Does CLIP's generalization performance mainly stem from high train-test similarity?

Prasanna Mayilvahanan[1,2,3*] **Thaddäus Wiedemer**[1,2,3*] **Evgenia Rusak**[1,2,3]

Matthias Bethge[1,2] **Wieland Brendel**[2,3]

[1]University of Tübingen [2]Tübingen AI Center
[3]Max-Planck-Institute for Intelligent Systems, Tübingen

{prasanna.mayilvahanan, thaddaeus.wiedemer}@uni-tuebingen.de

Abstract

Foundation models like CLIP are trained on hundreds of millions of samples and effortlessly generalize to new tasks and inputs. Out of the box, CLIP shows stellar zero-shot and few-shot capabilities on a wide range of out-of-distribution (OOD) benchmarks, which prior works attribute mainly to today's large and comprehensive training dataset (like LAION). However, it is questionable how meaningful terms like out-of-distribution generalization are for CLIP as it seems likely that web-scale datasets like LAION simply contain many samples that are similar to common OOD benchmarks originally designed for ImageNet. To test this hypothesis, we retrain CLIP on pruned LAION splits that replicate ImageNet's train-test similarity with respect to common OOD benchmarks. While we observe a performance drop on some benchmarks, surprisingly, CLIP's overall performance remains high. This shows that high train-test similarity is insufficient to explain CLIP's performance.

1 Introduction

A core characteristic of *Foundation Models* [Bommasani et al., 2021] is that they are trained on hundreds of millions or even billions of data points scraped from the internet. For example, OpenCLIP [Schuhmann et al., 2022], the open-source version of CLIP [Radford et al., 2021], is trained on LAION-400M, a web-scale dataset with a wide variety of image-text pairs [Schuhmann et al., 2021]. CLIP forms the backbone of generative models like DALL-E2 [Ramesh et al., 2022] and is known for its remarkable "zero-shot" and "few-shot" performance on a wide range of tasks, specifically on "out-of-distribution" (OOD) benchmarks like ImageNet-Sketch [Wang et al., 2019], ImageNet-R [Hendrycks et al., 2020], etc.

CLIP's stellar OOD performance stems mainly from its data distribution [Fang et al., 2022, Radford et al., 2021]. Nevertheless, it remains unclear which specific properties of the training distribution, such as its scale, diversity, density, or relation to the test set, drive performance. OOD benchmarks like ImageNet-Sketch and ImageNet-R were originally designed in reference to ImageNet-1k [Deng et al., 2009], which had served as the primary dataset driving progress in machine vision for several years before the emergence of web-scale datasets. ImageNet-Sketch, ImageNet-R, and others are considered OOD because they share the same content (i.e., classes) as ImageNet-1k but are *dissimilar* in terms of style, pose, scale, background, or viewpoint. Clearly, there is no guarantee that these datasets are also *dissimilar* to LAION-400M. We provide evidence in Fig. 1 where we choose samples from ImageNet-Sketch and ImageNet-R and examine their nearest perceptual neighbors in LAION-400M and ImageNet-R. We find highly *similar* neighbors and even exact duplicates in

Workshop on Distribution Shifts, 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

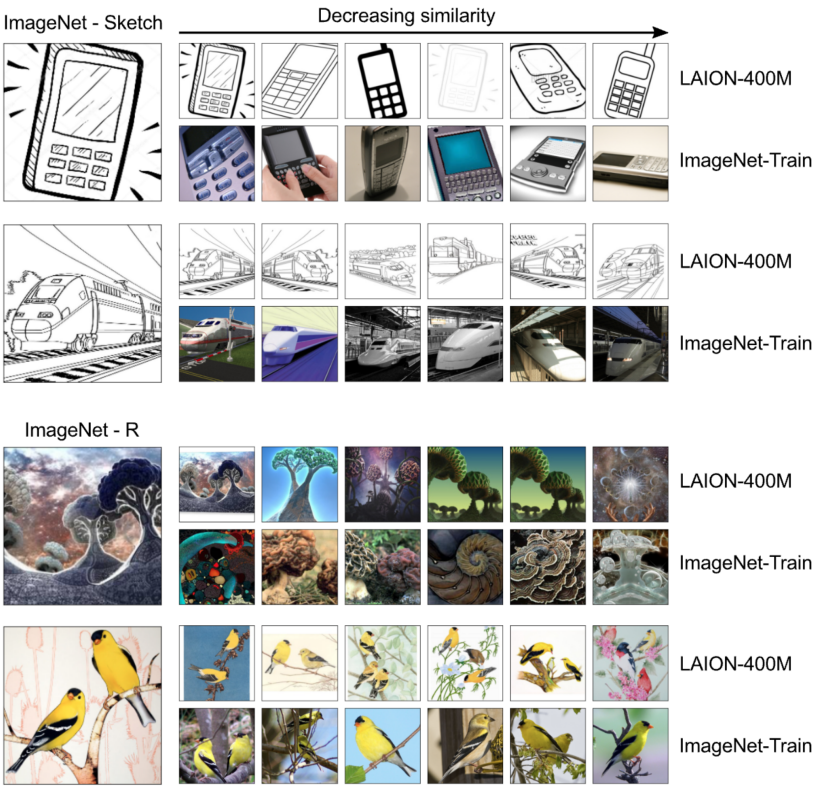

Figure 1: **Similarity of common benchmarks to LAION-400M and ImageNet-Train.** We show nearest neighbors of some ImageNet-Sketch, ImageNet-R and ImageNet-Val images in LAION-400M and ImageNet-Train ordered by decreasing *perceptual similarity*. We omit duplicates within these nearest neighbors. Perceptual similarity is computed in CLIP's image embedding space (see Sec. C) and can be thought of as measuring the "perceptual closeness" of images in terms of both content and style. LAION-400M clearly contains more similar images to samples from ImageNet-Sketch and ImageNet-R.

LAION-400M while neighbors in ImageNet-Train are relatively *dissimilar*. In other words, models trained on LAION-400M may perform well on conventional OOD benchmarks simply due to being trained on semantically and stylistically *similar* data points. Naturally, the question arises:

Does CLIP's accuracy on test sets mainly stem from highly similar images in its train set?

The goal of our work is to resolve this question. In order to assess the impact of *highly similar images* on CLIP's generalization performance, we introduce the notion of a *similarity gap* based on nearest neighbor distance (Sec. 3). Under this formalization, *highly similar images* of LAION-400M lie within the similarity gap of ImageNet-Train, *i.e.*, are more similar to test samples than any image in ImageNet-Train is. We then create a subset of LAION-400M for each OOD test set, ensuring that their similarity gap is consistent with that of ImageNet-Train (Sec. 4). The OOD test sets are as out-of-distribution to these curate subsets of LAION-400M as they are to ImageNet-Train. As our central result, we surprisingly find that training CLIP models on those subsets leads only to marginal losses on the corresponding OOD benchmarks (Tab. 1). We hence conclude that high train-test similarity cannot fully explain CLIP's remarkable OOD performance, and other properties of LAION-400M must play a role.

2 Abridged Related Work

ID vs. OOD generalization Large-scale language-image models such as CLIP [Radford et al., 2021], ALIGN [Jia et al., 2021], or BASIC [Pham et al., 2021] claim exceptional OOD generalization

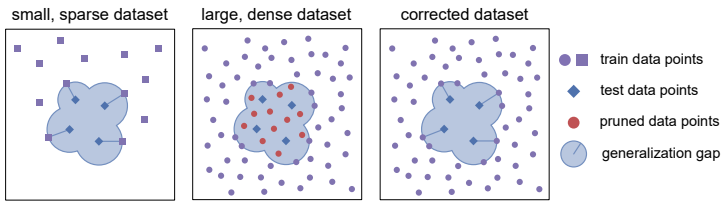

Figure 2: **Aligning the similarity gap of two datasets**. A larger, denser, more diverse dataset likely contains samples closer to given test points than a smaller, sparser one. To correct for this, we compute the nearest neighbor distance of each test point to the smaller dataset (left) and prune points from the larger dataset that lie within this hull (center).

and zero-shot capabilities. Fang et al. [2022] probe which aspects of the models—like language supervision, cost function, or training distribution—are related to a model's effective OOD robustness and find that differences in the distribution play a key role. Here, we aim to extend the findings of Fang et al. [2022] by evaluating whether high similarity between training and test set is the main driver of CLIP's claimed OOD performance, or whether CLIP is truly better at generalizing across larger distribution shifts.

3 Similarity gap

Our goal is to prune LAION in such a way, that the pruned dataset has the exact same perceptual distances between training and test samples as ImageNet-Train. We formalize this notion as *similarity gap* below and explain how to use it as a systematic measure to study generalization performance more fairly, even in the age of massive web-scale training datasets.

Let us for consider two training datasets, denoted as $\mathcal{D}_S$ (small, like ImageNet-Train) and $\mathcal{D}_L$ (large, like LAION-400M), along with a test dataset $\mathcal{T}$ (like ImageNet-Sketch). For the sake of simplicity, we assume that $\mathcal{D}_S$ is a subset of $\mathcal{D}_L$. We choose a distance measure $d(\cdot,\cdot) : \mathcal{T} \times \mathcal{D} \to \mathbb{R}^+$ in order to find the k nearest neighbors in the training set for each sample $t_i \in \mathcal{T}$. Although we choose $k = 1$ for simplicity, the approach can be generalized straightforwardly. We denote the nearest neighbor distance between t_i and a dataset $\mathcal{D}$ as $d_i(\mathcal{D})$, which is given by

$$d_i(\mathcal{D}) = \min_{x \in \mathcal{D}} d(t_i, x). \tag{1}$$

The set of nearest neighbor distances

$$G(\mathcal{D}, \mathcal{T}) = \left\{ d_i(\mathcal{D}) \mid i \in \left[1, |\mathcal{T}|\right] \right\} \tag{2}$$

then quantifies the *similarity gap* G between training and test set. We can think of this measure as a full characterization of the training set's similarity to any point in the test set; compare Fig. 2.

Based on the assumption that the large dataset contains all samples from the small dataset, it follows that $d_i(\mathcal{D}_S) \geq d_i(\mathcal{D}_L)$. In other words, the nearest neighbor distance to samples in the small training set is always larger than or equal to the distance to samples in the large training set. Consequently, the similarity gap $G(\mathcal{D}_L, \mathcal{T})$ between the test set and the large training set is strictly smaller than the similarity gap $G(\mathcal{D}_S, \mathcal{T})$ between the test set and the small training set (on a sample-by-sample basis). We aim to identify a maximally large subset $\tilde{\mathcal{D}}_L \subseteq \mathcal{D}_L$ of the large training set, such that it's similarity gap $G(\tilde{\mathcal{D}}_L, \mathcal{T}) = G(\mathcal{D}_S, \mathcal{T})$ is equal to the small dataset (on a sample-by-sample basis, meaning $d_i(\tilde{\mathcal{D}}_S) = d_i(\mathcal{D}_S)$ for all samples). To achieve this, we examine each test sample t_i and remove any sample $x \in \mathcal{D}_L$ for which the distance $d(t_i, x) < d_i(\mathcal{D}_S)$ is smaller. We illustrate this procedure in Fig. 2. Sec. 4 applies this framework to LAION-400M and ImageNet-Train. We compute perceptual distances in the embedding space of pretrained CLIP ViT-B/16+. This ensures that remove all data points in LAION that are semantically and stylistically more similar to the test data points than samples in ImageNet-Train. In this way, we obtain a version of LAION-400M that is equivalently far removed from the test sets as ImageNet-Train.

