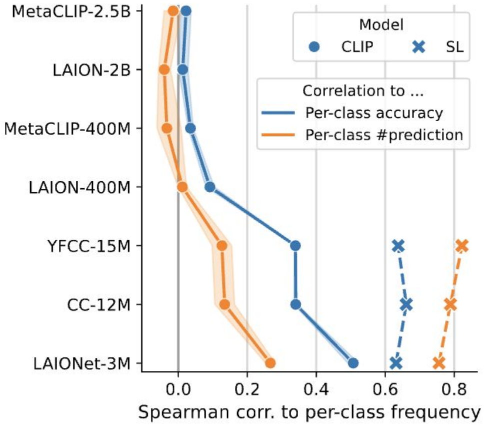


Figure 6: CLIP can benefit from open-world concepts. (a) Train on IN-Caps variants, and evaluate on 100 classes. (b) Train on YFCC-15M variants, and evaluate on 1K classes.



Wen et al., *What Makes CLIP More Robust to Long-Tailed Pre-Training Data? A Controlled Study for Transferable Insights,* NeurIPS 2024

Chapter 3:

Can we transfer these insights to other ML communities?

SL under extreme long-tail (or open-world recognition)

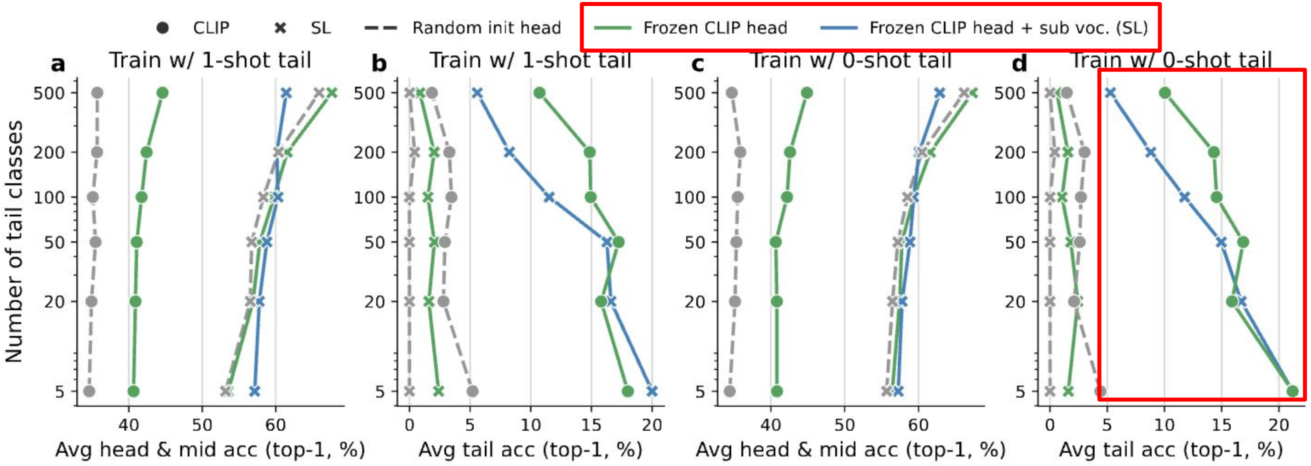


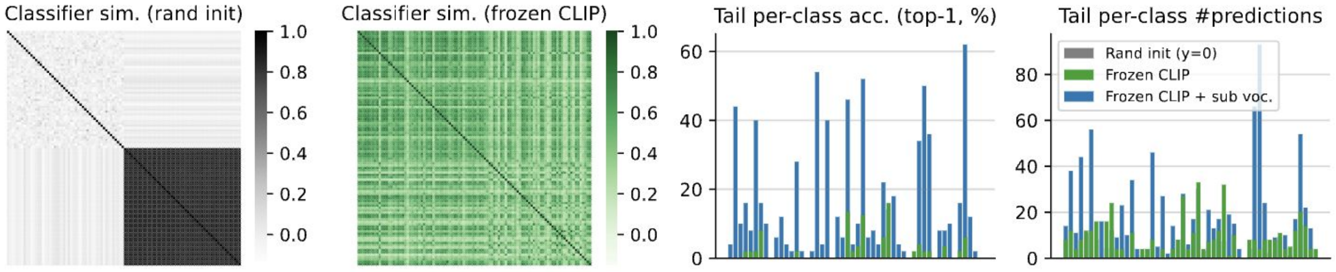
Figure 8: An extreme case: we train SL models on IN-Caps variants that have tail classes trimmed to only one shot (a & b) or even zero shot (c & d), and evaluate the accuracy on the tail and other classes. ● CLIP with a frozen pre-trained text encoder shows superior generalization, which can be acquired by a ✖ SL model with ✖ fixed class prototypes from CLIP and ✖ vocabulary subsampling.

