In Search of Forgotten Domain Generalization

Prasanna Mayilvahanan [*1 2] **Roland S. Zimmermann** [*1] **Thaddäus Wiedemer** [1] **Evgenia Rusak** [2 1] **Attila Juhos** [1]
Mattias Bethge [2] **Wieland Brendel** [1 3]

Abstract

Out-of-Domain (OOD) generalization is the ability of a model trained on one or more domains to generalize to unseen domains. In the ImageNet era of computer vision, evaluation sets for measuring a model's OOD performance were designed to be strictly OOD with respect to their style. However, the emergence of foundation models and expansive web-scale datasets has obfuscated this evaluation process, as datasets cover a broad range of domains and risk test data contamination. In search of the forgotten domain generalization, we create large-scale datasets subsampled from LAION—LAION-Natural and LAION-Rendition—that are strictly OOD to corresponding ImageNet and DomainNet test sets in terms of style. By training CLIP models on these datasets, we find that OOD generalization challenges from the ImageNet era still prevail. Furthermore, through a systematic exploration of combining natural and stylistic datasets at varying proportions, we identify optimal ratios for model generalization across several style domains. Our datasets and results re-enable meaningful assessment of OOD robustness at scale—a crucial prerequisite for improving model robustness. Overall, we make the sobering point that large-scale data merely obscures the issue of OOD generalization, which remains an unsolved problem.

1. Introduction

Foundation models have revolutionized our world, demonstrating remarkable capabilities in solving grade school math problems, writing creative essays, generating stunning images, and comprehending visual content. One notable example is CLIP (Radford et al., 2021), a vision-language model pre-trained on a vast dataset of image-text pairs, which forms the backbone of numerous other foundation models. CLIP has achieved unprecedented performance in various benchmarks across many domains—a stark difference to models in the ImageNet era, which struggled to generalize to unseen domains. This raises an important question:

Does CLIP solve out-of-domain generalization?

Out-of-domain (OOD) generalization refers to a model's ability to perform well on data from domains other than its training (or *source*) domain. A *domain* is usually not rigorously defined and rather arises from collecting data in different contexts or environments. Nevertheless, some domains like the domain of *natural* images or the domain of *renditions* are delineated sufficiently clearly to enable the collection of datasets like ImageNet-Sketch (Wang et al., 2019), ImageNet-R (Hendrycks et al., 2020), or DomainNet (Peng et al., 2019) for rigorous evaluation.

CLIP's impressive performance and generalization ability is primarily attributed to its extensive web-scale training set (Fang et al., 2022). Despite the large diversity of *natural* images in the training set, CLIP is likely to learn robust representations through exposure to many test domains during training. Indeed, Mayilvahanan et al. (2024) showed that CLIP's training distribution contains exact or near duplicates of all commonly used OOD datasets but were also able to demonstrate that CLIP's generalization performance remains high when correcting for this contamination. However, their analysis was only concerned with contamination on a data set level and failed to account for entire data domains. For example, even after their correction many *rendition* images remain in the training distribution (refer to Tab. 9). It is therefore unclear if CLIP will generalize to domain shifts if all datapoints from that domain are removed. We address this question with the following contributions:

- We develop a domain classifier that effectively distinguishes between *natural* images and *renditions*. We achieve this by labeling 19 000 random data points from LAION-400M for training and 6000 datapoints each from ImageNet and DomainNet test sets for evaluation.
- By applying the domain classifier to a deduplicated ver-

[*]Equal contribution [1]MPI for Intelligent Systems, Tübingen AI Center [2]University of Tübingen [3]ELLIS Institute Tübingen.
Project website: brendel-group.github.io/forgotten-domain-generalization

Published at ICML 2024 Workshop on Foundation Models in the Wild. Copyright 2024 by the author(s).

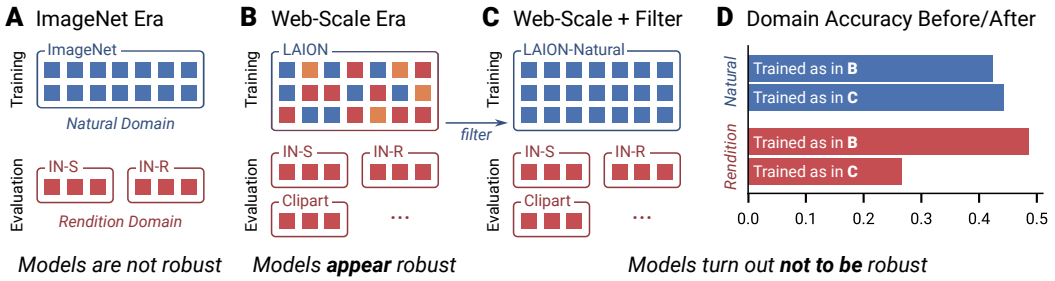

Figure 1: **Evaluated correctly, CLIP does not generalize across domains**. **A** Models used to be trained on a single domain like ***natural*** images from ImageNet (Russakovsky et al., 2015) and evaluated for out-of-domain (OOD) generalization on a different domain like ***renditions*** from test sets such as ImageNet-R (Hendrycks et al., 2020), ImageNet-Sketch (Wang et al., 2019). **B** Today, large foundation models like CLIP (Radford et al., 2021) are trained on web-scale datasets such as LAION-400M (Schuhmann et al., 2021) containing images from many domains. Tested on a specific domain like renditions, CLIP exhibits unprecedented performance and appears robust. **C** We subsample from a deduplicated LAION-400M (Abbas et al., 2023) to obtain LAION-Natural, web-scale data set containing only natural images, which re-enables a meaningful assessment of CLIP's generalization performance to renditions. **D** CLIP trained on LAION-Natural performs noticeably poorer on renditions, demonstrating that CLIP does not solve OOD generalization. The models are evaluated on refined test datasets containing samples only from their intended domains.

sion of LAION-400M, we create two datasets: LAION-Natural, containing 57 million natural images, and LAION-Rendition, with 16 million renditions of objects and scenes. Additionally, we use the domain classifier to refine common OOD benchmarks by removing a small number of samples from an incorrect domain.

