Data Filtering Networks

Alex Fang[*]
Apple

Albin Madappally Jose
Apple

Amit Jain
Apple

Ludwig Schmidt
University of Washington

Alexander Toshev
Apple

Vaishaal Shankar
Apple

Abstract

Large training sets have become a cornerstone of machine learning and are the foundation for recent advances in language modeling and multimodal learning. While data curation for pre-training is often still ad-hoc, one common paradigm is to first collect a massive pool of data from the Web and then filter this candidate pool down to an actual training set via various heuristics. In this work, we study the problem of learning a *data filtering network* (DFN) for this second step of filtering a large uncurated dataset. Our key finding is that the quality of a network for filtering is distinct from its performance on downstream tasks: for instance, a model that performs well on ImageNet can yield worse training sets than a model with low ImageNet accuracy that is trained on a small amount of high-quality data. Based on our insights, we construct new data filtering networks that induce state-of-the-art image-text datasets. Specifically, our best performing dataset DFN-5B enables us to train state-of-the-art CLIP models for their compute budgets: among other improvements on a variety of tasks, a ViT-H trained on our dataset achieves 83.0% zero-shot transfer accuracy on ImageNet, out-performing models trained on other datasets such as LAION-2B, DataComp-1B, or OpenAI's WIT. In order to facilitate further research in dataset design, we also release a new 2 billion example dataset DFN-2B and show that high performance data filtering networks can be trained from scratch using only publicly available data. Full version at `https://arxiv.org/abs/2309.17425`.

1 Introduction

Carefully curated datasets have driven progress in machine learning for decades, from early pattern recognition experiments in Bell Labs to recent developments like GPT-4, Stable Diffusion, and CLIP (Highleyman & Kamentsky, 1959; LeCun et al., 1989, 1998; Deng et al., 2009; Krizhevsky et al., 2009, 2012; Radford et al., 2019, 2021, 2022; OpenAI, 2023). Despite their crucial role, datasets themselves are rarely the subject of active research (Sambasivan et al., 2021).

Current approaches to improving performance on machine learning tasks have focused on scaling model capacity or training data volume. While scaling laws (Hestness et al., 2017; Kaplan et al., 2020; Aghajanyan et al., 2023; Cherti et al., 2023) have elucidated the relationship between model size, data size, and performance, little formal guidance exists on how to scale these quantities. On the model side, experimentation is straightforward - with enough compute, permutations of width, depth, normalization and training hyperparameters can be rigorously evaluated, leading to consistent modeling improvements over the years (Touvron et al., 2023a,b; Elsen et al., 2023).

We argue dataset design can leverage the same tools as model design. Almost all large-scale dataset construction can be broken down into two phases: uncurated data collection and dataset filtering. We focus our work on the latter, with the assumption that a large uncurated dataset exists. We show data

[*] Affiliated with University of Washington, work done while at Apple

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

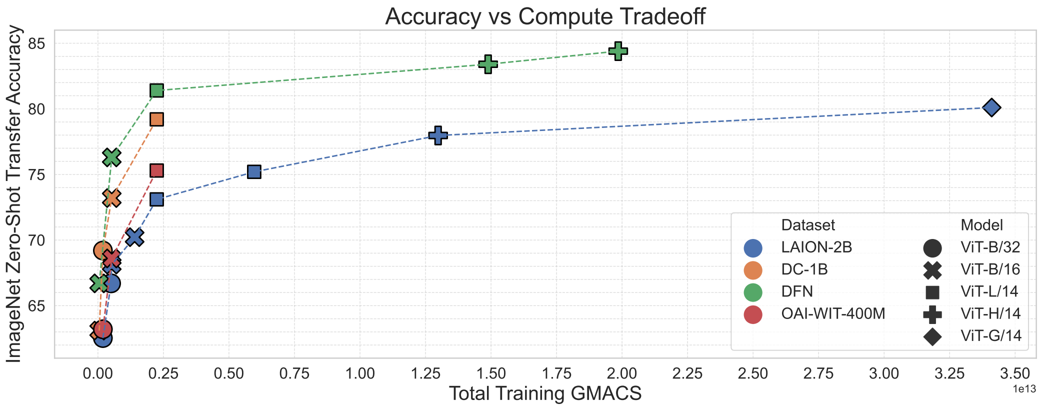

Figure 1: Compute scaling behavior of training CLIP models on various datasets. DFN-2B, the subset of CommonPool (DataComp-12.8B) chosen by our best performing data filtering networks, out-performs all other datasets including OpenAI's WIT and the previous state-of-the-art CLIP training dataset DataComp-1B. Our ViT-L outperforms a ViT-G trained on LAION with 18× more compute. Similarly, our ViT-B/16 outperforms OpenAI's ViT-L/14 trained with 4× more compute. Our ViT-H/14 achieves 83.0% on ImageNet, out-performing any model in its compute class.

filtering networks (DFNs) - neural networks designed to filter data - can induce massive, high-quality pre-training datasets. Unlike previous techniques relying on domain-specific heuristics, DFNs paired with a large unfiltered image-text pool produce billion-scale state-of-the-art datasets algorithmically. We demonstrate DFNs can be efficiently trained from scratch and improved with the same techniques as standard ML models.

The contributions of this work are as follows. First, we characterize the properties of data filtering networks that lead to high-quality datasets. We ablate properties of data filtering networks from supervision signal to training data quality. We find that a small contrastive image-text model trained on *only* high-quality data is sufficient to construct state-of-the-art datasets.