LT and OW commonly use pre-trained CLIP **as is**.

We find voc. sub. is necessary to acquire CLIP knowledge in downstream tasks.

Wen et al., *What Makes CLIP More Robust to Long-Tailed Pre-Training Data? A Controlled Study for Transferable Insights,* NeurIPS 2024

SL under extreme long-tail (or open-world recognition)



(a) Affinity matrices of the classification head. (b) Distributions of models' per-class statistics.

Figure 9: A case study of SL under the zero-shot tail setting. (a) SL models seek maximal margins between classifiers, and tail prototypes collapse together. Instead, CLIP has a healthier structure. (b) Using CLIP head solely is less effective, and voc. subsampling is needed for CLIP-like generalization.

LT and OW commonly use pre-trained CLIP **as is.**

We find voc. sub. is necessary to acquire CLIP knowledge in downstream tasks.

CLIP head is good, but CLIP-like loss is also needed in downstream.

Wen et al., *What Makes CLIP More Robust to Long-Tailed Pre-Training Data? A Controlled Study for Transferable Insights,* NeurIPS 2024

SSL (DINO) on uncurated web data *vs* ImageNet

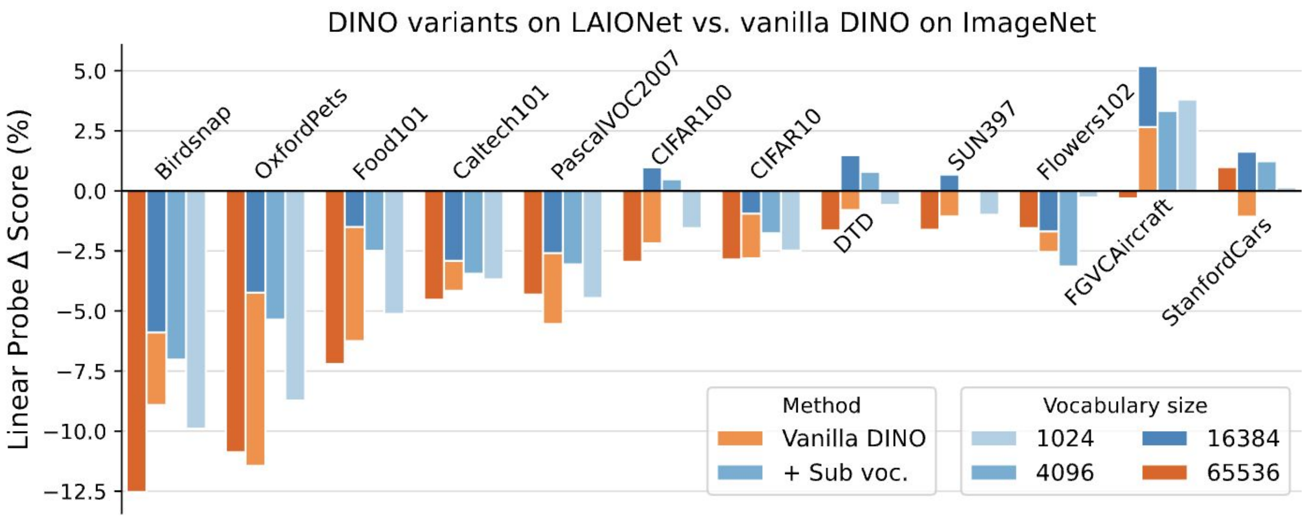
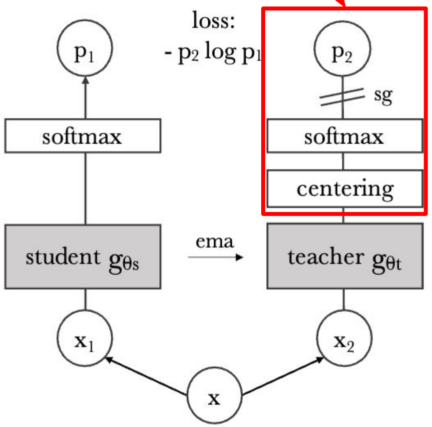


Figure 10: Transfer learning results of DINO variants pre-trained on LAIONet *vs*. vanilla DINO trained on ImageNet. Extreme data imbalance makes LAIONet much harder for DINO to learn transferrable representations. The ■ vocabulary subsampling strategy effectively helps ■ DINO alleviate such defects and generally match ImageNet-pretrained performance.