- Via our proposed LAION-Natural dataset, we demonstrate that CLIP trained on a single domain performs significantly worse on naturally-occurring domain shifts (see 1 for a summary). This indicates that CLIP's strong performance is due to domain-contamination of the training data, rather than an inherent ability to generalize OOD.

2. Abridged Related Work

On gauging the OOD generalization performance of CLIP, Mayilvahanan et al. (2024) remove images that are *highly similar* to the test sets to show that data contamination and high perceptual similarity between training and test data does not explain generalization performance. While their data pruning technique removes some samples from LAION-400M that are somehow *close* to the test datapoints they give no guarantee that all images of a given domain were removed. We refer the reader to Sec. B for a thorough literature review.

3. Building a Domain Classifier

Our work hinges on filtering out datapoints that belong to specific domains from web-scale datasets. There is no precise definition for what constitutes a *domain* in general. Still, the community has come to agree on an implicit demarcation of the *natural* image and *renditions* domains by virtue of ImageNet compared to ImageNet-Sketch and ImageNet-R as well as DomainNet-Real compared to DomainNet-Sketch, -Quickdraw, -Infograph, -Clipart, and -Painting. Derived from the overall quality of an image, there is an intuitive, texture-centric notion of style us humans use which we adopt in this work. Further, we borrow methods from prior work that successfully classify images into different domains. We defer the reader to Sec. D for a thorough description of how we train and test our domain classifiers.

3.1. Domain Composition of LAION-200M

We now deploy the chosen classifiers from Sec. 3 and label each sample in LAION-200M as *natural*, *rendition*, or *ambiguous*. We apply the classifiers with their strict thresholds at 98 % validation precision which yields a strong lower bound for the number of samples in each domain, as well as with their default thresholds which yields a more rounded estimate. From Tab. 9, it is clear that the LAION-200M contains a considerable portion of strictly stylistic images (with a lower bound of 7.90 % corresponding to 16 million images), and potentially many more images with some rendition elements are contained in the ambiguous group. We detail the domain composition of ImageNet-Train and datasets from (Mayilvahanan et al., 2024) in Sec. D.5.

3.2. Creating Single-Domain Datasets

To measure the true OOD performance of CLIP, we need to create a large dataset with only natural examples. We now use our trained domain classifiers at 98 % validation-precision to subsample LAION-200M. We obtain LAION-

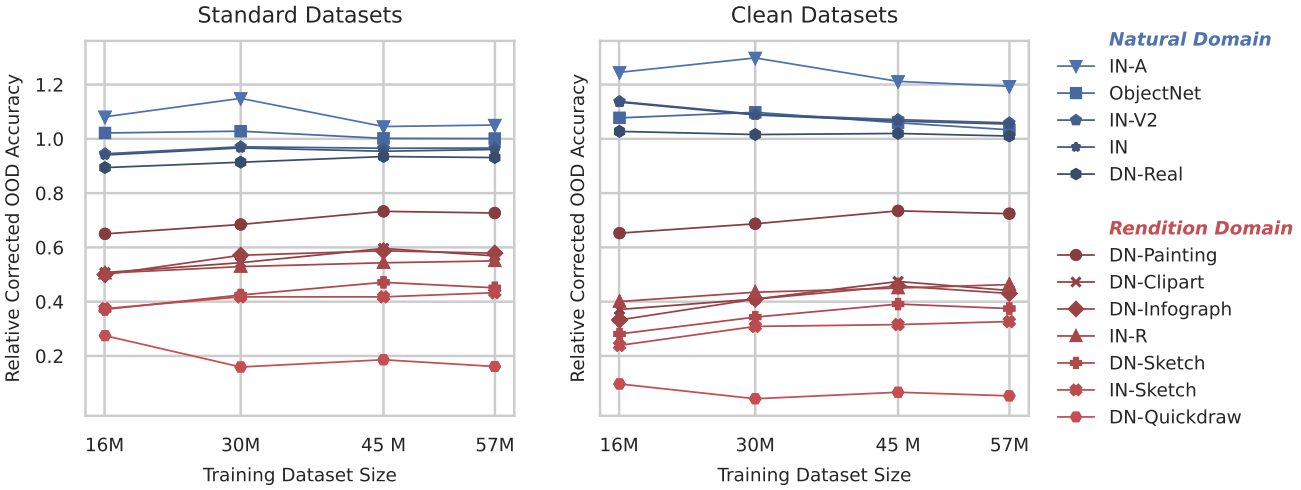

Figure 2: **Across scales, CLIP fails to generalize to unseen domains**. The *relative corrected OOD accuracy* shows performance losses or gains of a CLIP model trained exclusively on the ***natural domain*** via LAION-Natural to a CLIP model trained on a domain-contaminated dataset like LAION-200M. We evaluate on the original ImageNet and DomainNet test sets (left) and our cleaned versions of them (right, see Sec. 3.2 Without samples from the ***rendition*** domain, CLIP's domain generalization ability suffers significantly and consistently across scales.