Second, we use these properties to train DFNs and construct datasets that induce Contrastive Image-Text Pre-trained (CLIP) models that achieve high accuracy and present better compute accuracy tradeoff than any existing dataset in the literature as show in Figure 1. In particular we train a ViT-L/14 or 12.8B examples seen on our DFN induced dataset DFN-2B to 81.4 ImageNet zero-shot transfer accuracy, outperforming the previous best ViT-L trained on DataComp-1B by over 2 percentage points. We further train a ViT-H/14 on a larger DFN induced dataset DFN-5B to 83.0 ImageNet zero-shot transfer accuracy. We show that models trained on these datasets show consistent improvements on many tasks, including zero-shot classification, retrieval, and visual question answering, and maintain the favorable robustness properties of CLIP models.

Lastly, the above insights can be used as a recipe to construct high-quality datasets from scratch by using only public data [2] thus making strides towards democratization of large high-quality datasets. In addition, we release DFN-2B for the community to enable research on large image-text models.

2 Data Filtering Networks

Definitions Since our ultimate goal is to build functions that filter potentially trillions of examples efficiently, we restrict the scope of our study to DFNs that are only applied pointwise to elements of a larger data pool. Thus, processing a data pool with a DFN lends itself to parallelization and thus efficient application. For a given DFN and pool, we refer to the data pool we train the DFN on as a *filter dataset*. Furthermore, we refer to the dataset constructed by filtering the pool with the DFN the *induced dataset*. We then refer to a model trained (only) on that dataset the *induced model*. As introduced in Section A.2 a common choice for a DFN is a CLIP trained image-text model. Thus, a DFN can not only be used to induce a dataset but also be applied to common evaluation

[2] Most large public image-text datasets including LAION-5B and DataComp-1B are built using OpenAI's CLIP model

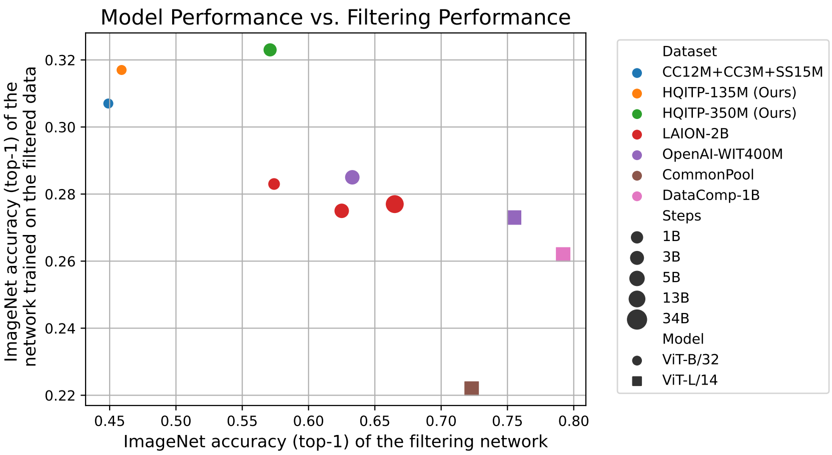

Figure 2: Filtering strength is uncorrelated with image task performance. The models are trained using CLIP, and the number of samples seen and the training data are displayed on the right hand side. Filtering performance is measured by filtering on DataComp medium.

problems such as zero-shot ImageNet classification. Inversely, a CLIP model can be both used for general recognition as well as as a DFN. When we use a CLIP model as a DFN, we define its *filtering performance* as the performance of the induced model, as evaluated on standard benchmarks, e.g. ImageNet top-1. We further explain our evaluation setup in Appendix B

Understanding Data Filtering Networks As open source CLIP-models improve on standard vision metrics such as ImageNet, the question arises whether we can replace the OpenAI CLIP model used in the dataset construction process with one of these better models. Figure 2 shows that ImageNet performance of CLIP models is not correlated with filtering performance. To measure filtering performance, we create a dataset by using the CLIP model to apply CLIP filtering on DataComp's medium raw pool, and measure ImageNet performance of models trained on the induced dataset. It is especially striking that a model with 30% less ImageNet performance than OpenAI's CLIP models can be as good when used as a filtering model.

We find that data quality is key to training good filtering models. To demonstrate this, we start with a high-quality pool of 10 million samples from Conceptual 12M (CC12M), and gradually replace it with unfiltered data from Common Crawl until this pool only contains Common Crawl. We train DFNs on these data mixes, and use these DFNs to CLIP filter a separate pool of 128 million Common Crawl samples from DataComp's medium scale. In Appendix Figure 5, we measure the ImageNet performance of both the DFNs and the induced models trained on datasets generated by each of the DFNs. While the ImageNet performance of the DFNs degrade steadily as they are trained on larger fractions of unfiltered data, their performance as filtering networks decreases immediately when the high-quality pool is "poisoned" with even a small portion of unfiltered data. Once the filtering training pool is poisoned, the dataset induced by the DFN is only slightly better than unfiltered data.

3 Creating Better Data Filtering Networks

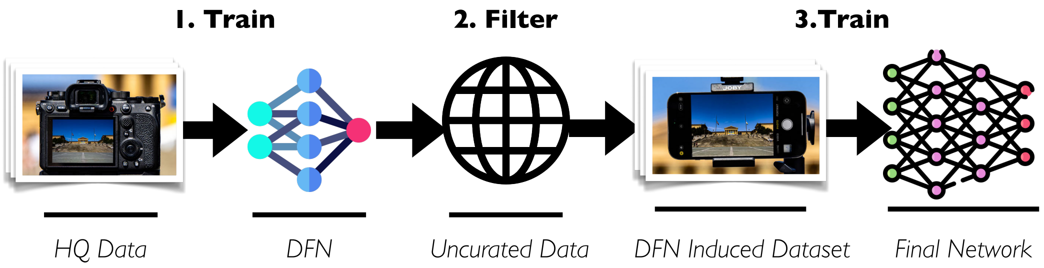

Figure 3: A high level overview of our pipeline for constructing datasets using DFNs

Table 1: Training on DFN-2B produces state-of-the-art CLIP models. Here we evaluate on the DataComp benchmark, comparing against LAION-2B, DC-1B, MetaCLIP and OpenAI WIT-400M. Additional comparisons can be found on the DataComp leaderboard.