Pseudo labels here are high imbalanced!

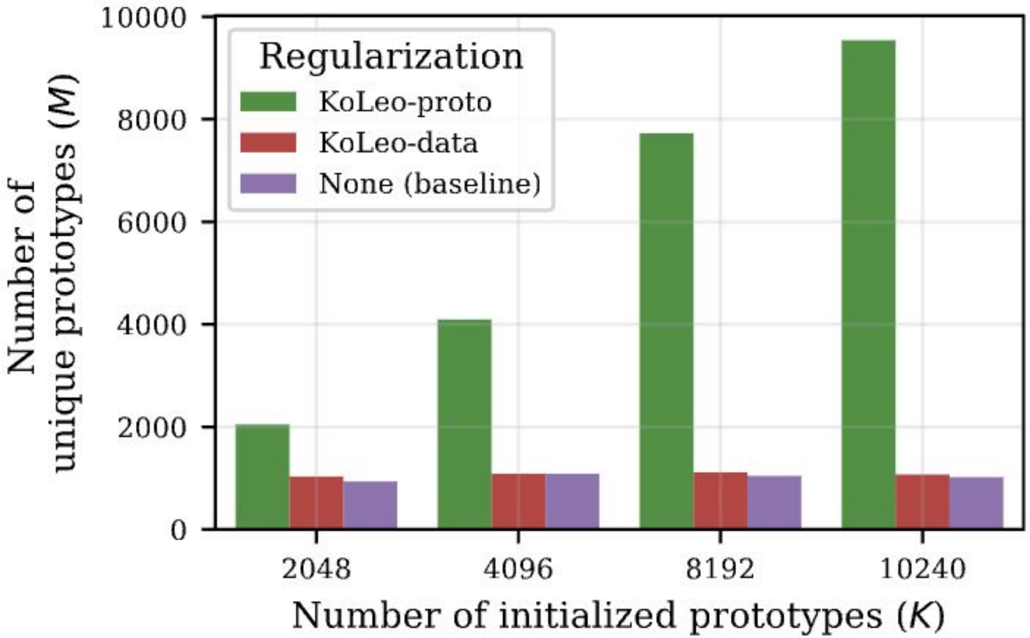


Caron et al., *Emerging properties in self-supervised vision transformers*, ICCV 2021
Wen et al., *What Makes CLIP More Robust to Long-Tailed Pre-Training Data? A Controlled Study for Transferable Insights,* NeurIPS 2024

65536 prototypes of DINO are far from good utilization

Table 1: Number of unique prototypes in existing models with $\epsilon = 0.025$ (default pre-training: ImageNet-1K, $*$: iNat-2018, $**$: ImageNet-22K)

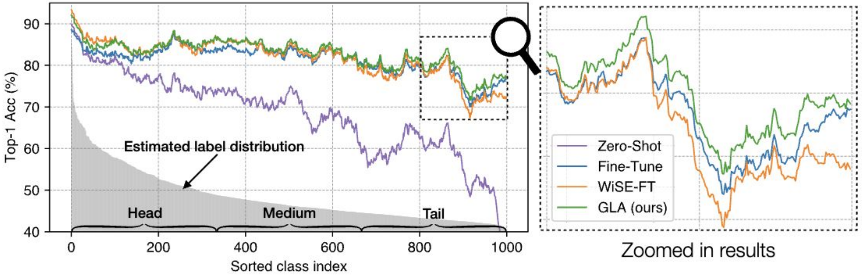
Backbone	Method	Initialized prototypes (K)	Unique prototypes (M)
ViT-S/16	DINO	65536	1078
ViT-B/16	DINO	65536	804
ViT-S/16	DINO-vMF	65536	1157
ViT-B/16	DINO-vMF	65536	939
ViT-S/16	iBOT	8192	3242
ViT-B/16	iBOT	8192	875
ViT-B/16	iBOT-vMF	8192	1170
ViT-L/16	iBOT	8192	969
ViT-B/16	iBOT**	8192	1241
ViT-L/16	iBOT**	8192	1037
ViT-S/16	MSN*	8142	3363
ViT-S/16	PMSN*	8142	3005



Govindarajan et al., *On Partial Prototype Collapse in the DINO Family of Self-Supervised Methods,* ICLR 2024 Submission