Table 1: **Domain composition of LAION-200M.** We apply our *natural* and *rendition* domain classifiers with their strict thresholds at 98 % validation precision to get a lower bound of samples from each domain and with their default thresholds to obtain a more balanced estimate. Irrespective of the thresholding, LAION-200M still contains a large amount of renditions.

Classifier Precision		**% Samples**		
Natural	*Rendition*	*Natural*	*Ambiguous*	*Rendition*
0.79	0.77	60.74	25.41	13.86
0.98	0.98	28.40	63.70	7.90

Natural with roughly 57 million samples and LAION-Rendition with roughly 16 million samples. Figure 7 shows random samples from both datasets, more samples are shown in Figs. 19 and 20. We also deploy the domain classifiers on the ImageNet and DomainNet test sets to remove the domain-contamination reported above. The exact number of datapoints and the number of classes for each test set are detailed in Tab. 11. These datasets enable us to fairly assess CLIP's domain generalization performance in the following sections.

4. Measuring CLIP's OOD performance

For all our experiments, we train CLIP ViT-B/32 (Dosovitskiy et al., 2020) from scratch for 32 epochs with a batch size of 16 384 on one node with either four or eight A100 GPUs (training takes several days, depending on dataset size). We use the implementation provided by Ilharco et al. (2021) and stick to their hyperparameters. We first train CLIP on the 57 M LAION-Natural and random subsets of it with 45 M, 30 M, and 16 M samples. We compare the classification accuracy of these models to that of CLIP models trained on random subsets of LAION-200M of the same sizes by reporting the accuracy ratio, which we refer to as *relative corrected OOD accuracy*. We measure this quantity on the original ImageNet and DomainNet test sets and their cleaned versions (see Sec. 3.2). Fig. 2 summarizes the results.

Across the board, we find that the relative corrected OOD accuracy on the clean datasets is around or above 1.0 for *natural* test sets, but drops to around 0.4 for most *rendition* test sets. This demonstrates that without domain-contamination of the training distribution, CLIP does not generalize across domains nearly as effectively as previously assumed. Notably, the relative corrected OOD accuracy is very consistent across dataset scales, allowing us to conjecture that this result holds also for CLIP models trained on much larger data sizes. To further reinforce this observation, we build LAION-Mix-nM by replacing n million samples from LAION-Natural with samples from LAION-Rendition. We show in Tab. 2 that adding 13 or 16 million renditions has little effect on performance on the *natural* domain, but greatly improves performance on the *rendition* domain, highlighting the effect of domain-contamination.

To put the corrected OOD accuracy in context, we evaluate effective robustness (Fang et al., 2022; Taori et al., 2020)

Table 2: **Performance on the *rendition* domain is driven by renditions in the training data**. We compare CLIP trained without renditions on LAION-Natural to CLIP trained on datasets of the same size with renditions: LAION-Mix-nM contains n million renditions, LAION-Rand is a random subset of LAION-200M with an estimated fraction of 7.9-13.86 % renditions (see Tab. 9). Training with renditions greatly impacts performance on the ***rendition*** domain.

	Standard Datasets top-1 Acc.		**Clean Datasets** top-1 Acc.	
Dataset	***Natural***	***Rendition***	***Natural***	***Rendition***
LAION-Natural	36.88 %	21.98 %	39.72 %	18.75 %
LAION-Mix-12M	37.28 %	40.48 %	38.97 %	43.09 %
LAION-Mix-16M	36.92 %	41.46 %	38.58 %	41.46 %
LAION-Rand-57M	37.63 %	40.66 %	36.99 %	41.32 %

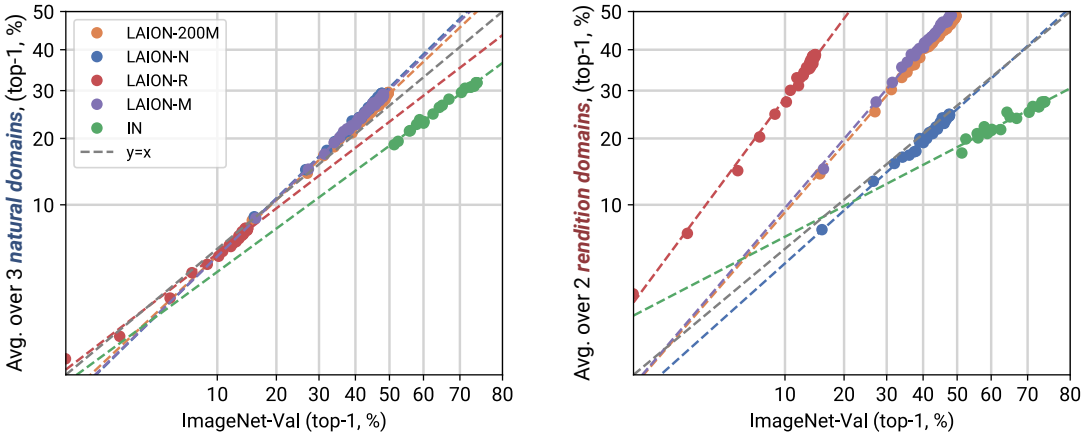

Figure 3: **CLIP's effective robustness to renditions is driven by domain-contamination**. We evaluate effective robustness (Fang et al., 2022; Taori et al., 2020) for models trained on different LAION-200M subsets. Most notably, CLIP trained on LAION-Natural matches the effective robustness of a LAION-200M-trained CLIP on the ***natural*** domain (left), but has significantly lower effective robustness on the ***rendition*** domain, indicating that CLIP requires rendition samples in its training distribution to perform well on this domain.