Dataset	DataComp Scale	IN	IN Shifts	VTAB	Retrieval	Average
DC-1B	medium	0.297	0.239	0.346	0.231	0.328
DFN-2B	medium	0.371	0.298	0.388	0.288	0.373
DC-1B	large	0.631	0.508	0.546	0.498	0.537
DFN-2B	large	0.678	0.540	0.555	0.534	0.560
LAION-2B	xlarge	0.731	0.603	0.586	0.589	0.601
OpenAI WIT-400M	xlarge	0.755	0.649	0.586	0.543	0.617
DC-1B	xlarge	0.792	0.679	0.652	0.608	0.663
DFN-2B	xlarge	0.814	0.688	0.656	0.649	0.669
LAION-2B	ViT-G/14-224px	0.801	0.691	0.646	0.635	0.667
DC-1B (CLIPA-v2)	ViT-G/14-224px	0.831	**0.740**	0.645	0.631	0.684
MetaCLIP	ViT-H/14-336px	0.805	0.700	0.640	0.652	0.667
WebLI	ViT-SO/400M-384px	0.831	0.734	0.648	**0.698**	0.692
DFN-2B	ViT-L/14-224px	0.822	0.679	0.664	0.666	0.678
DFN-5B	ViT-H/14-224px	0.834	0.713	0.675	0.684	0.698
DFN-5B	ViT-H/14-378px	**0.844**	0.738	**0.685**	0.695	**0.710**

Equipped with a better understanding of CLIP models as data filtering networks, we aim to create better data filtering networks. DFNs can be trained and modified in the same ways as standard machine learning models. We start by training a CLIP model on a high-quality dataset, and then we can fine-tune the filtering network on subsequent datasets that we want to do especially well on. We use weight ensembling to reduce overfitting on the fine-tuned datasets. Standard machine learning techniques such as augmentation, using a different initialization, and training for more steps with a larger batch size seem to improve the filtering model. We demonstrate the effect of these interventions in Appendix Table 7. On the other hand, using a different model size seems to have limited benefits, while model ensembling increases filtering costs without bringing gains. Compared to previous datasets such as DataComp-1B (DC-1B) which involved combining CLIP filtering with clustering-based heuristics, DFNs simplify the data filtering process into a single pipeline while also reducing computational costs.

To create our best DFN, we train a ViT-B/32 CLIP model on High-Quality Image-Text Pairs (HQITP-350M), which is a high-quality dataset of 357 million image-text samples with human-verified captions. This dataset is similar to the HQITP-135M used in Ranasinghe et al. (2023), but expanded to 357M examples. We initialize the weights with OpenAI's checkpoint. We then fine-tune on the combined MS COCO training set, Flickr30k training set, and ImageNet 1k with OpenAI templates as the captions. We use additional augmentation at both training and fine-tuning time. Additional training details are in Appendix H. We create our dataset DFN-2B by applying this DFN on DataComp's full 12.8 billion sample CommonPool, with a threshold equivalent to taking the top 15% of samples.

Our DFN induces datasets that achieve state-of-the-art results on multiple scales in DataComp. In particular at xlarge, we train a ViT-L/14 on DFN-2B for 12.8B samples seen to achieve 81.4% zero-shot accuracy on ImageNet, and a 0.669 average over 38 DataComp evaluation datasets. As shown in Table 1, in terms of ImageNet zero-shot improvement, this is a 2.2% improvement over DC-1B, a 5.9% improvement over OpenAI WIT-400M, and a 8.3% improvement over LAION-2B. These improvements are beyond ImageNet, as we can see similar trends across the DataComp evaluation suite in distribution shifts, retrieval, VTAB, and average performance, as well as on VQA benchmarks (Appendix E). We also train a ViT-L/14 on DFN-2B for 39B samples seen. Lastly, we train DFN-5B on a ViT-H/14 for 39B samples seen at 224×224 resolution, and 5B samples at 378×378 resolution – achieving 84.4% zero-shot transfer accuracy on ImageNet, and 0.710 average on the DataComp evaluation suite. We find that models trained our DFN produced datasets outperform all other models on the evaluation suite regardless of pre-training dataset: MetaClip, WebLI or DataComp-1B (Xu et al., 2023; Zhai et al., 2022a; Gadre et al., 2023), archiectural improvements such as shape-optimized ViTs (Alabdulmohsin et al., 2023), a more performant sigmoid loss (Zhai et al., 2023), or pre-training performance optimizations such as those in Li et al. (2023b).

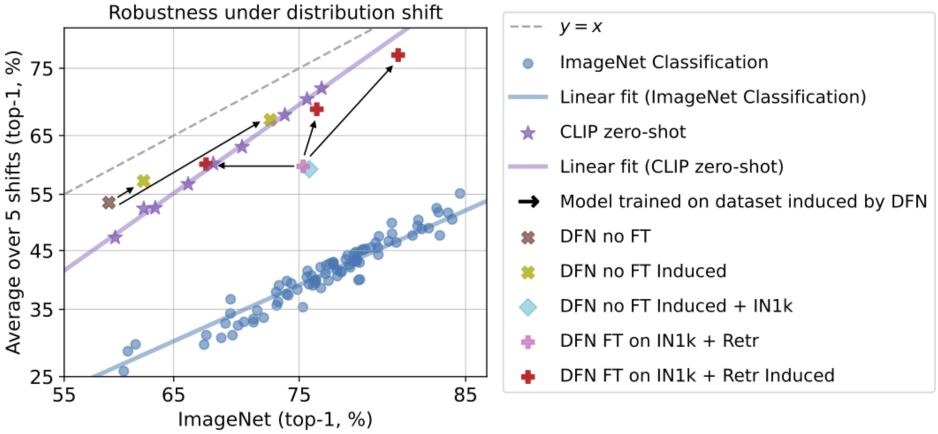

Figure 4: Datasets induced by DFNs can be robust to distribution shift. DFNs can be fine-tuned to maintain robustness of induced datasets, unlike directly training on ImageNet (IN). DFNs are not performing distillation because induced datasets lead to higher performing models than the original DFN. Distribution shifts used are IN-V2, ObjectNet, IN-Sketch, IN-R, and IN-A.