Chapter 4: Discussions

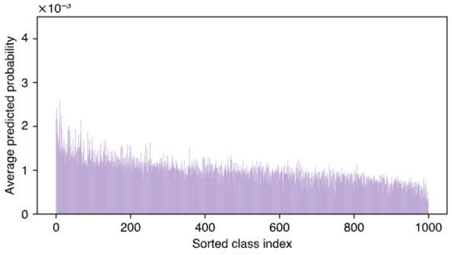
CLIP’s per-class accs are still biased, but weakly correlated to data distribution, and debiasing techniques in LT learning can be applied



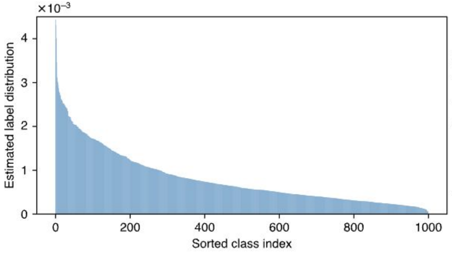
(a) Per class accuracy of different models on ImageNet

Model	Head	Medium	Tail	All
Zero-Shot	78.0	69.8	57.2	68.3
Fine-Tuned	83.6	83.0	77.3	81.3
WiSE-FT	**85.3** (+1.7)	83.7 (+0.7)	76.4 **(-0.9)**	81.8 (+0.5)
GLA (ours)	**85.2** (+1.6)	**84.3** (+1.3)	**78.8 (+1.5)**	**82.8** (+1.5)

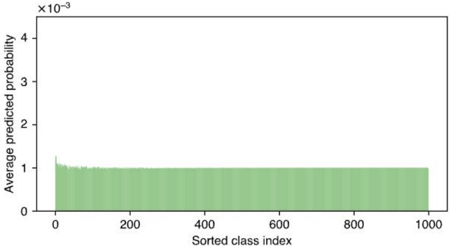
(b) Beak-down performance of different models on ImageNet



(a) Averaged outputs of zero-shot model



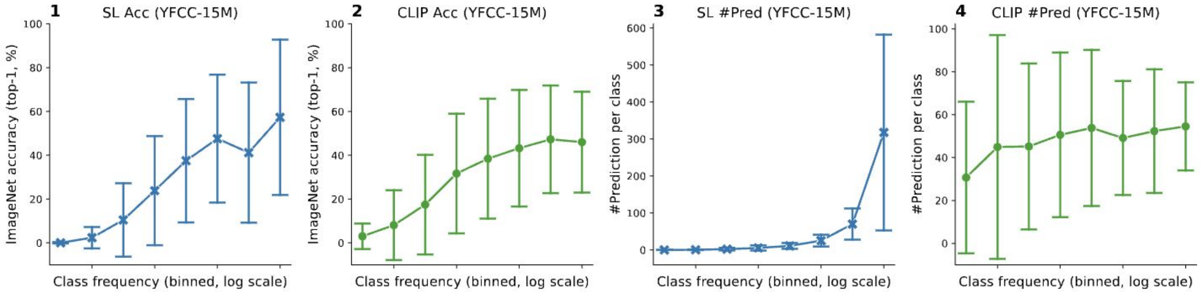
(b) Estimated label distribution $\mathbf{q}$



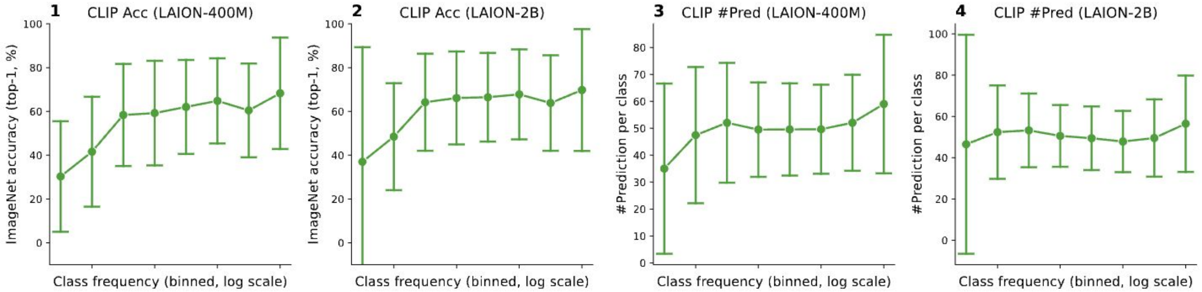
(c) Averaged outputs of debiased zero-shot model

Wang et al., Debiased learning from naturally imbalanced pseudo-labels., CVPR 2022
Zhu et al., Generalized logit adjustment: Calibrating fine-tuned models by removing label bias in foundation models, NeurIPS 2023

Looking at the overall trend, weak correlation can still be spotted; still, much better than SL



(a) Binned statistics of models pre-trained on YFCC-15M (avg ± std).