on the *natural* and *rendition* domain. To this end, Fig. 3 shows the top-1 classification accuracy of multiple CLIP models trained on LAION-200M, LAION-Natural, LAION-Rendition, LAION-Mix-13M, and ResNets trained on ImageNet (see Sec. F for details). We include LAION-Mix-13M as opposed to LAION-Mix-16M since it matches the effective robustness results for LAION-200M most closely. As usual, models with the same training regimen lie on a line and the y-distance of a model to the ImageNet line indicates it's effective robustness. While all LAION-trained models achieve a similar effective robustness on the *natural* domain (Fig. 3 left), effective robustness on the *rendition* domain varies greatly and is notably lowest for LAION-Natural-trained models. Effective robustness plots on the individual datasets can be found in App. G. Together, the findings in this section demonstrate that CLIP's unprecedented OOD generalization performance is a direct result of the domain-contamination of its training distribution. We defer a detailed discussion of Comparison of LAION training to ImageNet-training, Short-cut Learning, Domain Classification and Ambiguous Datapoints to Appx. C.

5. Conclusion

With the emergence of models trained on enormous web-scale datasets containing abundant samples from seemingly all possible domains, the study of domain generalization mostly came to a halt. Hence, the question of how dataset scale actually effects the ability of models to generalize between domains remains mostly unanswered. Here, we try to answer this question thoroughly by fully controlling the domain of training samples models are trained on. By creating clean subsets of LAION containing either natural images or renditions, and by training models on various mixtures and dataset sizes, we show that the generalization performance of CLIP trained on only one domain drops to levels similar to what we observe for ImageNet-trained models. Hence, we conclude that the domain generalization problem remains unsolved even for very large-scale datasets.

Reproducibility Statement

We describe the methodology to create all of the datasets we use in Sec. 3.2, D.1, D.2. We also sketch the training details of all our models in Sec. D.3,4, F. This should be sufficient to reproduce all our datasets and experimental results. We aim to host our datasets and models shortly. The code to train the domain classifiers is available at https://anonymous.4open.science/r/clip-dg-68D1/.

Acknowledgements

This work was supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. WB acknowledges financial support via an Emmy Noether Grant funded by the German Research Foundation (DFG) under grant no. BR 6382/1-1 and via the Open Philantropy Foundation funded by the Good Ventures Foundation. WB is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645. This research utilized compute resources at the Tübingen Machine Learning Cloud, DFG FKZ INST 37/1057-1 FUGG. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting PM, RSZ, TW, ER, and AJ.

References

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.

Zechen Bai, Yuta Nakashima, and Noa Garcia. Explain me the painting: Multi-topic knowledgeable art description generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5422–5432, 2021.

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.

Wei-Ta Chu and Yi-Ling Wu. Image style classification based on learnt deep correlation features. *IEEE Transactions on Multimedia*, 20(9):2491–2502, 2018.

Benjamin Cohen-Wang, Joshua Vendrow, and Aleksander Madry. Ask your distribution shift if pre-training is right for you. *arXiv preprint arXiv:2403.00194*, 2024a.

Benjamin Cohen-Wang, Joshua Vendrow, and Aleksander Madry. Ask your distribution shift if pre-training is right for you, 2024b.

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.

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.

Noa Garcia and George Vogiatzis. How to read paintings: semantic art understanding with multi-modal retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style. *ArXiv*, abs/1508.06576, 2015. URL https://api.semanticscholar.org/CorpusID:13914930.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization, 2020.

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.

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.

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.

Akshay Joshi, Ankit Agrawal, and Sushmita Nair. Art style classification with self-trained ensemble of autoencoding transformations. *arXiv preprint arXiv:2012.03377*, 2020.

Zhuang Liu and Kaiming He. A decade's battle on dataset bias: Are we there yet? *arXiv preprint arXiv:2403.08632*, 2024.

Prasanna Mayilvahanan, Thaddäus Wiedemer, Evgenia Rusak, Matthias Bethge, and Wieland Brendel. Does clip's generalization performance mainly stem from high train-test similarity?, 2024.

Orfeas Menis-Mastromichalakis, Natasa Sofou, and Giorgos Stamou. Deep ensemble art style recognition. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. Quality not quantity: On the interaction between dataset design and robustness of clip. *Advances in Neural Information Processing Systems*, 35:21455–21469, 2022.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1406–1415, 2019.

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.

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.

Evgenia Rusak, Steffen Schneider, Peter Vincent Gehler, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Imagenet-d: A new challenging robustness dataset inspired by domain adaptation. In *ICML 2022 Shift Happens Workshop*, 2022. URL https://openreview.net/forum?id=LiC2vmzbpMO.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Catherine Sandoval, Elena Pirogova, and Margaret Lech. Two-stage deep learning approach to the classification of fine-art paintings. *IEEE Access*, 7:41770–41781, 2019.

Catherine Sandoval Rodriguez, Margaret Lech, and Elena Pirogova. Classification of style in fine-art paintings using transfer learning and weighted image patches. In *2018 12th International Conference on Signal Processing and Communication Systems (ICSPCS)*, pages 1–7. IEEE, 2018.

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.

Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models. *arXiv preprint arXiv:2404.01292*, 2024.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

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.

Sheng-Yu Wang, Alexei A Efros, Jun-Yan Zhu, and Richard Zhang. Evaluating data attribution for text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7192–7203, 2023.