Creating better datasets not only improves model performance, but also improves model efficiency. Performance that was once only achievable by larger models can be matched with a smaller model trained on a better dataset. Our ViT-L/14 trained on DFN-2B surpasses a ViT-G/14 trained on LAION-2B for 34B samples seen by 1.5% zero-shot accuracy on ImageNet, and by 0.002 average performance, despite using 16x less computational cost [3]. Similarly, we can train a ViT-B/16 on DFN-2B for 12.8B samples seen to achieve competitive performance with OpenAI's ViT-L/14, representing a 4x computational cost reduction.

The key to training good DFNs is using high-quality data for training the filtering network. Collecting verified high-quality data is expensive, as it often requires human annotations, and is thus difficult to scale to large quantities. But given a sizable high-quality dataset, we can explore if there are benefits to directly training on it instead of using it to train a DFN. In Appendix Table 9, we compare models trained on datasets induced by our DFNs with a model trained on HQITP-350M combined with the dataset induced by CLIP filtering CommonPool with OpenAI's ViT-B/32. Models trained on DFN induced datasets outperform the baseline on all major categories of the DataComp evaluation suite. Furthermore, training on the combination of HQITP-350M and DFN-2B seems to have little improvement when compared to just training on DFN-2B. By training a DFN instead of directly training on high-quality data, we demonstrate a successful recipe for leveraging high-quality data for creating large-scale high-quality datasets.

We also explore the differences between fine-tuning a DFN and directly training on the fine-tuning dataset. In Figure 4 and Table 5, we compare a dataset induced by a baseline DFN, a dataset induced by the baseline DFN fine-tuned on ImageNet, and a dataset induced by the baseline DFN without fine-tuning on ImageNet combined with ImageNet. While directly training on ImageNet leads to higher performance on ImageNet and ImageNet-V2, it does not improve upon the baseline for the ObjectNet, ImageNet-Sketch, and ImageNet-R (Recht et al., 2019; Barbu et al., 2019; Wang et al., 2019; Hendrycks et al., 2021a,b). However, the DFN fine-tuned on ImageNet induces a dataset that improves over the baseline on ImageNet and all of its distribution shifts. Fine-tuning on DFNs acts as a regularizer to induce datasets similar to the fine-tuning dataset, while maintaining strong robustness properties that come with drawing from a more distributionally diverse candidate pool.

Publicly Reproducible DFNs Scientific research benefits from results that can be reproduced by anyone from scratch. Though OpenAI's internal dataset and HQITP-350M are not publicly accessible, we demonstrate that a competitive DFN can be trained on public data sources. We train a ViT-B/32 on Conceptual Caption12M, Conceptual Captions 3M, and Shutterstock 15M (Changpinyo et al., 2021; Sharma et al., 2018; Nguyen et al., 2023). As shown in Appendix Table 10 , this DFN surpasses OpenAI's ViT-B/32 in terms of filtering performance at DataComp's medium, large, and xlarge scales. Additionally, this DFN can be modified as described above to further improve filtering performance.

[3] calculation does not take into account patch dropout used to train ViT-G/14 on LAION-2B

References

Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. Scaling laws for generative mixed-modal language models. *arXiv preprint arXiv:2301.03728*, 2023.

Ibrahim Alabdulmohsin, Xiaohua Zhai, Alexander Kolesnikov, and Lucas Beyer. Getting vit in shape: Scaling laws for compute-optimal model design. *arXiv preprint arXiv:2305.13035*, 2023.

Apple. axlearn, July 2023.

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. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32. Curran Associates, Inc., 2019. `https://proceedings.neurips.cc/paper/2019/file/97af07a14cacba681feacf3012730892-Paper.pdf`.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3558–3568, 2021.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. `https://ieeexplore.ieee.org/abstract/document/5206848`.

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.

Erich Elsen, Augustus Odena, Maxwell Nye, Sağnak Taşırlar, Tri Dao, Curtis Hawthorne, Deepak Moparthi, and Arushi Somani. Releasing Persimmon-8B, 2023. URL `https://www.adept.ai/blog/persimmon-8b`.

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), 2022.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023.

Xinyang Geng, Hao Liu, Lisa Lee, Dale Schuurams, Sergey Levine, and Pieter Abbeel. Multimodal masked autoencoders learn transferable representations. *arXiv preprint arXiv:2205.14204*, 2022.

Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, and Aditi Raghunathan. Finetune like you pretrain: Improved finetuning of zero-shot vision models, 2022.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. `https://arxiv.org/abs/1512.03385`.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021b.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically, 2017.

W. H. Highleyman and L. A. Kamentsky. A generalized scanner for pattern- and character-recognition studies. In *Papers Presented at the the March 3-5, 1959, Western Joint Computer Conference*, IRE-AIEE-ACM '59 (Western), pp. 291–294, New York, NY, USA, 1959. Association for Computing Machinery. ISBN 9781450378659. doi: 10.1145/1457838.1457894. URL `https://doi.org/10.1145/1457838.1457894`.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering, 2019.

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`.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009. `https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf`.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012. `https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf`.

Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989. doi: 10.1162/neco.1989.1.4.541.

Yann LeCun, Yann LeCun, and Yann LeCun. The mnist database of handwritten digits, 1998. `http://yann.lecun.com/exdb/mnist/`.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023a.

Xianhang Li, Zeyu Wang, and Cihang Xie. Clipa-v2: Scaling clip training with 81.1accuracy within a $10,000 budget; an extra $4,000 unlocks 81.82023b.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 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, 2023.

OpenAI. Gpt-4 technical report, 2023.

Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, Mingxing Tan, and Quoc V. Le. Combined scaling for zero-shot transfer learning, 2023.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. `https://arxiv.org/abs/2103.00020`.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022.

Kanchana Ranasinghe, Brandon McKinzie, Sachin Ravi, Yinfei Yang, Alexander Toshev, and Jonathon Shlens. Perceptual grouping in contrastive vision-language models, 2023.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400. PMLR, 09–15 Jun 2019. `http://proceedings.mlr.press/v97/recht19a.html`.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2021.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

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 (NeurIPS)*, pp. 10506–10518, 2019.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*, 2019.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models, 2022.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data, 2023.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers, 2022a.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning, 2022b.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023.

A Background and Related Work

A.1 Contrastive Image Language Pre-training (CLIP)

CLIP has altered the use of cheaply available image-alt-text datasets by demonstrating the practicality of large-scale training on web-scraped image-text pairs to build state-of-the-art image representations. CLIP consists of separate vision and text encoders, and uses contrastive loss during training to push the representations of related images and text pairs together, and unrelated pairs apart. Crucial to this process is a large dataset of *aligned* image-text pairs - images paired with semantically relevant text. The release of CLIP was followed by several other image-text models such as ALIGN, BASIC, LiT and Open-CLIP all of which we will refer to in this work as CLIP models (Jia et al., 2021; Pham et al., 2023; Zhai et al., 2022b; Ilharco et al., 2021). CLIP models generally come in 3 canonical sizes of vision transformer: ViT-B/32, ViT-B/16 and ViT-L/14; since then, the open source community has extended these to 3 larger variants ViT-H/14, ViT-g/14 and ViT-G/14 (Dosovitskiy et al., 2020; Zhai et al., 2022a). Generally the larger models exhibit better zero-shot generalization and transfer properties. CLIP models have been trained on a variety of datasets from OpenAI's WiT, Google's WebLI and JFT-3B, LAION, COYO and DataComp-1B.

Prior work has also studied how to fine-tune CLIP models to improve performance in targeted directions. CLIP models can be fine-tuned on image classification tasks by using templates to transform labels to text (Fang et al., 2022; Goyal et al., 2022). Additionally, practitioners often use weight ensembling to preserve robustness properties of the pre-trained model while reaping the benefits of fine-tuning (Wortsman et al., 2022). We take advantage of these techniques in order to improve the filtering models we train in this work.

A.2 Dataset Construction

Prior to CLIP, datasets most commonly used in computer vision were *supervised* with human labels (Deng et al., 2009; Krizhevsky et al., 2009). Though these older dataset construction pipelines were quite intricate and did not scale beyond a few million examples, they share some similarity with current constructions. Classical datasets such as ImageNet and CIFAR started with a large *roughly curated* pool of images paired with metadata, and used humans to either label or filter the data.

Modern dataset pipelines have a similar procedure but at a much higher scale. The initial pool of images can contain up to 100 billion images, and the dataset filtering is purely automated, often with a set of rules and heuristic filtering stages (Jia et al., 2021). Past work in natural language processing has used binary filters as an initial step to remove low quality documents (Wenzek et al., 2019; Brown et al., 2020), but contain multiple components to their filtering pipelines.

One of the first publicly available web-scale image-text datasets is LAION. LAION-400M and LAION-2B were constructed by collecting image-text pairs from Common Crawl, filtering by English, and keeping pairs whose image and text are well *aligned*. This alignment is performed using a procedure known as *CLIP filtering*, which uses an existing image-text model (in LAION's case OpenAI CLIP ViT-B/32), and removes samples whose cosine similarity between image and text are below some threshold. We show pseudocode of the basic CLIP filtering operation below.

```
def clip_filter(image, text, threshold=0.3):
    # compute image and text representations
    image_features = clip.encode_image(image_input)
    text_features = clip.encode_text(text_input)
    # compute alignment
    dot_product = image_features.T @ text_features
    norm_a = image_features.norm()
    norm_b = text_features.norm()
    similarity = dot_product / (norm_a * norm_b)
    # filter by alignment
    return similarity > threshold
```

While CLIP filtering is convenient it is dependent on a existing trained CLIP model, and perhaps limited on the top-line performance of any model trained using it as a filter. For example, despite LAION-2B being five times larger than OpenAI's dataset, models trained on it could only match OpenAI's ImageNet zero-shot performance with a significantly larger compute budget.

To better facilitate the study of image-text datasets, researchers created the DataComp benchmark (Gadre et al., 2023). The benchmark provides 12.8 billion image-text pairs from Common Crawl so that researchers can study the effect of various data filtering techniques. DataComp fixes the computational budget used to train the resulting models, fixing the compute budget of the largest scale to match the cost of training OpenAI's ViT-L/14 CLIP model. These models are then evaluated on a suite of 38 downstream tasks, which includes ImageNet and distribution shifts, VTAB, and retrieval tasks. We use this benchmark as our primary method of evaluating the datasets created by our data filtering networks.

The authors of DataComp also released a baseline dataset, DataComp-1B (DC-1B) that improved upon LAION-5B, by combining CLIP filtering with an ImageNet based clustering approach to improve dataset quality on a variety of benchmarks. However this dataset still relies on the OpenAI CLIP model for CLIP filtering and imposes a costly ImageNet specific clustering step in the pipeline.