(b) Binned statistics of models pre-trained on LAION-400M and LAION-2B (avg ± std).

Wen et al., *What Makes CLIP More Robust to Long-Tailed Pre-Training Data? A Controlled Study for Transferable Insights,* NeurIPS 2024

Can we make a better estimation of concept frequency? There are works using LLM and/or VLM (GLIP)

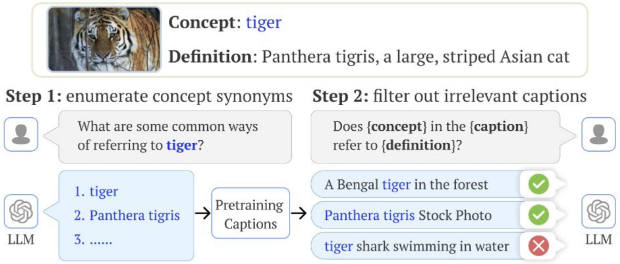
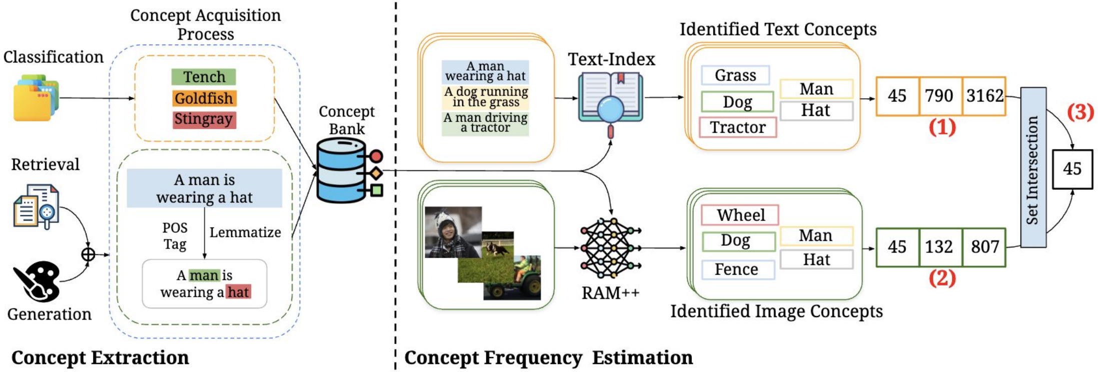


Figure 2. **Using large language models (LLMs) to estimate concept frequency in a VLM's pretraining dataset.** We conduct the frequency estimation using publicly available LAION [37] datasets. First, since a visual concept can be expressed in various ways, we ask an LLM (e.g., ChatGPT [31]) to enumerate all its synonyms to search for potentially relevant pretraining texts. For example, for `tiger`, we retrieve all captions that contain not only "`tiger`" but also its synonyms such as "`Panthera tigris`". Second, we filter out irrelevant captions that do not refer to the target concept by its definition. For example, although "*tiger shark swimming in water*" contains "*tiger*", it actually refers to a type of shark, not the animal `tiger` as defined by "Panthera tigris, a large, striped Asian cat". We conduct the filtering process by querying an LLM Llama-2 [42] (cf. Section 3).



Parashar et al., *The Neglected Tails of Vision-Language Models,* CVPR 2024
Udandarao et al., No "Zero-Shot" Without Exponential Data: Pretraining Concept Frequency Determines Multimodal Model Performance, arXiv 2024