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

Table 1: **Corrected zero-shot performance of CLIP ViT-B/32.** 'X-pruned' represents a pruned dataset from LAION-200M + ImageNet such that the similarity gap to 'X' is the same as the similarity gap of ImageNet to 'X'. The sizes of these subsets are subtracted from the LAION-200M + ImageNet's size. Here, 'X' is one of the 6 standard ImageNet test sets. 'combined-pruned' splits ensure a similarity gap of LAION-200M and ImageNet-Train to all 6 test sets. CLIP's corrected zero-shot performance drops the most on ImageNet-Sketch and ImageNet-R with a relative performance drop of 10.8 % and 4.8 % respectively. Red color indicates a drop in performance on the respective test set, and blue represents a rise. Overall high performance indicates that highly similar images do not play a key role in explaining CLIP's generalization ability.

4 Correcting for highly similar images

As described in Sec. 3, we first compute the similarity gaps of the smaller dataset, *i.e.*, ImageNet-Train, to the samples in each of the six test sets. We work LAION-200M, which is a deduplicated version of LAION-400M with 200M samples Abbas et al. [2023]. To ensure that ImageNet-Train and LAION-200M have the same similarity gap to the test sets, we include all ImageNet-Train images in LAION-200M with the caption `"a photo of a {object class}"` [Radford et al., 2021]. Pruning LAION-200M to these similarity gaps leaves us with six different base splits as shown in Tab. 1. We also generate a 'combined-pruned' split that ensures an ImageNet-Train-like similarity gap to all test sets at the same time. We can now train CLIP from scratch on these splits to obtain a corrected zero-shot performance and compare it to the accuracy of CLIP trained by OpenAI and OpenClip [Ilharco et al., 2021, Radford et al., 2021].

The first important point to note in Tab. 1 is that for 'sketch-pruned' and 'r-pruned' datasets, we prune 8.3M and 5.7M samples, respectively. For all other datasets, we prune only around 250K-380K samples. The number of pruned samples also also highly related with the respective accuracies. For CLIP trained on the 'r-pruned' dataset and CLIP trained on the 'sketch-pruned' dataset, we observe a $4.8\,\%$ relative performance decrease on ImageNet-R and $10.8\,\%$ relative performance decrease on ImageNet-Sketch compared to the baseline. There is also a considerable performance change on ImageNet-R for 'sketch-pruned' and on ImageNet-Sketch for 'r-pruned'. This is reasonable as there is some style overlap in ImageNet-Sketch and ImageNet-R. For the other four base splits, we see less than $1\,\%$ relative performance change on all six evaluation sets. The performance of the CLIP model trained on the 'combined-pruned' split is lower than the baseline on all six eval sets, with sizeable drops in ImageNet-R and ImageNet-Sketch. We also observe similar trends when we do not add ImageNet-Train to the pruned datasets (refer to Tab. 4 in the Appx.).

5 Discussion

We now return to our original question: *Does CLIP's accuracy on OOD benchmarks mainly stem from highly similar images in its train set?* To give a definitive answer, we take a closer look at the CLIP model trained on 'sketch-pruned'. This model's training set is as dissimilar to ImageNet-Sketch as is ImageNet-Train. It features an accuracy of $68.34\,\%$ on ImageNet-Val. According to ImageNet-Train's

effective robustness line [Fang et al., 2022], at this performance level, we would expect an accuracy of roughly $14\,\%$ on ImageNet-Sketch. Instead, we find an accuracy of $44.78\,\%$. In other words, training on a much larger dataset while keeping the similarity gap constant drastically increases generalization performance for CLIP (in this case, by a staggering 30 percentage points). This effect is even higher for other datasets. *This indicates that CLIP's impressive performance is not so much the result of a high train-test similarity but that CLIP leverages its dataset scale and diversity to learn more generalizable features.*

What drives generalization? Generalization of vision-language models is a complex subject where several factors like architectural choices, caption quality, training procedures, and data distribution play a role. We focus on the training distribution since prior works have studied the effect of the aforementioned factors on CLIP's generalization performance [*e.g.*, Santurkar et al., 2022, Mintun et al., 2021] and identified it as a prominent factor [Fang et al., 2022]. Many distribution properties could contribute to generalization performance, but based on raw visualizations of the involved datasets, highly similar images are clearly *a* factor. Our results only show that it is not the most important factor and a large chunk of performance remains to be explained. We leave the scrutiny of other likely factors like data diversity and density for future work. Our work should be interpreted as a step towards finding specific properties of data that dictate generalization, a topic that will remain relevant in the foreseeable future.

Similarity metric To our knowledge, *perceptual similarity* as measured in CLIP ViT-B/16+'s image embedding space is a leading metric to capture the semantic and stylistic similarity between images. While we found this metric to align well with our intuitive notion of similarity and believe it to have captured the vast majority of highly similar images (see also Appx. E.2 where we visualize the pruned datasets), we cannot guarantee that all highly similar images were removed. While we believe, based on prior work [Fu et al., 2023, Abbas et al., 2023, Gadre et al., 2023, Zhang et al., 2021], that CLIP's embedding space sufficiently captures relevant features and do not expect to see a drastic change in the trends we observed, our findings might be refined in the future with more precise notions of image similarity.

Highly similar images We would like to further clarify the notion of *highly similar images*. In Secs. C.1 and C.4, when we use the notion of *highly similar images* to a given image sample, we refer to images with high perceptual similarity values with no precise constraint. In contrast, in Secs. 3 and 4 we impose a constraint that defines *highly similar images* to a sample as images that are closer to LAION-200M than ImageNet-Train based on our perceptual similarity metric.

Coreset In Sec. C.4, we identify a 100M coreset of LAION-400M, which, when trained on, leads to a CLIP model that nearly matches the performance of a LAION-400M trained CLIP model on the six test datasets. We list details and further performance comparisons in Appx. D.2. We do not suggest this coreset as an alternative to other pruning or deduplication methods that are largely agnostic to the downstream test datasets. Instead, we here created a coreset that is specifically designed to perform well on six OOD test sets to facilitate further research into what aspects drive generalization.

6 Conclusion

CLIP has demonstrated unprecedented performance on common OOD benchmarks designed originally for ImageNet. Given that the training dataset of CLIP is so large and diverse, it is natural to wonder whether its performance stems from the sheer similarity of many training samples to the benchmarks. To the best of our knowledge, we are the first to systematically account for train-test similarity in order to more fairly assess generalization performance in the era of foundation models. In our work, we address this by pruning away samples from the training set that are more similar to the test sets than ImageNet samples. Models trained on the pruned dataset do not significantly lose performance and still exhibit stellar generalization capabilities far beyond performance-matched ImageNet-trained models. This indicates that high similarity to the test sets alone can not explain CLIP's generalization ability. We hope that this result will prompt the community to investigate other factors that allow models to learn more generalizable features from web-scale datasets.

References

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=M3Y74vmsMcY.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. *CoRR*, abs/2006.16241, 2020.

Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data determines distributional robustness in contrastive language image pre-training (clip). In *International Conference on Machine Learning*, pages 6216–6234. PMLR, 2022.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al. Combined scaling for zero-shot transfer learning. *arXiv preprint arXiv:2111.10050*, 2021.

Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*, 2023.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL https://doi.org/10.5281/zenodo.5143773. If you use this software, please cite it as below.

Shibani Santurkar, Yann Dubois, Rohan Taori, Percy Liang, and Tatsunori Hashimoto. Is a caption worth a thousand images? a controlled study for representation learning. *arXiv preprint arXiv:2207.07635*, 2022.

Eric Mintun, Alexander Kirillov, and Saining Xie. On interaction between augmentations and corruptions in natural corruption robustness. *Advances in Neural Information Processing Systems*, 34:3571–3583, 2021.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023.

Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019.

Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. Quality not quantity: On the interaction between dataset design and robustness of clip. In *Advances in Neural Information Processing Systems*, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

A Experimental details

For computing image-to-image similarity, measuring duplicates, and pruning data points, we use CLIP ViT-B/16+'s image embedding space. For all our pruning experiments, we train CLIP ViT-B/32 [Dosovitskiy et al., 2020] for 32 epochs with a batch size of 33,600 on one node with eight A100 GPUs (training takes several days, depending on the dataset size). We use the implementation provided by Ilharco et al. [2021] and stick to their settings for learning rate, weight decay, etc. Our downloaded version of LAION-400M contains only 377M images overall due to missing or broken links, compared to the original 400M used in OpenCLIP [Ilharco et al., 2021]. Abbas et al. [2023] show that pruning exact duplicates, near duplicates, and semantically very similar samples *within* LAION-400M (not yet taking any test sets into account) can reduce dataset size by up to $50\,\%$ without performance degradation. We reimplement their method to generate our baseline LAION split containing 199M samples, which we refer to as LAION-200M. This step is important to make training multiple instances of CLIP feasible, and we observe that the incurred drop in performance is negligible (compare Tab. 1). The test datasets used in our experiments and analyses are ImageNet-Val, ImageNet-Sketch, ImageNet-V2, ImageNet-R, ImageNet-A, and ObjectNet.

B Related Work

Measuring OOD generalization To assess expected model performance in the wild, researchers use different test sets that are considered OOD with respect to the training distribution. The terms OOD generalization, (distributional) robustness, or just generalization are used interchangeably by the community. This work mainly focuses on standard datasets that share classes with ImageNet. They include: image renditions (ImageNet-R; Hendrycks et al., 2020), unusual camera views and object positions (ObjectNet; Barbu et al., 2019), images selected to be difficult for ImageNet-trained ResNet-50s (ImageNet-A; Hendrycks et al., 2021) and sketches of ImageNet classes (ImageNet-Sketch Wang et al., 2019). We also consider two datasets commonly considered in-distribution, namely ImageNet-Val, which is just the test dataset of ImageNet [Deng et al., 2009], and a similarly sampled but newer test set for ImageNet (ImageNet-V2 Recht et al., 2019).