Recent work (Xu et al., 2023) has demystified the CLIP dataset construction process and demonstracted high quality dataset construction is possible by simple keyword based sampling and global balancing. While their work does create competitive datasets, the reliance on sampling heuristics from the original CLIP paper (Radford et al., 2021) allows for accurate dataset reproduction, our work focuses on *improving* model performance using dataset construction.

B Data Filtering Networks Evaluation Setup

With these definitions in place, we now address how we evaluate DFNs. In our context, the quality of a DFN is determined by the strength of models it can induce. We build on the evaluation framework proposed by DataComp (Gadre et al., 2023). DataComp provides a multi-scale evaluation framework for datasets by measuring CLIP model zero-shot performance. It provides 4 nested unfiltered image-text pair pools of increasing size. In this work, we use the medium (128M datapoints), large (1.28B datapoints) and xlarge(12.8B datapoints) pools. We also follow the DataComp guidelines of model hyperparameters for each of these pools, which are ViT-B/32 for medium, ViT-B/16 for large and ViT-L/14 for XL. Exact hyperparameters can be found in Table 4. We additionally expand our DFN to a larger pool of 42B images by combining 30B non-DataComp web-scraped images with the DataComp XL pool. We denote the dataset induced using this pool and our DFN as DFN-5B, which we use to train a ViT-H/14 model.

For evaluation we use 38 zero-shot classification and retrieval tasks in the DataComp benchmark. We denote the average performance on these benchmarks simply as "Average" performance, but we also track various subsets: ImageNet performance (IN), ImageNet distribution shift performance (IN shifts), Visual Task Adapation Benchmark (VTAB), Retrieval performance (COCO, Flickr, WinoGAViL).

Our actual training runs on both Nvidia A100s and TPU v4s. We use OpenClip and AXlearn to train our CLIP models on GPUs and TPUs respectively (Ilharco et al., 2021; Apple, 2023) .

C Filtering Model Type Ablation

We explore using filtering models beyond CLIP models. While DFNs can use any model that can be reduced to a binary function, intuitively it makes sense to use CLIP models. By filtering with a similarity score between the image and text, we encourage keeping samples where the image and text are aligned.

In order to verify this intuition we consider a few other options to produce a DFN. One is to train a binary classifier that can distinguish between ImageNet or CC12M data as positives and Common Crawl as negatives. We consider both ResNet (He et al., 2016) as well as frozen OpenAI CLIP embeddings for this filter. Another option is to use M3AE (Geng et al., 2022) trained on CC12M as a DFN that takes into account both images and text. We can use reconstruction loss as the filtering criterion, as it is a reasonable proxy for how similar samples are to the high-quality data used to train the filtering model.

The filtering performance of all these options, including CLIP models, are summarized in Table 2, where the CLIP model outperform the other backbones. A key difference between the binary classifier and CLIP filters is that the binary filter makes an explicit assumption on what qualifies as a good

Table 2: Filtering Performance of various filtering models, after filtering DataComp medium scale (ViT-B/32, 128M samples seen). We present results on ImageNet top-1 as well as “Average" set of tasks (see Sec. B for details.)

DFN Type	Filter Dataset	ImageNet	Average
No Filter Baseline	None	0.176	0.258
ResNet-34 Image Binary Filter	ImageNet	0.242	0.292
OpenAI ViT-B/32 Image Binary Filter	ImageNet	0.266	0.295
ResNet-34 Image Binary Filter	CC12M	0.203	0.257
OpenAI ViT-B/32 Image Binary Filter	CC12M	0.218	0.276
M3AE ViT-B/16	CC12M	0.237	0.297
CLIP ViT-B/32	CC12M	0.289	0.335

distribution, while CLIP filters are more flexible. Although the M3AE and CLIP filtering models both are trained on CC12M and examine both modalities, M3AE performs much worse, potentially due to a combination of CLIP encouraging image-text alignment and the difficulty of text reconstruction from just CC12M. We conclude that CLIP models are the most practical and performant models for image-text DFNs.

D Effect of Data Quality on Model versus Filtering Performance

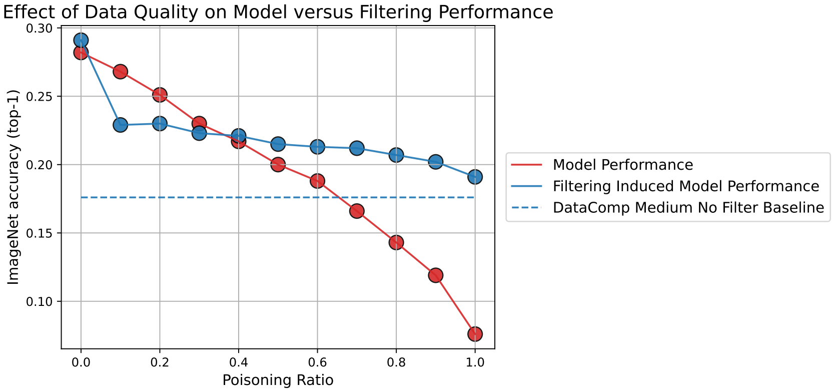

Figure 5: Data quality determines the filtering performance of models. We create these filter training datasets of various quality by having a set pool size of 10 million samples, and interpolating between CC-12M (high quality) and CommonPool (low quality). We then train models induced by the DFN filtering DataComp medium.

E Better Datasets Beyond Vision Tasks: VQA

Table 3: Performance of BLIP-2 variants with different visual encoder training datasets. The DFN-2B trained ViT-L provides consistent improvements across multiple zero-shot VQA tasks.