Comparing to more advanced estimations, our results are mostly coherent

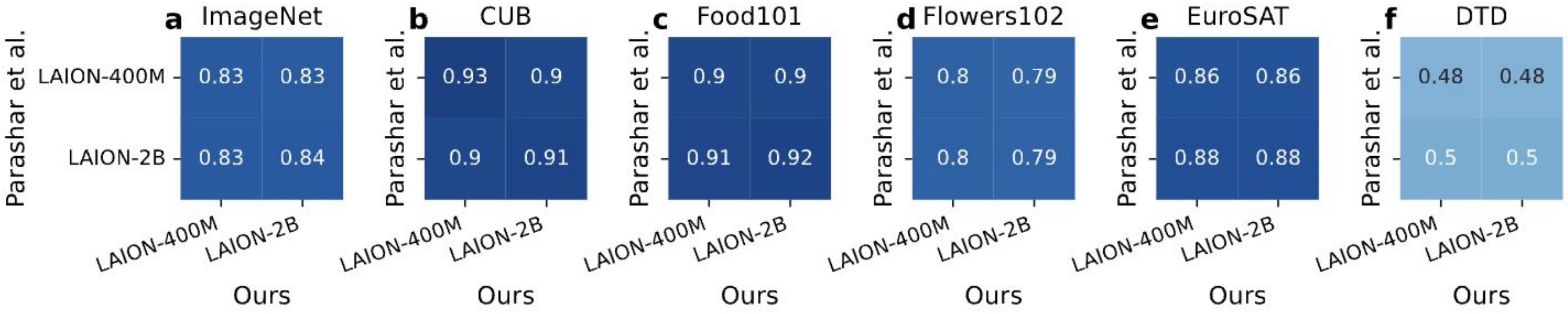


Figure 14: Correlation between class frequency statistics of our estimations and concurrent results of Parashar et al. [61]. There is an agreement on most concept sets except DTD [15], which is about descriptive textures and can be more semantically ambiguous [61].

Wen et al., *What Makes CLIP More Robust to Long-Tailed Pre-Training Data? A Controlled Study for Transferable Insights*, NeurIPS 2024

Beyond 1000 ImageNet classes, CLIP is still robust

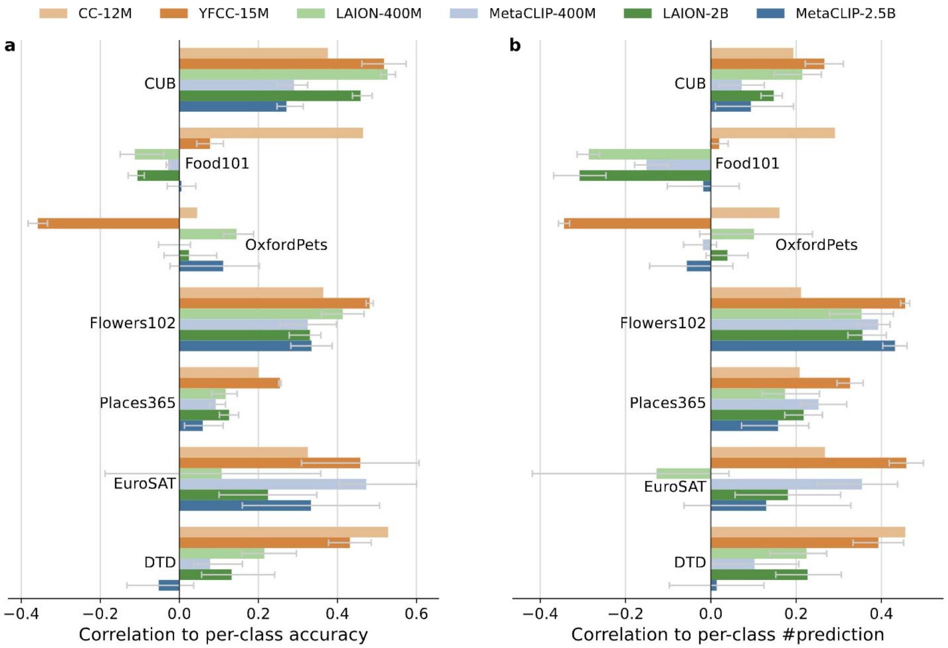


Figure 12: Correlation statistics of CLIP evaluated on broader sets of concepts. Models pre-trained at scale ($\geq$ 400M) remain robust on most datasets except fine-trained (*e.g.*, CUB and Flowers) and domain-specific ones (*e.g.*, EuroSAT). These data might be relatively rare on the web or have significant gaps with other data, thus hard to benefit from scaling or generalization from existing data.

Wen et al., *What Makes CLIP More Robust to Long-Tailed Pre-Training Data? A Controlled Study for Transferable Insights,* NeurIPS 2024

Why discussing CLIP in 2024?

FROMAGe, BLIP-2, MiniGPT-4, LLaVA, Instruct-BLIP still base on frozen CLIP.

We use the publicly available OPT model (Zhang et al., 2022) with 6.7B parameters as our LLM. Past work indicates that findings at the 6.7B scale are likely to generalize to larger model sizes (Dettmers et al., 2022), and large enough to exhibit the few-shot and in-context learning abilities that we are interested in (Radford et al., 2019). For the visual model, we use a pretrained CLIP ViT-L/14 model (Radford et al., 2021) for its ability to produce strong visual representations for vision-and-language tasks (Merullo et al., 2022).