Ross Wightman, Hugo Touvron, and Hervé Jégou. Resnet strikes back: An improved training procedure in timm. *arXiv e-prints*, pages arXiv–2110, 2021.

Yihao Xue, Siddharth Joshi, Dang Nguyen, and Baharan Mirzasoleiman. Understanding the robustness of multimodal contrastive learning to distribution shift, 2024.

A. Understanding Domain Mixtures

We now expand on the experiment from Tab. 2 to understand how different ratios of domains in the training data affect downstream performance, and whether this effect transfers across scales. To this end, we show performance on the *natural* and *rendition* domain for models trained on LAION-Mix of different proportions and scales in Fig. 4, left and middle. The possible mixing ratios at larger scales are limited by the size of LAION-Rendition (16 million images), but we can nonetheless observe that the optimal mixing ratio is consistent across scales. Interestingly, as we slowly increase the fraction of natural / renditions samples starting from purely renditions / natural datasets, the performance steeply increases on natural / renditions shifts while remaining stable on the other domain.

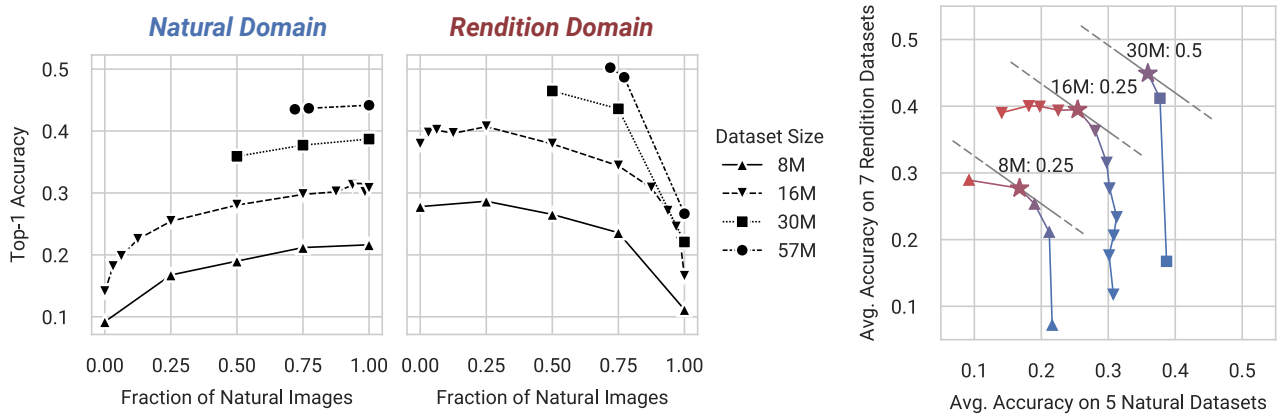

Figure 4: **Optimal data mixture transfers across scales**. We show the average accuracy on the ***natural*** and ***rendition*** domains for models trained with LAION-Mix of different absolute sizes and ratios. As expected, performance on each domain increases with the number of samples from that domain (left). The optimal mixing ratio for each scale is found at the intersection with the highest overall average accuracy iso-line. This ratio seems to be consistent across scales at 0.25, but our analysis is limited by the number of LAION-Rendition samples used for mixing (16 million images).

B. Related Work

Measuring the OOD Generalization of CLIP Models. We aim to understand the OOD generalization capabilities of CLIP from a data-centric viewpoint. While multi-modal training with rich language captions does seem to contribute to robustness against distribution shifts (Xue et al., 2024), Fang et al. (2022) demonstrated that the nature of CLIP's training distribution (as opposed to its mere size, its specific training objective, or natural language supervision) causes strong performance on various distribution shifts.

However, it is unclear what aspects of the data distribution drive the robustness gains. Mayilvahanan et al. (2024) remove images that are *highly similar* to the test sets to show that data contamination and high perceptual similarity between training and test data does not explain generalization performance. While their data pruning technique removes some samples from LAION-400M that lie outside the natural image domain, they do not address domain generalization: They only account for the part of a domain covered by existing test sets and give no guarantee that all images of a given domain were removed. In another line of work, Nguyen et al. (2022) discover that a model's effective robustness (Fang et al., 2022; Taori et al., 2020) on a test set interpolates when training data is compiled from various sources. While they combine different training datasets covering a mixture of domains, the authors have not analyzed the changes in effective robustness on a distributional similarity level. In this work, we take their analysis further and show that mixing two data sources similar to the test datasets interpolates the effective robustness. Our study's title is inspired by Gulrajani and Lopez-Paz (2020), who studied generalization from multiple distinct source domains. In contrast, we focus on generalization from single or mixed source domains to unseen domains.

Domain Classification. The primary goal of our work necessitates creating web-scale datasets of different domains. This entails building a robust domain classifier that can reliably distinguish *natural* images from *renditions*. This task can be regarded as classifying the style of an image, which Gatys et al. (2015) proposed to measure using Gram Matrices and which

has been widely explored since then (Sandoval et al., 2019; Menis-Mastromichalakis et al., 2020; Sandoval Rodriguez et al., 2018; Joshi et al., 2020; Garcia and Vogiatzis, 2018; Chu and Wu, 2018; Bai et al., 2021). More recently, Cohen-Wang et al. (2024a) use a fine-tuned CLIP model from OpenCLIP (Ilharco et al., 2021) to distinguish between ImageNet and shifted versions of ImageNet, such as ImageNet-Sketch, ImageNet-R, and ImageNet-V2 (Recht et al., 2019). Wang et al. (2023) and Somepalli et al. (2024) develop a dataset classifier using a backbone trained by self-supervised learning and classification through retrieval via a database. Liu and He (2024) report high performance when training image classifiers to distinguish between different large-scale and diverse classifiers.