Visual Encoder Training Dataset	Architecture	VQAv2 Acc. (%)	GQA Acc. (%)	OKVQ Acc. (%)
OAI-WIT-400M	ViT-L	45.5	30.0	19.1
DFN-2B	ViT-L	48.3	31.3	21.9
LAION-2B	ViT-g	48.7	31.1	24.5

Just like how machine learning models would ideally generalize across many tasks, we would also like our datasets to generalize across diverse tasks. We show that our datasets not only lead to better models when evaluated on vision tasks, but also lead to better visual question answering (VQA) models. We train a BLIP2 model (Li et al., 2023a) which takes as input a CLIP visual encoder and is trained for zero-shot VQA on COCO and Visual Genome, to measure zero-shot VQA performance on

VQVA2, GQA, and OKVQA (Goyal et al., 2017; Hudson & Manning, 2019; Marino et al., 2019). In Table 3, we compare the performance on BLIP2 between the standard OpenAI ViT-L visual encoder and the ViT-L trained on DFN-2B. The DFN-2B model consistently outperforms the OpenAI ViT-L encoder and is competitive with a much larger EVA ViT-g model trained on LAION-2B[4]

F Further Discussion

The simplicity of the data filtering network pipeline makes it a flexible tool to integrate into existing workflows. As DFNs operates on individual samples, this approach scales linearly with candidate pool size, enabling the creation of datasets orders of magnitude larger than those that we introduce in this work. Additionally, the DFNs we train in this work are relatively small neural networks which allows for filtering to be directly integrated into training procedures of much larger networks for minimal marginal cost. DFNs can then filter batches of raw data that are then trained on, reducing the need for complex data pre-processing procedures.

As useful as DFNs are in building performant models in this work, there are still many unanswered questions to address in future work. We still do not know exactly how to optimize directly for dataset quality, and thus opt for weak proxies such as alignment. It is not even clear what that proxy would be for other domains where DFNs could be applied such as speech, text or video data. We hope that these open questions and the bridge DFNs build between modeling work and dataset work can lead to fruitful new avenues of research.

G Training Hyperparameters

Table 4: We follow the hyperparameter settings of the DataComp paper for the medium, large and xlarge pool

Dataset	Model	Pool size and # seen samples	Batch Size	Max LR	Weight Decay	Warmup	Beta2
DataComp-medium	ViT-B/32	128M	4096	5e-4	0.2	500	-
DataComp-large	ViT-B/16	1.28B	8192	5e-4	0.2	500	-
DataComp-xlarge	ViT-L/14	12.8B	90112	1e-3	0.2	10000	0.95
DFN-5B-pool	ViT-H/14	39B	79872	2e-3	0.1	10000	0.95

H DFN Hyperparameters

DFNs trained for ablations use DataComp large scale hyperparameters with a ViT-B/32 instead of a ViT-B/16. Final DFNs that induce DC-2B train for 5.12B samples, 16,384 batch size, and 2,000 steps of warmup.

I Robustness of Using ImageNet at Filtering vs. Training Time

Table 5: Fine-tuning a DFN on ImageNet induces datasets with nice robustness properties that are lost when directly training on ImageNet. Ran at DataComp large scale (ViT-B/16, 1.28B samples).

Dataset	IN	IN-V2	ObjectNet	IN-Sketch	IN-R	IN-A	VTAB
Baseline DFN	0.624	0.547	0.511	0.510	0.724	0.257	0.551
Baseline DFN FT on ImageNet	0.678	0.594	0.536	0.536	0.743	0.284	0.555
Baseline DFN + IN	0.757	0.652	0.509	0.512	0.703	0.272	0.543

[4] EVA's ViT-g has an additional pre-training procedure trained on ImageNet-21k, COCO, Objects365 and Conceptual Captions 12M

J Full Experimental Evaluation & Model Release

Below we provide links to checkpoints and detailed evaluation results of models in Table 1 on each of the 38 DataComp evaluation datasets

Model Link	ImageNet	Average
DFN5B-CLIP-ViT-H-14-378	0.844	0.710
DFN5B-CLIP-ViT-H-14	0.834	0.698
DFN2B-CLIP-ViT-L-14	0.814	0.669
DFN2B-CLIP-ViT-B-16	0.762	0.609

Table 6: Links to checkpoints and detailed evaluation results

K Figures measuring average performance instead of ImageNet

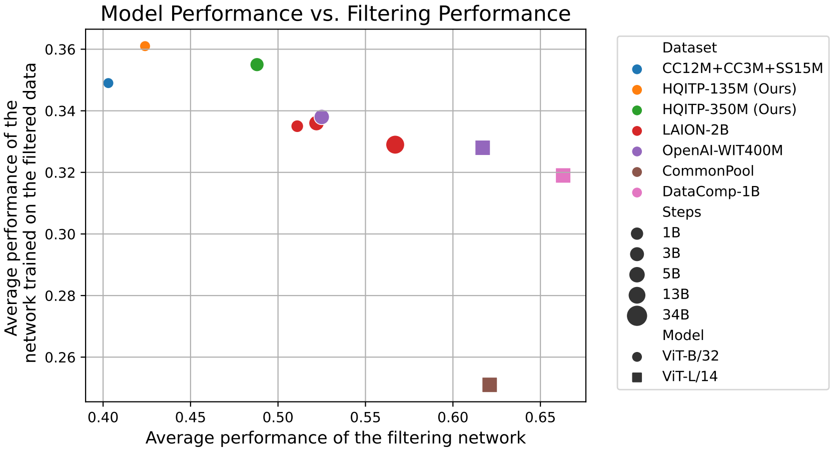

Figure 6: Average accuracy version of Figure 2.