ID vs. OOD generalization While researchers treat the test sets listed above as OOD with respect to the training distribution when they study robustness, this core assumption is rarely scrutinized. Mintun et al. [2021] show for CIFAR10-C and ImageNet-C that the perceptual similarity between augmentations used during training and a target corruption is highly predictive of the final model performance on the target corruption. Large-scale language-image models such as CLIP [Radford et al., 2021], ALIGN [Jia et al., 2021], or BASIC [Pham et al., 2021] claim exceptional OOD generalization and zero-shot capabilities. Fang et al. [2022] probe which aspects of the models—like language supervision, cost function, or training distribution—are related to a model's effective OOD robustness and find that differences in the distribution play a key role. Further, Nguyen et al. [2022] find that combining data from multiple sources for training interpolates the model's effective robustness on an OOD test set between the performance of the model trained on either data source. In DINOv2 [Oquab et al., 2023], the authors curate a large web-scale dataset by keeping images similar to query images from a variety of different datasets, thereby aiming to build a diverse training set. Here, we aim to extend the findings of Mintun et al. [2021], Fang et al. [2022], and Nguyen et al. [2022] by evaluating whether high similarity between training and test set is the main driver of CLIP's claimed OOD performance, or whether CLIP is truly better at generalizing across larger distribution shifts. We propose a principled methodology to account for this confounder and evaluate OOD performance of CLIP more fairly.

C The similarity hypothesis

In this section, we elucidate our line of reasoning to arrive at the hypothesis that CLIP's performance may stem from highly similar images.

C.1 Comparing LAION and ImageNet's test-set similarity

The popular ImageNet-1k and LAION-400M datasets differ not only in scale but also in their diversity. ImageNet-1k consists of 1.2 million samples, the vast majority of which are natural images.

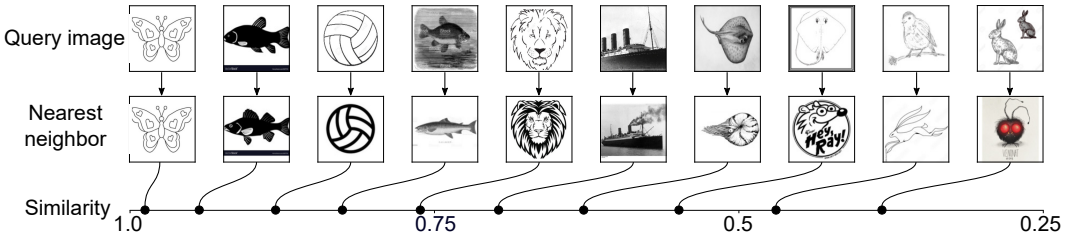

Figure 3: **Relation between *perceptual similarity* and visual closeness of nearest neighbors**. Query images are sampled from ImageNet-Sketch (top row) and are connected to their nearest neighbor in LAION-400M (bottom row). As in Fig. 1, perceptual similarity is measured in CLIP's image embedding space.

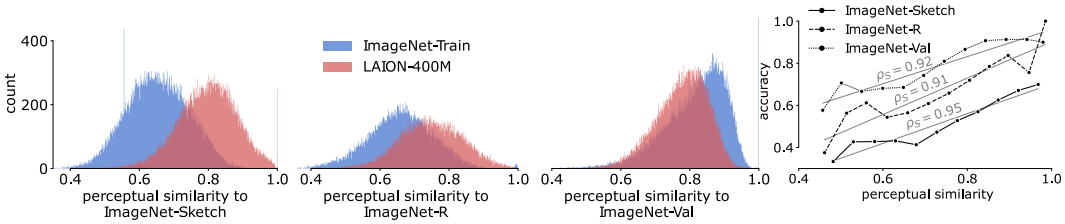

Figure 4: **Similarity of nearest neighbors to test sets varies between LAION-400M and ImageNet-Train and is correlated with performance**. **Left**: The histograms display the similarity of samples in ImageNet-Sketch (left), ImageNet-R (center), and ImageNet-Val (right) to their nearest neighbors in LAION-400M (red) and ImageNet-Train (blue). LAION-400M is overall more similar to ImageNet-Sketch and ImageNet-R, while ImageNet-Train is more similar to ImageNet-Val. **Right**: The strong correlation between perceptual similarity of test points to nearest neighbors in LAION-400M samples and CLIP's top-1 classification accuracy indicates that differences in similarity can be expected to drastically impact the performance of LAION-trained models. Data points in the correlation plot are averaged over bins (interval = 0.05) of the red histograms in the left plot.

In comparison, LAION-400M comprises roughly 400 million samples scraped from the Common Crawl dataset[1]. The Common Crawl dataset comprises petabytes of data collected over more than a decade of web crawling; LAION-400M, therefore, contains an extremely high diversity and density of images.

The most common datasets used for testing OOD generalization performance are ImageNet-Sketch, ImageNet-R, ImageNet-V2, ImageNet-A, and ObjectNet. To quantify similarity of samples from the aforementioned datasets to data points in ImageNet-1k (for which they were designed) and LAION-400M, we here use a common *perceptual similarity* metric based on the cosine similarity in the image embedding space of a LAION-400M-trained CLIP ViT-B/16+ model [Ilharco et al., 2021]. Abbas et al. [2023] demonstrated that nearest neighbors in the image embedding space share *semantic* and *stylistic* characteristics. We illustrate this in Fig. 3, where we plot samples from ImageNet-Sketch and their neighbors in LAION-400M for different similarity values. Visually, the similarity scores correlate well with the semantic and stylistic closeness of the image pairs. This is corroborated by other works that demonstrate high perceptual alignment between CLIP's embedding distance and human perception [Fu et al., 2023] and use it to sample ImageNet-like data points from a large dataset [Gadre et al., 2023] or to build similarity-based classifier for few-shot classification [Zhang et al., 2021]. We consequently use it as the main similarity measure throughout this paper. We interchangeably use the terms *similarity* and *perceptual similarity* to refer to this metric. Likewise, *nearest neighbors* or *similar images to a query image* denotes the closest images in the CLIP image-embedding space. In Fig. 1 and Appx. E, we use perceptual similarity to illustrate that often nearest neighbors in LAION-400M are semantically and stylistically more similar to random samples from the OOD datasets than nearest neighbors from ImageNet-Train are. Specifically, for ImageNet-Sketch and ImageNet-R images, it is evident that neighbors in LAION-400M share both semantics and style, whereas neighbors in ImageNet-Train are merely natural images of the same class. In contrast,

[1] https://commoncrawl.org/

for ImageNet-Val, nearest neighbors in ImageNet-Train are more similar than the neighbors in LAION-400M. To quantify this observation, we measure the distribution of perceptual similarities of all ImageNet-Sketch samples to their nearest neighbors in LAION-400M and ImageNet-Train (Fig. 4, left). While ImageNet-Train's similarity distribution has a mode of roughly 0.65 and is fairly narrow, LAION's similarity distribution peaks at around 0.8 and is much broader. The same trend can be observed for ImageNet-R. In contrast, the distribution of nearest neighbor similarities to ImageNet-Val (Fig. 4, right) indicates that samples are generally closer to ImageNet-Train than to the LAION-400M, but the difference in modes is smaller. We refer the reader to Appx. E and C.2 for nearest neighbor and perceptual similarity histograms for ImageNet-V2, ImageNet-A, and ObjectNet. For these datasets, the similarity distributions of LAION-400M and ImageNet-Train barely differ.

Moreover, in Appx. C.3, we detail the number of *near duplicates* (duplicates up to small shifts or crops) of the OOD datasets in LAION-400M and ImageNet-Train. While we found $3.1\,\%$ of ImageNet-Sketch images to have duplicates in LAION-400M, there are only $0.04\,\%$ ImageNet-Sketch duplicates in ImageNet-Train. On the other hand, ImageNet-Train contains duplicates of $2.67\,\%$ ImageNet-Val images as opposed to just $0.14\,\%$ ImageNet-Val images in LAION-400M.

The rightmost plot of Fig. 4 illustrates that similarity between test samples and their nearest neighbors is a good predictor for a model's performance on these samples. Each data point here corresponds to a bin from one of the red histograms to the left. We report CLIP's average top-1 classification accuracy over all samples in the same bin. We observe a clear correlation between similarity and classification performance, ranging from $35\,\%$ accuracy for sketches that have no similar counterparts in LAION-400M (similarity 0.38) up to $69\,\%$ accuracy for sketches that are duplicated in LAION-400M (similarity close to 1).

Based on these observations, we expect that for ImageNet-Sketch and ImageNet-R, LAION-400M will contain a large number of highly similar images. Because of the high correlation between train-test similarity and accuracy, it is a natural hypothesis that this similarity largely explains CLIP's high performance.

C.2 Nearest neighbor similarity between LAION / ImageNet-Train and other OOD datasets

As an extension of our analysis in Sec.C, we plot the nearest neighbor similarity between ImageNet-Train / LAION-400M and other OOD test sets, namely ImageNet-A [Hendrycks et al., 2021], ObjectNet [Barbu et al., 2019] and ImageNet-V2 [Recht et al., 2019], and display our results in Figure 5. There are no significant differences in nearest neighbor similarity for all of these test sets. Similar to our results in Figure 4 (right), we find a strong correlation between perceptual similarity to LAION-400M and the top-1 accuracy of our LAION-trained model.

C.3 Duplicates

We do a duplicate analysis in Tab. 2. To estimate the number of test points with *near duplicates*, we project the test set and LAION-400M to CLIP's image embedding space and check if any point in LAION-400M lies in the vicinity ($\epsilon = 0.05$) of each of the query test point.

Table 2: **Number of test points of OOD datasets for which we find *near duplicates* in ImageNet-Train and LAION-400M**. A data point is considered *near duplicate (semantic duplicate)* if the distance in the CLIP embedding space is less than 0.05 [Abbas et al., 2023].