C. Discussion

Comparison to ImageNet To the best of our knowledge, this work is the first to cleanly transfer the evaluation of domain generalization from the ImageNet era into the era of foundation models. While we do observe a somewhat similar generalization gap, it is difficult to quantitatively compare models trained on LAION and ImageNet for (at least) two reasons: For one, the distribution shifts from ImageNet-Val to LAION and ImageNet-Train are very different. Second, we are comparing a very noisy unsupervised learning method (CLIP + LAION) with a clean supervised learning method (CE + ImageNet), which is why LAION-trained models need $50\times$–$100\times$ more samples to reach the same ImageNet-Val accuracy as ImageNet-trained models.

Short-cut Learning Parts of the domain generalization gap of ImageNet models has been attributed to short-cut learning: models learn to solve a given task (like image classification) using features (like textures) that are misaligned to how humans solve the same task (like focusing on shape). The widely echoed notion of emergent abilities that models acquire at larger model and dataset sizes have fueled hopes that some parts of short-cut learning get mitigated simply by training on much larger and more diverse data. While some effect cannot be ruled out, our results also show that just adding more natural samples is unlikely to mitigate the effects of short-cut learning.

Domain Classification By labeling a small subset of images, we built a classifier that separates images into three categories: natural, artificial renditions, and ambiguous images. While accuracy and recall of our classifier was high, it should be noted that we did no further controls in potential biases (like favouring specific classes within domains) or the overall class distribution across all training and test sets. We also leave it to future work to study domain classifiers that distinguish between more domains, thus enabling a more fine-grained study of domain generalization.

Ambiguous Datapoints Our work does not examine the impact of ambiguous samples, i.e., samples exhibiting elements of both *natural* and *rendition*. To gain a clearer understanding of their effect, it is essential to distinguish between such ambiguous samples and those that exhibit neither. We anticipate that the former category significantly enhances performance and sample efficiency, while the latter does not contribute substantially. A more thorough analysis of this distinction is left for future work.

D. More Details on the Domain Classifier

We describe our labeling procedure based on this demarcation in Sec. D.1 and explore different ways to train a domain classifier on the resulting dataset in Sec. D.3. In Sec. D.5, we employ the best-performing classifier to analyze the composition of different training and test sets and finally use it to subsample LAION-Natural and LAION-Rendition in Sec. 3.2. For the remainder of this work, we substitute LAION-400M by LAION-200M, which we obtain by de-duplicating LAION-400M based on perceptual similarity as introduced by Abbas et al. (2023). They demonstrate that CLIP trained on LAION-200M obtains comparable downstream performance while greatly increasing data efficiency.

D.1. Labeling

LAION-200M contains diverse images from a multitude of sources. The images vary from naturally occurring to synthetically generated. We encourage the reader to glance at Fig. 19 to get a sense of the dataset and the difficulty of determining the domain of each image. As explained above, we aim to classify images belonging to the *natural* image or *rendition* domain. We also add an *ambiguous* class for images with elements of both domains and edge-cases.

We provide the human annotator with a comprehensive set of guidelines derived from analyzing the existing OOD test sets, which we outline in App. D.2. In general, we adopt a *texture*-centric approach to distinguish renditions of a scene or object

Figure 5: **Labeled *Natural*, *ambiguous*, and *rendition* samples from different data sets**. *Natural* images are photos or high-quality renders with minor filters that preserve *fine-grained textures*, while *renditions* are typically sketches, paintings, or graphics with *flat or simplified textures*. Images with elements of both, such as collages or natural images with large stylized elements, and images that mainly contain text are labelled as *ambiguous*.

from their natural depictions. That is, depictions where *fine-grained texture information* is preserved are generally considered *natural*, while depictions with *simplified or flat textures* are considered *renditions*. Fig. 5 illustrates this demarcation on samples from LAION-200M, ImageNet test sets and DomainNet test sets.

Overall, we label 19 000 random images from LAION-200M and 1000 images from each of the ImageNet and DomainNet distribution shifts (12 000 in total). Notably, almost all ImageNet and DomainNet test sets that are usually assumed to contain only images of a single domain exhibit some domain contamination. We discuss this in detail in Sec. D.5. Tab. 3 contains a detailed breakdown of labels for each data set. We show more samples grouped by domain for each data set in Figs. 22- 33.

D.2. Labeling

As mentioned in Sec. D.1, we take a *texture-centric* approach in domain labeling. We resolve further ambiguities with respect to labeling in the following way:

- Natural objects with watermark or text, infographs with natural objects, signs with human symbol (eg. walking signal), objects with common logos (eg. Nike), naturalistic books or movie covers, images that are retro / low resolution / blurry / grainy / or with fake background but with texture information preserved, graphically altered natural images with significant texture information, and real objects with fake backgrounds **are all classified as natural**.
- Stylistic: Infographs with stylized objects, stylized books or movie covers, retro / low resolution / blurry / grainy /graphically altered images with significant loss in texture information, stylized objects on plain or common natural background (eg. wall, bedsheet etc.) **are all classified as stylistic**.
- Ambiguous: Tattoos where hand / back is very visible, sculpture with real objects around, real images with distinct drawing of logos with objects, images that are retro / low resolution / blurry / grainy / or with fake background but with little texture information preserved **are all classified as ambiguous**.