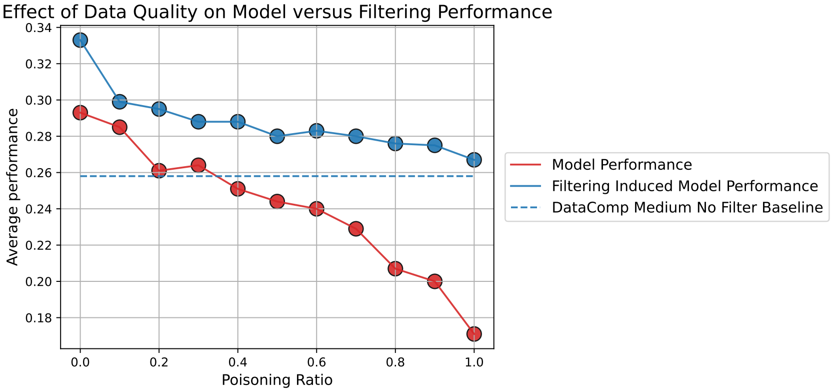

Figure 7: Average accuracy version of Figure 5.

L Log-Scale plot of Figure 1

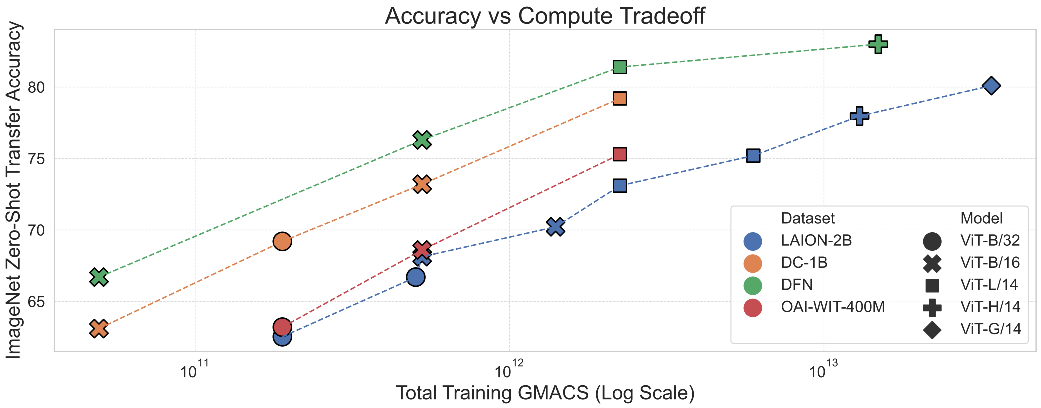

Figure 8: Compute scaling behavior of training CLIP models on various datasets (log scale)

M Additional Tables

Table 7: Standard interventions used to improve models can be used on DFNs to induce stronger datasets, leading to better models. DFNs are used to filter and train DataComp large (ViT-B/16, 1.28B samples seen).

Intervention		IN	IN Shifts	VTAB	Retrieval	Average
Augmentation	✗	0.620	0.493	0.534	0.515	0.536
	✓	0.626	0.501	0.534	0.516	0.542
Samples Seen / Batch Size	2.56B / 4096	0.626	0.506	0.536	0.511	0.545
	5.12B / 8192	0.624	0.508	0.551	0.517	0.550
Fine-tune	✗	0.624	0.508	0.551	0.517	0.550
	✓	0.678	0.540	0.555	0.534	0.560
OAI-Init	✗	0.674	0.535	0.533	0.529	0.548
	✓	0.678	0.540	0.555	0.534	0.560

Table 8: We can produce high-quality DFNs completely from scratch. Specifically, we do not use any OpenAI CLIP models for results in this table. We also use HQITP-135M for DFN training, a subset of HQITP-350M that we use in the rest of the paper

Dataset	DataComp Scale	IN	IN Shifts	VTAB	Retrieval	Average
Induced by DFN HQITP-135M with FT on IN1k, no OAI-Init, no Aug., samples 1B, BS 4096	xlarge	0.805	0.665	0.641	0.639	0.663

Table 9: High-quality data is best used to train the filtering model rather than the end model. Training DFNs with HQITP-350M induces a dataset that outperforms the dataset induced by a worse DFN combined with HQITP-350M.

Dataset	Model	IN	IN Shifts	VTAB	Retrieval	Average
OAI ViT-B/32 Induced Dataset + HQITP-350M	ViT-B/16	0.706	0.572	0.582	0.575	0.596
DFN without FT Induced Dataset	ViT-B/16	0.729	0.599	0.604	0.597	0.612
DFN-2B	ViT-B/16	0.762	0.623	0.598	0.611	0.609
OAI ViT-B/32 Induced Dataset + HQITP-350M	ViT-L/14	0.774	0.654	0.643	0.616	0.654
DFN without FT Induced Dataset	ViT-L/14	0.787	0.670	0.654	0.631	0.666
DFN-2B	ViT-L/14	0.814	0.688	0.656	0.649	0.669
DFN-2B + HQITP-350M	ViT-L/14	0.813	0.691	0.662	0.656	0.670

Table 10: DFNs are trained with a ViT-B/32, then used to filter DataComp pools. Conceptual 12M, Conceptual Captions 3M, and Shutterstock 15M are publicly available datasets, demonstrating that large-scale high-quality datasets can be constructed with only publicly available resources.

DFN Training Data	DataComp Scale	IN	IN Shifts	VTAB	Retrieval	Average
CC12M + CC3M + SS15M	medium	0.307	0.253	0.359	0.274	0.349
OpenAI WIT-400M	medium	0.285	0.240	0.355	0.253	0.338
CC12M + CC3M + SS15M	large	0.591	0.481	0.522	0.503	0.532
OpenAI WIT-400M	large	0.578	0.466	0.525	0.475	0.527
CC12M + CC3M + SS15M	xlarge	0.770	0.656	0.663	0.624	0.658
OpenAI WIT-400M	xlarge	0.764	0.640	0.628	0.599	0.638