		Duplicates	
Dataset	**Size**	**ImageNet-Train**	**LAION-400M**
ImageNet-Val	50000	1336	70
ImageNet-Sketch	50889	18	1553
ImageNet-R	30000	104	297
ImageNet-A	7500	10	5
ImageNet-V2	10000	10	24
ObjectNet	18574	0	0

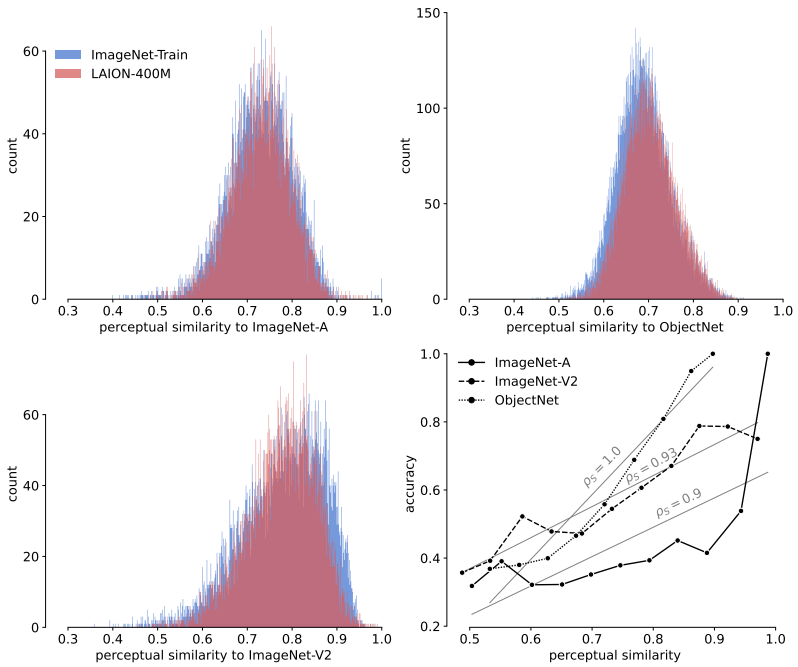

Figure 5: **Similarity of nearest neighbors to test sets varies between LAION-400M and ImageNet-Train and is correlated with performance.** Histograms over the nearest neighbor similarity of test sets ImageNet-A (top left), ObjectNet (top right) and ImageNet-V2 (bottom left) to training sets LAION-400M (red) and ImageNet-Train (blue). There are no significant differences in nearest neighbor similarity for all of test sets. The strong correlation between nearest neighbor similarity of test points to LAION-400M samples and top-1 classification accuracy (bottom right) indicates that this can be expected to drastically boost the performance of LAION-trained models. data points in the correlation plot are averaged over bins (interval = 0.05) of the red histograms.

C.4 Impact of pruning (dis)similar images

In the last subsection, we found that a sample's similarity to its nearest neighbors in the training set is highly predictive of a model's performance on this sample. We can observe this effect even more directly by pruning training samples from LAION that are highly *similar* to test data points. We then train CLIP on the pruned dataset and evaluate its performance on the test set to see how much CLIP's performance is influenced by train-test set similarity. To this end, we order all samples in LAION by their distance to the perceptually most similar test sample. We then prune those samples in LAION that are most similar to the test set. To ease analysis, we perform this analysis on LAION-200M (see Sec. A). We prune the dataset to 150M, 100M, and 50M (collectively called 'near-pruned' splits), making the splits progressively more dissimilar to the test sets.

We also examine the effect of removing points from LAION-200M that are highly *dissimilar* to the test data points. We refer to the resulting datasets as 'far-pruned'. As baseline, we also prune random points to generate 'rand-pruned' data splits. We separately use ImageNet-Sketch and ImageNet-Val as test datasets to generate 'near-pruned' and 'far-pruned' splits.

We train CLIP on 15 different subsets of LAION-200M which differ in their pruning strategy ('near-pruned', 'far-pruned', or 'rand-pruned'), the target test set (ImageNet-Val or ImageNet-Sketch) and their pruning strength (150M, 100M, and 50M). We plot the resulting zero-shot accuracy on the target test sets in Fig. 6. We find that the 'near-pruned' accuracy curve drops quickly with decreasing dataset size. This reiterates that a large chunk of CLIP's performance directly stems from similar images to the test set that are removed here. In contrast, the performance of models trained on the 'far-pruned' datasets remains stable up until a dataset size of 100M. In fact, the performance even slightly surpasses that of the baseline model, further indicating that dissimilar samples do not contribute to CLIP's performance and instead act more like noise in the training data. We extend our analysis to other test sets in Appx. D.1 (Fig. 7).

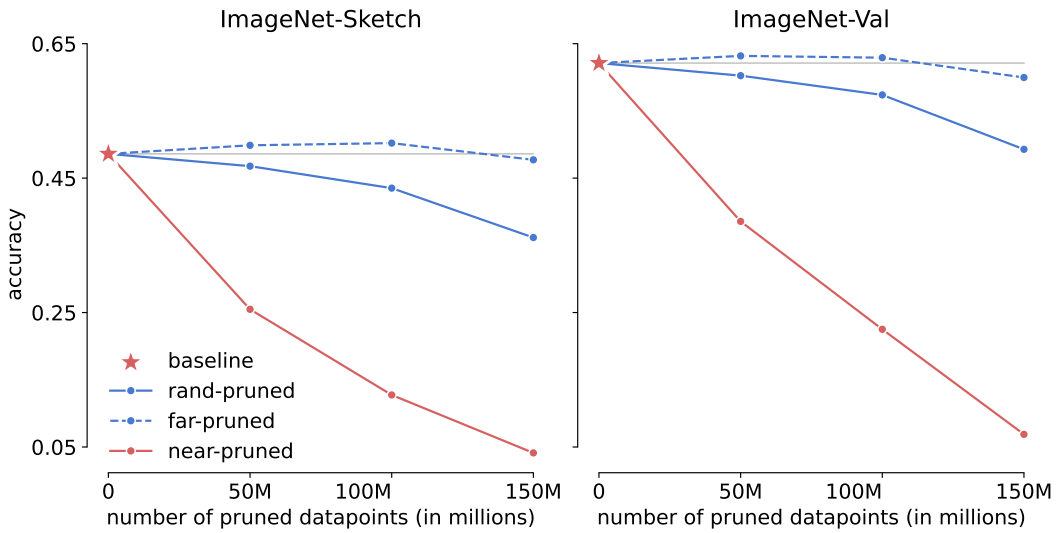

Figure 6: **Effect of pruning similar and dissimilar points to a given test set**. The baseline model is trained on de-duplicated LAION-400M, which we call LAION-200M. To generate the 'near-pruned' datasets, we remove images in decreasing order of similarity to ImageNet-Sketch or ImageNet-Val (based on CLIP image-embedding similarity). In contrast, the 'far-pruned' datasets are generated by pruning images in the increasing order of similarity values to the respective test sets. For the 'rand-pruned' datasets, we prune random points. Clearly, pruning similar images adversely affects performance in comparison to pruning dissimilar or random images. CLIP's 'near-pruned' and 'far-pruned' performance across all six test sets is shown in Fig. 7.

Motivated by the performance increase of CLIP trained on 'far-pruned' datasets, we aim to a build a 'coreset' of lower size that constitutes of samples needed to generalize well to the six test sets. To this end, we far-prune LAION-200M based on all six test sets to obtain an 'all-pruned' dataset of size 100M. In other words, we remove samples from LAION that are highly dissimilar to all samples in the six OOD datasets we consider. A CLIP model trained on the 'all-pruned' dataset performs significantly better than a CLIP model trained on a de-duplicated dataset of size 100M and roughly matches the performance of a LAION-200M trained model (see Appx. D.2).

C.5 Putting it all together

In Sec. C.1, we saw that the nearest neighbors in LAION-400M to some test sets were perceptually more similar than the neighbors in ImageNet-Train (Fig. 1). We also noted the differences in nearest neighbor similarity distributions and discovered a high correlation between nearest neighbor similarity and accuracy (refer to Fig. 4), concluding that these highly similar nearest neighbors may drive CLIP's performance. In Sec. C.4, we unconstrainedly prune similar points to verify that they drive performance.Likewise, we prune away dissimilar images and show that there's barely any drop in performance. Given these observations, it is plausible that the stellar performance on these test sets is mostly explained by higher relative similarity of LAION-400M to the OOD test sets.one cannot help but question whether LAION's vast scale really leads to the emergence of more generalizable representations or whether it is only the increased similarity to almost any test image data that lets foundation models like CLIP reach their stellar performance.

D Additional experimental results

D.1 Impact of near/far-pruning on all datasets

In Sec. C.4 by using each of the test datasets ImageNet-Val and ImageNet-Sketch and near/far-pruning LAION-200M, we trained models and reported the performance on the test datasets respectively. We now plot the performance of these models on all the six datasets in Figure 7. 'Near-pruning' ('far-pruning') with ImageNet-Sketch results in lower (higher) performance than 'near-pruning'

('far-pruning') with ImageNet-Val on ImageNet-R and ImageNet-Sketch. Likewise, 'near-pruning' ('far-pruning') with ImageNet-Val results in lower (higher) performance than 'near-pruning' ('far-pruning') with ImageNet-Sketch on ImageNet-Val, ImageNet-V2, and ObjectNet. This is expected because ImageNet-Sketch is characteristically closer to ImageNet-R, and ImageNet-Val is closer to ImageNet-V2, ObjectNet, and ImageNet-A.

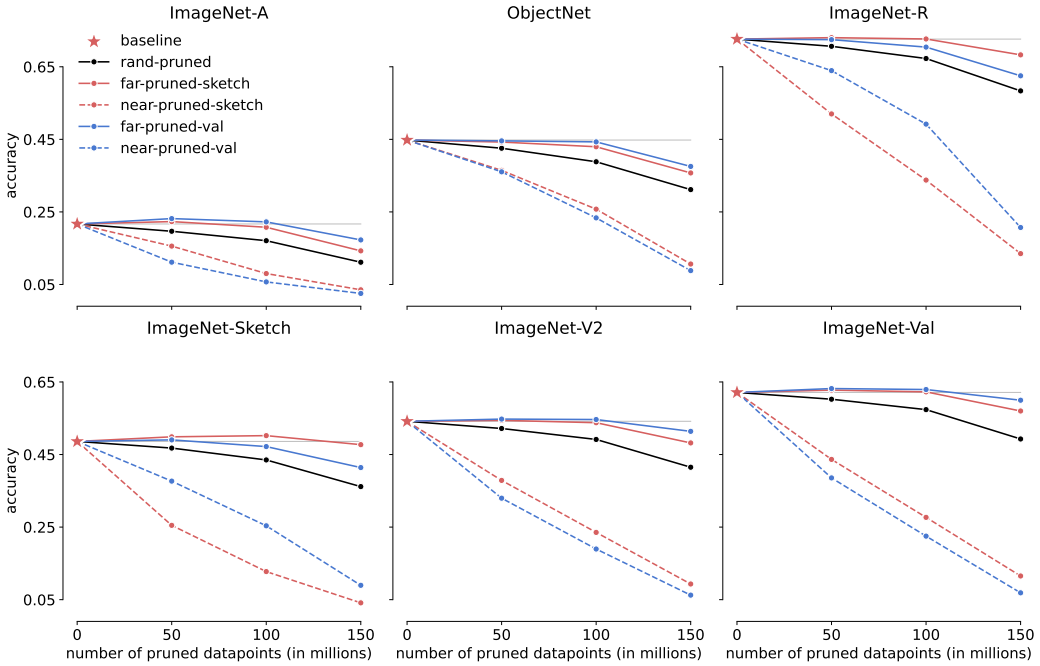

Figure 7: **The effect of 'near-pruning' and 'far-pruning' with ImageNet-Sketch or ImageNet-Val as the query dataset on the performance of all six test sets.** CLIP's zero-shot accuracy as a function of number of pruned points from LAION-200M. The baseline model is trained on de-duplicated LAION-400M which we call LAION-200M. To generate the'near-pruned' datasets we remove the images in the decreasing order of similarity (based on CLIP image-embedding similarity) to the each of the test sets ImageNet-Sketch and ImageNet-Val respectively. In contrast, the 'far-pruned' datasets are generated by dropping images in the increasing order of similarity values to the respective test sets. For the 'rand-pruned' datasets we prune random points.