To further ease the labeling procedure, we first build a rough binary classifier by fine-tuning CLIP ViT-L/14 with a linear readout to differentiate between some of the *natural* ImageNet and DomainNet test sets (namely, ImageNet-Val, ObjectNet (Barbu et al., 2019), ImageNet-V2 (Recht et al., 2019), ImageNet-A (Hendrycks et al., 2021), and DomainNet-Real) and *stylistic* test sets (namely, ImageNet-Sketch, ImageNet-R, DomainNet-Painting, DomainNet-Sketch, and DomainNet-Clipart). We use this classifier to roughly pre-label samples and provide the annotator with 25 images from the same group

at a time. This setup is shown in Fig. 6. The labeling was done by one labeler who labeled about 750-1000 images per hour. The labeler also did a checking of these labels by regrouping and going over them again. Below we visualize our labeling setup:

Final labeled images breakdown:

Table 3: **Number of labeled data points from several datasets and their domain-wise breakdown**. For training our domain classifier, we use the LAION-200M (Train), and LAION-200M (Val) for validation, and everything else to evaluate the final test performance.

Dataset	Natural	Stylistic	Ambiguous	Total
LAION-200M (Train)	7268	2978	2754	13000
LAION-200M (Val)	1000	1000	1000	3000
LAION-200M (Test)	1000	1000	1000	3000
ImageNet-A	974	7	19	1000
ObjectNet	917	2	81	1000
ImageNet-R	22	859	119	1000
ImageNet-Sketch	49	937	14	1000
ImageNet-V2	945	5	50	1000
ImageNet-Val	934	16	50	1000
DomainNet-Clipart	48	933	19	1000
DomainNet-Infograph	134	720	146	1000
DomainNet-Painting	101	795	104	1000
DomainNet-Quickdraw	0	1000	0	1000
DomainNet-Real	836	111	53	1000
DomainNet-Sketch	24	942	34	1000

D.3. Training and Choosing the Domain Classifier

With the domain-labeled dataset, we can train a domain classifier to partition all of LAION-200M into *natural* images, *renditions*, or *ambiguous* images. Since we aim to obtain datasets that contain only images from a single domain we need a domain classifier that is as precise as possible. To this end, we train classifiers on $13\,000$ labelled LAION-200M images, retaining 3000 samples each for a validation and test set. From the domain classification literature discussed in Sec. B, we evaluate four methods with publicly available code that we outline below. All methods build on CLIP ViT-L/14 pretrained on LAION-2B, which we choose for its balance between accuracy and inference speed.

Contrastive Style Descriptors (CSD) (Somepalli et al., 2024) fine-tune pre-trained backbones via multi-label supervised contrastive learning and self-supervised learning with only style-preserving augmentations (random flips, resize, rotation). The resulting final-layer embeddings serve as style descriptors: During inference, they find the k stylistically nearest neighbors in a database of labelled images (e.g., the training set) by computing pairwise embedding-similarities to the test images. An image is classified as belonging to a style if at least one of the k neighbors has that style. We can directly set up their method using the $13\,000$ labelled LAION-200M images as both the training set and the database for inference. From that, we obtain two binary classifiers, CSD-N (classifying natural vs. non-natural) and CSD-R (classifying renditions vs. non-renditions) that, together, can be used for our ternary classification.

Density Ratios (Cohen-Wang et al., 2024b) aim to estimate the probability that a given sample is drawn from a reference distribution p_{ref}. Since high dimensional density estimation is challenging, they build a classifier to distinguish between a reference and a shifted distribution and compute the density ratio $\frac{p_{\text{ref}}}{p_{\text{shifted}}}$ which they threshold at 0.2 to classify a given sample. We deploy their method unchanged to our task. We again obtain two binary classifiers, DR-N (classifying natural vs. non-natural) and DR-R (classifying renditions vs. non-renditions).

Centroid Embeddings Inspired by the baselines in (Somepalli et al., 2024), we implement a simple model (embedding model plus linear readout) where we take the pretrained CLIP ViT-L/14 as the embedding model and create a linear readout

Table 4: **We chose the best *natural* classfier and the best *rendition* classifier** amongst binary classifiers based on Contrastive Style Descriptors (CSD) (Somepalli et al., 2024) and Density Ratios (DR) (Cohen-Wang et al., 2024b) as well as ternary classifiers using a linear readout based on either each domain's centroid embedding (CE) or a fine-tuned CLIP (FT). All models use CLIP ViT-L/14 pretrained on LAION-2B. We report precision and recall on for the ***natural*** class (top) and ***rendition*** class (bottom) on ImageNet (IN) and DomainNet (DN) test sets and average performance across all test sets. Model hyperparameters are chosen for a validation precision of 98 % if possible. For each class, we select the classifier with the highest recall on the validation.