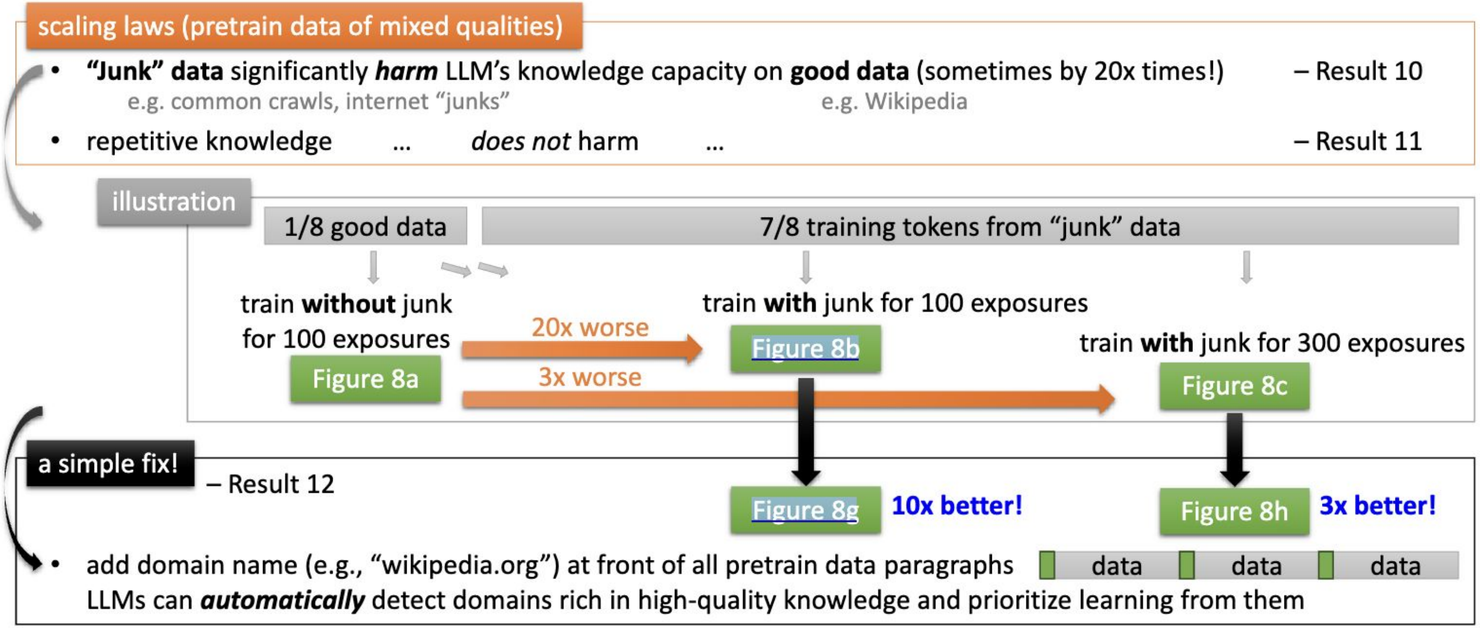
Pre-trained image encoder and LLM. For the frozen image encoder, we explore two state-of-the-art pre-trained vision transformer models: (1) ViT-L/14 from CLIP (Radford et al., 2021) and (2) ViT-G/14 from EVA-CLIP (Fang et al., 2022). We remove the last layer of the ViT and uses the second last layer's output features, which leads to slightly better performance. For the frozen language model, we explore the unsupervised-trained OPT model family (Zhang et al., 2022) for decoder-based LLMs, and the instruction-trained FlanT5 model family (Chung et al., 2022) for encoder-decoder-based LLMs.

To substantiate our hypothesis, we present a novel model named MiniGPT-4. It utilizes an advanced large language model (LLM), Vicuna [8], which is built upon LLaMA [32] and reported to achieve 90% of ChatGPT's quality as per GPT-4's evaluation, as the language decoder. In terms of visual perception, we employ the same pretrained vision component of BLIP-2 [16] that consists of a ViT-G/14 from EVA-CLIP [13] and a Q-Former. MiniGPT-4 adds a single projection layer to align the encoded visual features with the Vicuna language model and freezes all the other vision and language components. MiniGPT-4 is initially trained for 20k steps using a batch size of 256 on 4 A100 GPUs, leveraging a combined dataset that includes images from LAION [26], Conceptual Captions [5, 27], and SBU [20] to align visual features with the Vicuna language model. However, simply aligning the visual features with the LLM is insufficient to train high-performing model with visual conversation abilities like a chatbot, and the noises underlying the raw image-text pairs may result in incoherent language output. Therefore, we collect another 3,500 high-quality aligned image-text pairs to further fine-tune the model with a designed conversational template in order to improve the naturalness of the generated language and its usability.