D.2 Coreset of 100M

Motivated by the performance increase of the 'far-pruning' technique in the previous results, we now build several core-sets of 100M which when trained on roughly match the performance of CLIP trained on LAION-400M. Instead of pruning from the farthest point to samples in just a single test set in CLIP ViT-B/16+'s embedding space, we now prune from the farthest point to samples from a collection of test sets (all six ImageNet-1k OOD test sets). We do far-pruning with all of the test sets on both LAION-200M and LAION-400M to obtain datasets that we call 'all-far-pruned'. For comparison, we also add the performance of CLIP trained on far pruned datasets with query datasets as ImageNet-Sketch and ImageNet-Val which we call 'sketch-far-pruned' and 'val-far-pruned', respectively.

We report the results in Tab. 3 and observe that models trained on all of the splits are within $3\,\%$ average accuracy range of CLIP trained on LAION-400M. The model with the highest average accuracy is trained on 'all-far-pruned (L-200M)', which is a dataset generated by pruning far or dissimilar images in LAION-200M with all 6 test datasets as query datasets. This model also performs better than a model trained on dataset of the same size generated by the pruning technique SemDeDup [Abbas et al., 2023]. SemDeDup aims to prune data that is semantically similar with minor loss in test performance. Overall our pruning technique has identified a dataset that can be considered to

Table 3: **Performance of 'far-pruned' CLIP (ViT-B/32) on the six test sets.** We do 'far-pruning' on LAION-400M with all 6 test sets as query sets and obtain the dataset 'all-far-pruned (L-400M).' Similarly, we do 'far-pruning' on LAION-400M with all with all 6 test sets as query sets, ImageNet-Sketch, and ImageNet-Val to get the datasets 'all-far-pruned (L-200M)', 'sketch-far-pruned (L-200M)', and 'val-far-pruned (L-200M)' respectively. These models are compared to off the shelf CLIP model [Ilharco et al., 2021], model trained on LAION-200M, and a model trained on SemDeDup [Abbas et al., 2023] dataset of size 100M.

		Top-1 Accuracy						
Dataset	**Size**	**Val**	**Sketch**	**A**	**R**	**V2**	**ON**	**Avg.**
L-400M	400M	62.94	49.39	21.64	73.48	55.14	43.94	***51.09***
L-200M	199.8M	62.12	48.61	21.68	72.63	54.16	44.80	*50.67*
all-far-pruned (L-400M)	100M	61.90	48.11	19.43	70.14	53.11	39.30	*48.67*
all-far-pruned (L-200M)	100M	62.80	49.23	21.6	72.3	54.72	43.64	***50.71***
val-far-pruned (L-200M)	100M	62.79	47.53	21.65	70.40	54.35	43.70	*50.07*
sketch-far-pruned (L-200M)	100M	62.27	50.21	20.77	72.67	53.77	42.95	*50.44*
SemDeDup	100M	52.19	41.70	16.71	67.05	44.96	39.59	*43.7*

Table 4: **Corrected zero-shot performance of CLIP ViT-B/32.** 'X-pruned' represents a pruned dataset from LAION-200M such that the similarity gap to 'X' is the roughly the same as the similarity gap of ImageNet to 'X'. The sizes of these subsets are subtracted from the LAION-200M's size. Here, 'X' is one of the 6 standard ImageNet test sets. 'combined-pruned' splits ensure a similarity gap of LAION-200M and ImageNet-Train to all 6 test sets. CLIP's corrected zero-shot performance drops the most on ImageNet-Sketch and ImageNet-R with a relative performance drop of $11.08\,\%$ and $5.99\,\%$ respectively. Red color indicates a drop in performance on the respective test set, and blue represents a rise. Overall high performance indicates that highly similar images do not play a key role in explaining CLIP's generalization ability.

			Top-1 Accuracy					
Model	**Dataset**	**Size**	**Val**	**Sketch**	**A**	**R**	**V2**	**ObjectNet**
ViT-B/32	OpenAI	400M	63.38	42.32	31.44	69.24	55.96	44.14
ViT-B/32	L-400M	413M	62.94	49.39	21.64	73.48	55.14	43.94
ViT-B/32	L-200M	199,824,274	62.12	48.61	21.68	72.63	54.16	44.80
ViT-B/32	├─ val-pruned	-377,340	62.12	48.38	21.45	72.2	54.76	42.79
ViT-B/32	├─ sketch-pruned	-8,342,783	61.55	43.22	22.28	69.6	53.53	42.77
ViT-B/32	├─ a-pruned	-138,852	62.49	48.49	21.63	72.15	54.38	43.25
ViT-B/32	├─ r-pruned	-5,735,749	61.73	45.66	21.67	68.28	54.1	42.90
ViT-B/32	├─ v2-pruned	-274,325	62.48	48.62	22.13	72.3	53.83	43.38
ViT-B/32	├─ objectnet-pruned	-266,025	62.30	49.03	22.64	72.90	54.21	42.80
ViT-B/32	└─ combined-pruned	-12,352,759	61.5	41.97	21.72	67.25	53.65	42.23
ResNet-101	ImageNet-1k	1.2M	77.21	27.58	4.47	39.81	65.56	36.63

contain all data points that impact OOD performance and should consequently greatly ease further study.

D.3 Main experiments without ImageNet

We repeat the experiments in Sec. 4 without adding ImageNet-Train to LAION-200M and report results in Tab. 4. We observe the same trends as we did in Tab. 1.

E Nearest neighbor visualizations

E.1 LAION-400M vs ImageNet-Train

Just like in Figure 1, we plot the nearest neighbors in LAION-400M and ImageNet-Train of random query images from each of the six datasets in Figures 8, 9, 10, 11, 12, and 13.

Figure 8: Nearest neighbors of *randomly* sampled ImageNet-Sketch queries in LAION-400M and ImageNet-Train ordered by decreasing perceptual similarity. We omit duplicates within the nearest neighbors. Perceptual similarity is computed in CLIP's image embedding space and can be thought of as measuring the "perceptual closeness" of images in terms of both content and style.

E.2 After pruning

Percentage of images in each of the six datasets that have higher similarity to LAION-200M/LAION-400M than ImageNet-Train are reported in Tab. 5. For each of the six test sets, we randomly sample query images from them that are more similar to LAION-200M than ImageNet-Train and plot the

Figure 9: Nearest neighbors of *randomly* sampled ImageNet-Val queries in LAION-400M and ImageNet-Train ordered by decreasing perceptual similarity. We omit duplicates within the nearest neighbors. Perceptual similarity is computed in CLIP's image embedding space and can be thought of as measuring the "perceptual closeness" of images in terms of both content and style.

nearest neighbors in ImageNet-Train, LAION-200M, and LAION-200M after being pruned by the respective test in Figures 14, 15, 16, 17, 18, and 19.

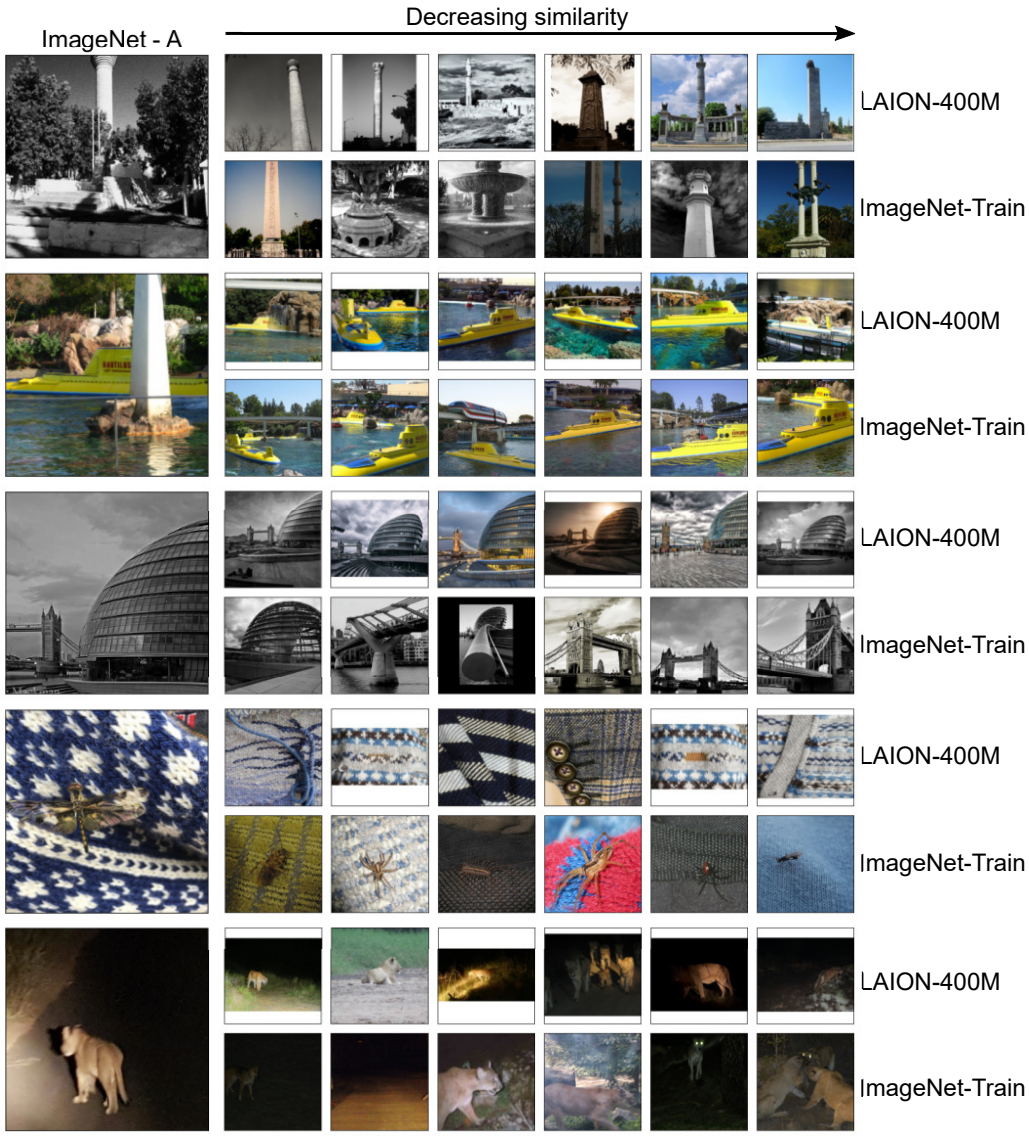

Figure 10: Nearest neighbors of *randomly* sampled ImageNet-A queries in LAION-400M and ImageNet-Train ordered by decreasing perceptual similarity. We omit duplicates within the nearest neighbors. Perceptual similarity is computed in CLIP's image embedding space and can be thought of as measuring the "perceptual closeness" of images in terms of both content and style.

Table 5: **Percentage (%) of points in the test datasets for which the nearest neighbor is in LAION-400M or LAION-200M and not in ImageNet-Train.**

Dataset	Size	LAION-400M	LAION-200M
ImageNet-Val	50000	16.80	14.88
ImageNet-Sketch	50889	97.94	97.45
ImageNet-R	30000	87.88	86.74
ImageNet-A	7500	47.39	45.53
ImageNet-V2	10000	38.95	35.48
ObjectNet	18574	63.24	61.62

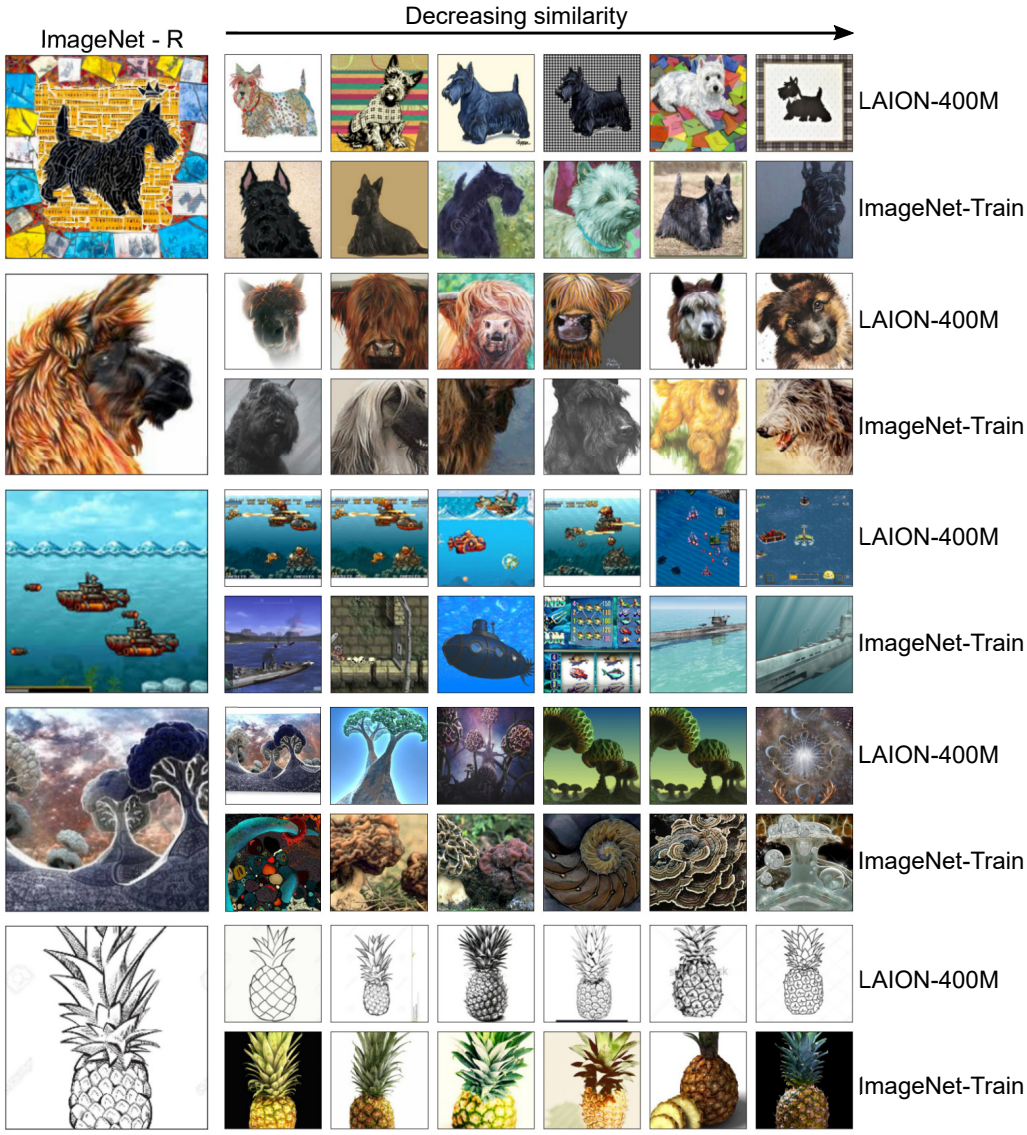

Figure 11: Nearest neighbors of *randomly* sampled ImageNet-R queries in LAION-400M and ImageNet-Train ordered by decreasing perceptual similarity. We omit duplicates within the nearest neighbors. Perceptual similarity is computed in CLIP's image embedding space and can be thought of as measuring the "perceptual closeness" of images in terms of both content and style.

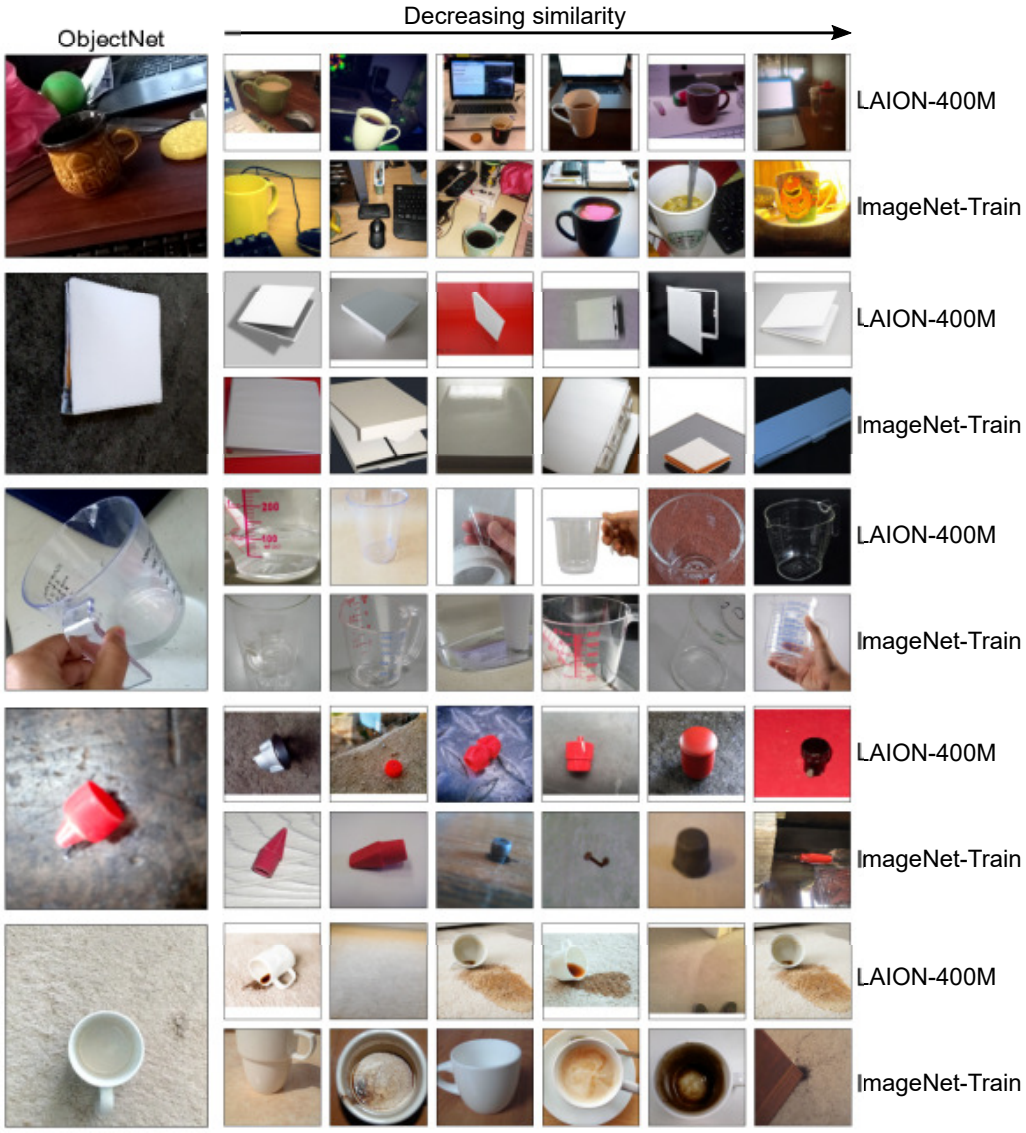

Figure 12: Nearest neighbors of *randomly* sampled ObjectNet queries in LAION-400M and ImageNet-Train ordered by decreasing perceptual similarity. We omit duplicates within the nearest neighbors. Perceptual similarity is computed in CLIP's image embedding space and can be thought of as measuring the "perceptual closeness" of images in terms of both content and style.

Figure 13: Nearest neighbors of *randomly* sampled ImageNet-V2 queries in LAION-400M and ImageNet-Train ordered by decreasing perceptual similarity. We omit duplicates within the nearest neighbors. Perceptual similarity is computed in CLIP's image embedding space and can be thought of as measuring the "perceptual closeness" of images in terms of both content and style.