Architecture. We implement InstructBLIP in LAVIS library [19]. Thanks to the flexibility enabled by the modular architectural design of BLIP-2, we can quickly adapt the model to incorporate various LLMs. In our experiments, we adopt four variations of BLIP-2 with the same image encoder (ViT-g/14 [10]) but different frozen LLMs, including FlanT5-XL (3B), FlanT5-XXL (11B), Vicuna-7B and Vicuna-13B. FlanT5 [7] is an instruction-tuned model based on the encoder-decoder Transformer T5 [33]. Vicuna [2], on the other hand, is a recently released decoder-only Transformer instruction-tuned from LLaMA [40]. During vision-language instruction tuning, we initialize the model from pre-trained BLIP-2 checkpoints, and only finetune the parameters of Q-Former while keeping both the image encoder and the LLM frozen. Since the original BLIP-2 models do not include Vicuna as LLMs, we perform pre-training with Vicuna following the same procedure as BLIP-2.

- *Large multimodal models*. We develop a large multimodal model (LMM), by connecting the open-set visual encoder of CLIP [36] with the language decoder LLaMA, and fine-tuning them end-to-end on our generated instructional vision-language data. Our empirical study validates the effectiveness of using generated data for LMM instruction-tuning, and suggests practical tips for building a general-purpose instruction-following visual agent. With GPT-4, we achieve state-of-the-art performance on the Science QA [30] multimodal reasoning dataset.

Performance of LLM can be easily poisoned by overwhelming "junk data"; their fix is allow the LLM to detect bad data



Allen-Zhu et al., *Physics of Language Models: Part 3.3, Knowledge Capacity Scaling Laws,* arXiv 2024

Performance of VLMs are also apparently biased by VL data; their fix is FT on balanced cls. data

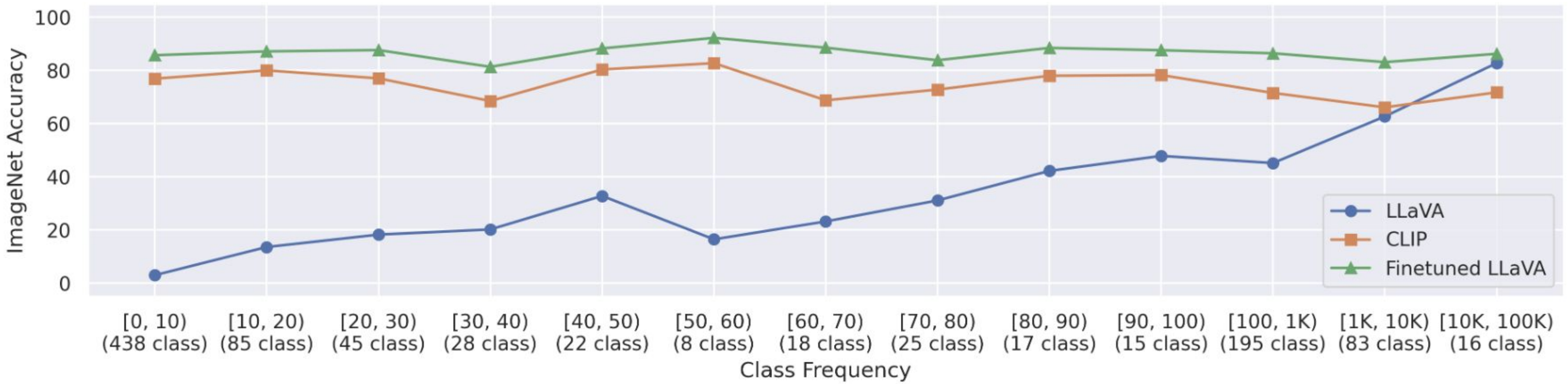


Figure 3: **Analysis of VLMs from the data perspective.** We study the relation between the ImageNet class frequency in the VLM training data and the VLM's classification performance on those classes. A strong correlation is observed, indicating that data determines VLM classification performance.

Zhang et al., *Why are Visually-Grounded Language Models Bad at Image Classification?*, arXiv 2024

Takeaways

- **Data matters**
  - Good data always helps
- **Data is not the silver bullet**
  - Re-balancing mechanisms of CLIP is one key factor of it to benefit from data scaling
- **We still do not fully understand contrastive models**
  - And they still can outperform generative models
- **Controlled experiment matters for a study**

Thanks!

Xin Wen, HKU
16 Aug, 2024