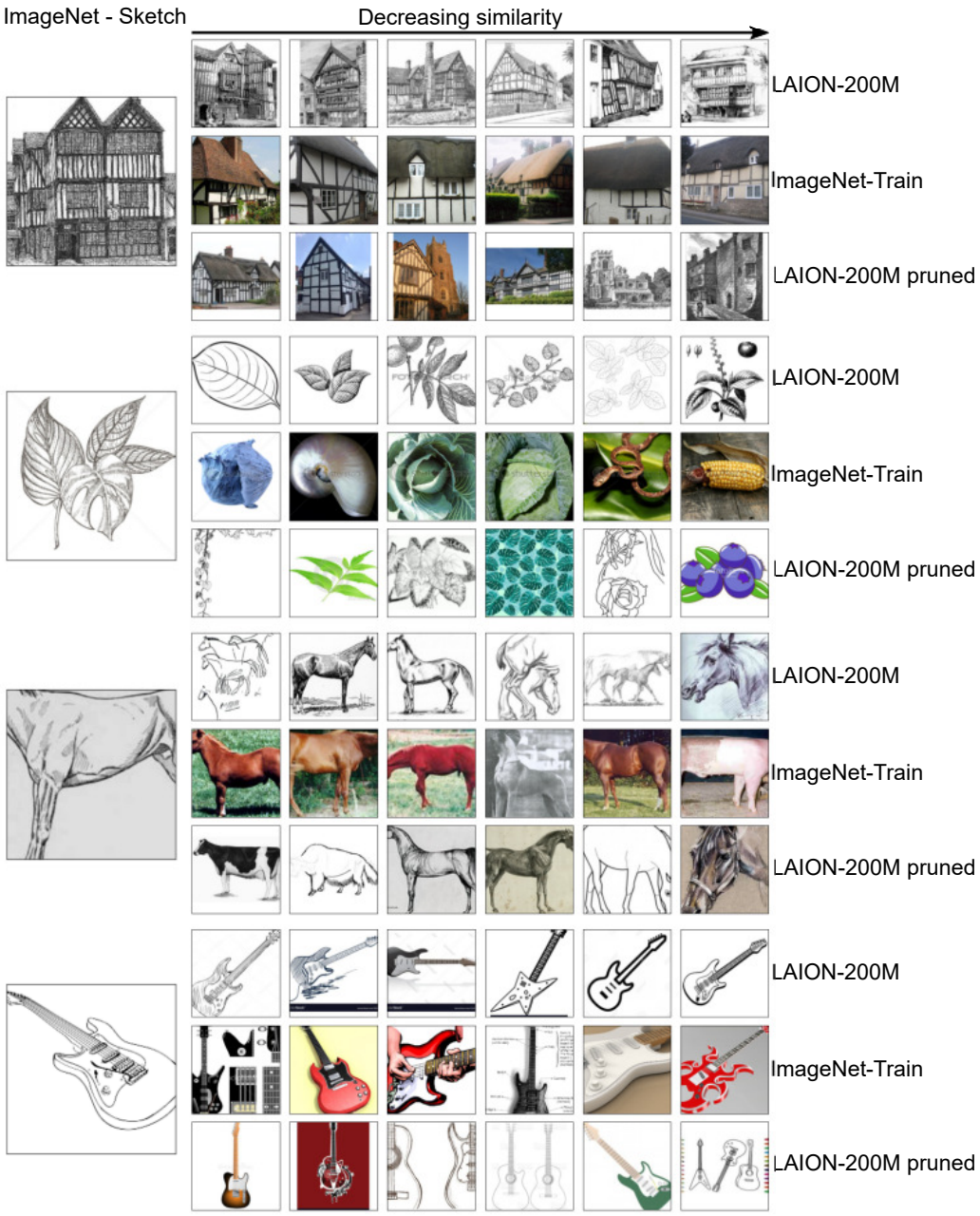

Figure 14: Nearest neighbors of ImageNet-Sketch images in LAION-200M, ImageNet-Train, and 'sketch-pruned' (LAION-200M pruned) ordered by decreasing perceptual similarity. The query (base) images are *randomly* sampled from the set of images that are more similar to LAION-200M than ImageNet-Train to see the effect of pruning (see Tab. 5). We omit duplicates within the nearest neighbors. Perceptual similarity is computed in CLIP's image embedding space and can be thought of as measuring the "perceptual closeness" of images in terms of both content and style. LAION-200M clearly contains more similar images to samples in the test set compared to ImageNet-Train or 'sketch-pruned'.

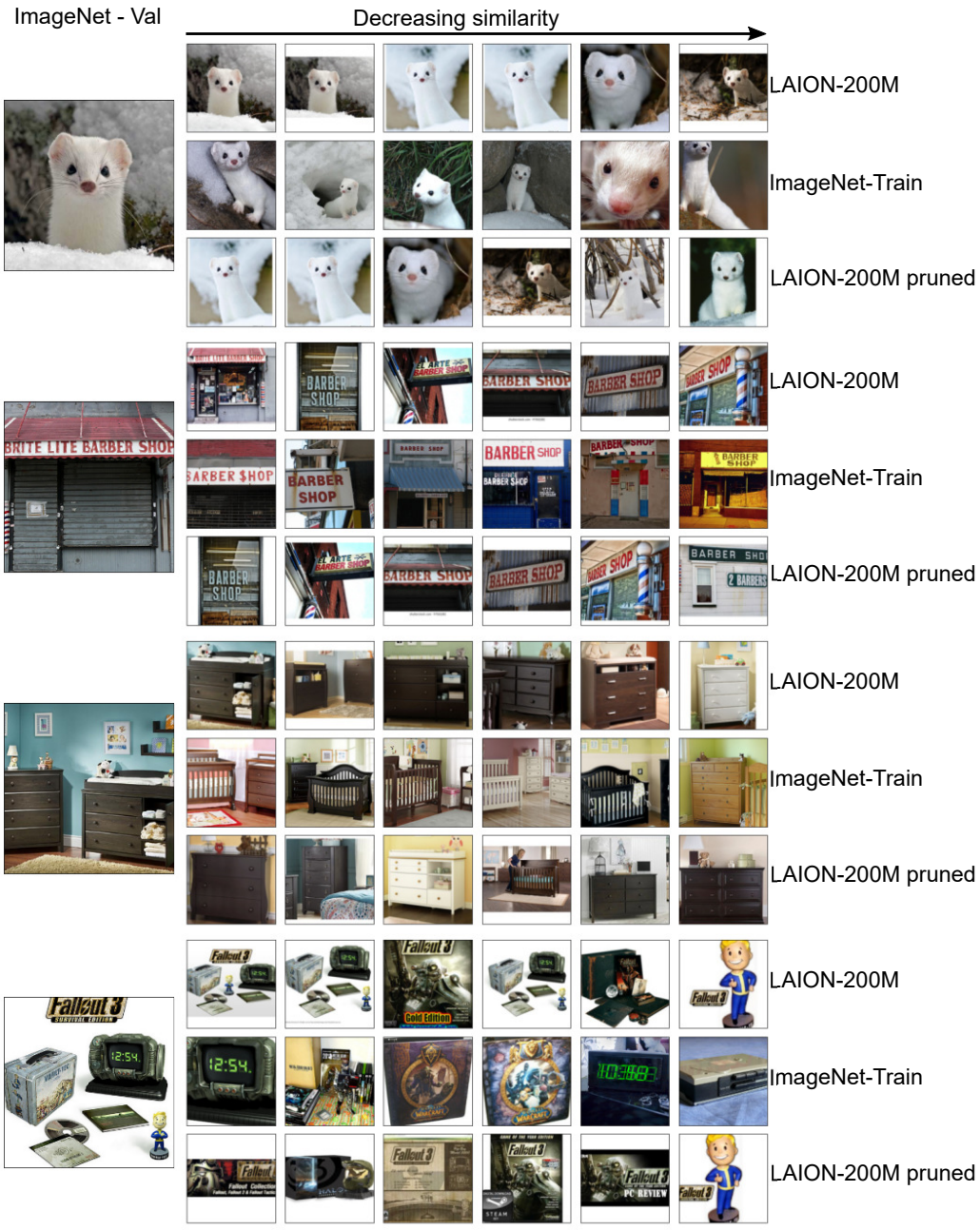

Figure 15: Nearest neighbors of ImageNet-Val images in LAION-200M, ImageNet-Train, and 'val-pruned' (LAION-200M pruned) ordered by decreasing perceptual similarity. The query (base) images are *randomly* sampled from the set of images that are more similar to LAION-200M than ImageNet-Train to see the effect of pruning (see Tab. 5). We omit duplicates within the nearest neighbors. Perceptual similarity is computed in CLIP's image embedding space and can be thought of as measuring the "perceptual closeness" of images in terms of both content and style. LAION-200M clearly contains more similar images to samples in the test set compared to 'val-pruned'; ImageNet-Train images are in-distribution to ImageNet-Val and, therefore, contain similar samples.

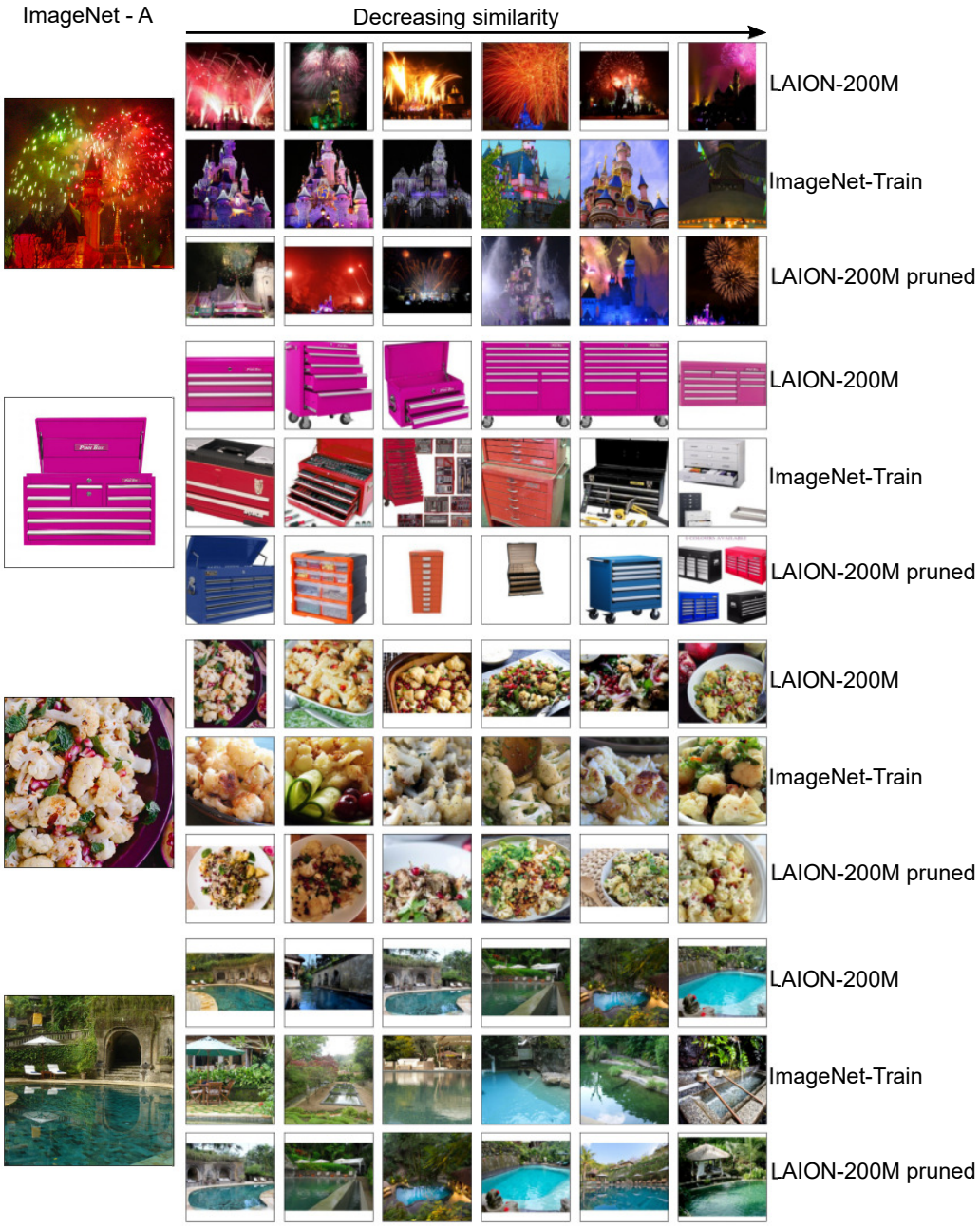

Figure 16: Nearest neighbors of ImageNet-A images in LAION-200M, ImageNet-Train, and 'a-pruned' (LAION-200M pruned) ordered by decreasing perceptual similarity. The query (base) images are *randomly* sampled from the set of images that are more similar to LAION-200M than ImageNet-Train to see the effect of pruning (see Tab. 5). We omit duplicates within the nearest neighbors. Perceptual similarity is computed in CLIP's image embedding space and can be thought of as measuring the "perceptual closeness" of images in terms of both content and style. LAION-200M clearly contains more similar images to samples in the test set compared to 'val-pruned'; ImageNet-Train images are in-distribution to ImageNet-Val and, therefore, contain similar samples.

Figure 17: Nearest neighbors of ImageNet-R images in LAION-200M, ImageNet-Train, and 'r-pruned' (LAION-200M pruned) ordered by decreasing perceptual similarity. The query (base) images are *randomly* sampled from the set of images that are more similar to LAION-200M than ImageNet-Train to see the effect of pruning (see Tab. 5). We omit duplicates within the nearest neighbors. Perceptual similarity is computed in CLIP's image embedding space and can be thought of as measuring the "perceptual closeness" of images in terms of both content and style. LAION-200M clearly contains more similar images to samples in the test set compared to 'val-pruned'; ImageNet-Train images are in-distribution to ImageNet-Val and, therefore, contain similar samples.

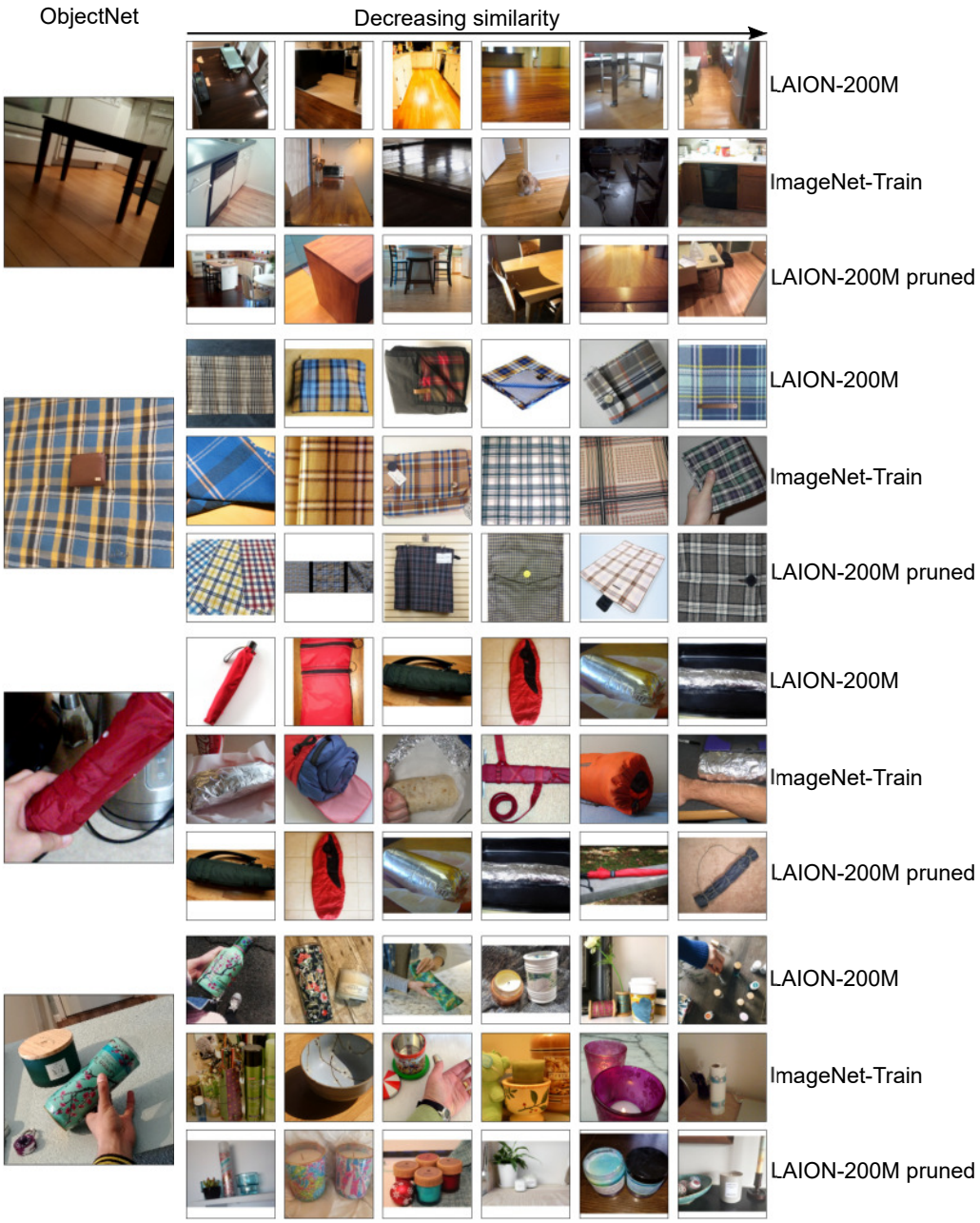

Figure 18: Nearest neighbors of ObjectNet images in LAION-200M, ImageNet-Train, and 'v2-pruned' (LAION-200M pruned) ordered by decreasing perceptual similarity. The query (base) images are *randomly* sampled from the set of images that are more similar to LAION-200M than ImageNet-Train to see the effect of pruning (see Tab. 5). We omit duplicates within the nearest neighbors. Perceptual similarity is computed in CLIP's image embedding space and can be thought of as measuring the "perceptual closeness" of images in terms of both content and style. LAION-200M clearly contains more similar images to samples in the test set compared to 'val-pruned'; ImageNet-Train images are in-distribution to ImageNet-Val and, therefore, contain similar samples.

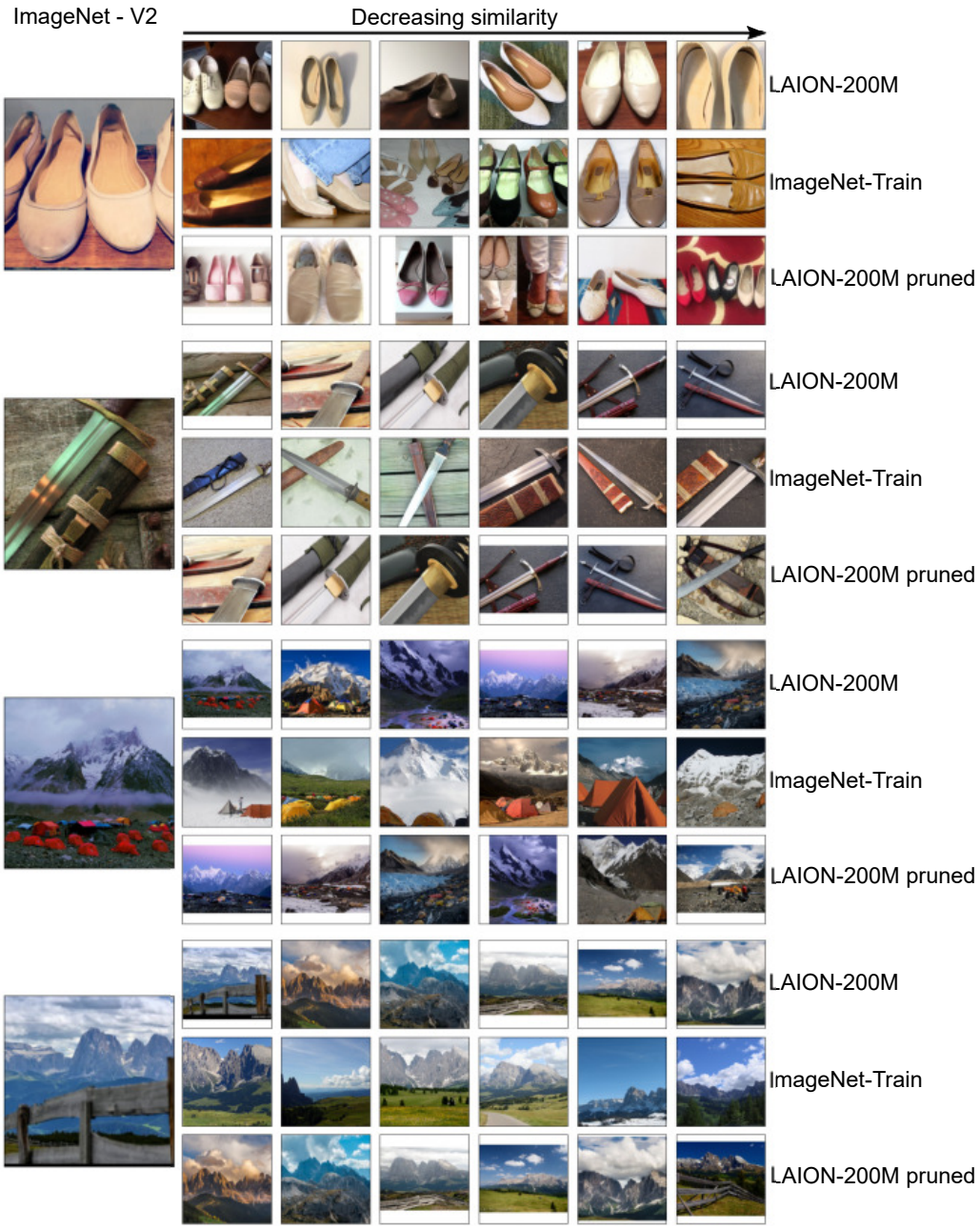

Figure 19: Nearest neighbors of ImageNet-V2 images in LAION-200M, ImageNet-Train, and 'objectnet-pruned' (LAION-200M pruned) ordered by decreasing perceptual similarity. The query (base) images are *randomly* sampled from the set of images that are more similar to LAION-200M than ImageNet-Train to see the effect of pruning (see Tab. 5). We omit duplicates within the nearest neighbors. Perceptual similarity is computed in CLIP's image embedding space and can be thought of as measuring the "perceptual closeness" of images in terms of both content and style. LAION-200M clearly contains more similar images to samples in the test set compared to 'val-pruned'; ImageNet-Train images are in-distribution to ImageNet-Val and, therefore, contain similar samples.