`cls=`***natural***	**Val**		**Test**		**IN-Val**		**IN-v2**		**IN-A**		**ON**		**DN-R**		***Average***	
Model	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
CSD-N `k=1`	0.61	0.85	0.58	0.85	0.96	0.93	0.97	0.92	0.98	0.91	0.93	0.94	0.92	0.88	0.85	0.90
CSD-R `k=23`	0.98	0.26	0.99	0.29	1.00	0.22	1.00	0.27	1.00	0.27	1.00	0.59	0.99	0.32	0.99	0.32
DR-N	0.98	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.21	0.00
DR-R	0.98	0.08	0.72	0.08	1.00	0.00	1.00	0.00	1.00	0.00	0.95	0.20	1.00	0.00	0.95	0.05
CE	0.98	0.35	0.89	0.33	0.95	0.02	1.00	0.04	1.00	0.02	0.99	0.16	0.99	0.11	0.97	0.15
FT	0.98	0.41	0.95	0.44	1.00	0.36	0.99	0.40	1.00	0.46	0.99	0.53	1.00	0.42	0.99	0.43

`cls=`***rendition***	**Val**		**Test**		**IN-R**		**IN-S**		**DN-S**		**DN-Q**		**DN-P**		**DN-C**		**DN-I**		***Average***	
Model	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
CSD-N `k=6`	0.98	0.26	0.99	0.24	1.00	0.20	1.00	0.18	1.00	0.25	0.00	0.00	1.00	0.24	1.00	0.22	0.98	0.34	0.88	0.21
CSD-R `k=1`	0.64	0.56	0.68	0.60	0.93	0.62	0.98	0.63	0.98	0.62	0.00	0.00	0.92	0.59	0.98	0.63	0.82	0.46	0.77	0.52
DR-N	0.98	0.20	0.98	0.23	1.00	0.29	1.00	0.20	1.00	0.27	1.00	0.01	1.00	0.28	1.00	0.28	0.98	0.11	0.99	0.21
DR-R	0.98	0.35	0.98	0.41	1.00	0.60	1.00	0.71	1.00	0.74	1.00	0.33	0.99	0.60	1.00	0.65	0.98	0.39	0.99	0.53
CE	0.98	0.11	0.99	0.12	0.99	0.43	1.00	0.39	1.00	0.30	1.00	0.09	0.98	0.47	1.00	0.38	1.00	0.01	0.99	0.26
FT	0.98	0.27	0.95	0.26	1.00	0.38	1.00	0.57	1.00	0.61	1.00	0.68	1.00	0.21	1.00	0.50	1.00	0.30	0.99	0.42

by comparing to the centroid embeddings for each domain. We use this as a ternary untrained nearest-neighbor classifier, dubbed CE.

Fine-Tuning We fine-tune the pretrained CLIP ViT-L/14 with a linear readout on the training dataset to obtain a ternary classifier, dubbed FT.

For the baselines (Cohen-Wang et al., 2024b; Somepalli et al., 2024), we simply use the training code detailed in their works and their public code. For the FT (Finetuning) model, as mentioned in Sec. D.3, we finetune a CLIP ViT-L/14 pretrained on LAION-2B with a linear readout. We finetune all models on 4 A100 GPUs, using a batch size of 256, weight decay of $5e-4$, using an SGD optimizer, with step scheduler (0.1 every 20 epochs), at a learning rate of 0.1, for 50 epochs. All models converge. Each model took about 2 A100 GPU hours to train, therefore all the models took around 30 A100 GPU hours. The storage requirement for these datasets were less than 100 GB memory.

We use the validation set to determine the two best classifiers, one for natural images and one for renditions. Since the domain classifier should maximize precision above all else, we set the confidence threshold for each model such that it achieves 98 % per-class precision. For CSD, we instead choose k to reach this precision. We then pick the classifier with the highest per-class recall to minimize the number of datapoints that are discarded when subsampling LAION-200M to build LAION-Natural and LAION-Rendition. We end up with FT, the fine-tuned ternary classifier, as our classifier for natural images, and DR-R, the binary classifier using density ratios as our rendition classifier. We use these classifiers for all subsequent experiments. Tab. 4 reports each model's precision and recall on the *natural* and *rendition* class across ImageNet and DomainNet test sets. For raw accuracy numbers of all models, which in general are high for most, please refer to Tabs. 5 and 6 in App. D.4.

D.4. Domain Classifier Performance without Precision Thresholding

In Sec.D.3 we only compute the precision and recall obtained from the threshold at which we get 98% precision on LAION-200M Val domain dataset. We here report the accuracy of these classifiers on these test sets at their own standard

precision of these models. We also train additional classifiers binary and ternary classifiers and by balancing the dataset sizes. To compare with the models from Cohen-Wang et al. (2024b), we train binary classifiers where we club natural with ambiguous and differentiate it from rendition (we name this FT-R), or we club rendition with ambiguous and differentiate it from natural (we name this FT-N). Further, we create several subsets for each of the ternary and the binary classification problem by balancing the number of datapoints in each class. We add the prefix '(balanced)' to these models.

Table 5: **Accuracy on each of the natural test sets on class natural without thresholding.** Some classifiers give the illusion of being good but have very low precision or recall(see Sec. D.3).

Model	(Val)	(Test)	IN-Val	IN-V2	IN-A	ON	DN-R	DN-I
FT	0.90	0.89	0.93	0.94	0.96	0.95	0.94	0.72
CE	0.75	0.78	0.80	0.84	0.86	0.95	0.81	0.19
FT-N	0.89	0.90	0.94	0.95	0.97	0.97	0.93	0.49
DR-N (balanced)	0.89	0.91	0.94	0.94	0.95	0.98	0.92	0.50
DR-R	0.98	0.97	0.99	0.99	1.00	1.00	0.97	0.90
FT (balanced)	0.78	0.82	0.84	0.86	0.86	0.88	0.83	0.46
FT-R	0.96	0.95	0.93	0.95	0.97	0.98	0.96	0.90
FT-N (balanced)	0.85	0.85	0.92	0.95	0.96	0.95	0.91	0.43
DR-R (balanced)	0.93	0.92	0.93	0.94	0.95	0.99	0.90	0.75
FT-R (balanced)	0.86	0.86	0.88	0.88	0.90	0.89	0.88	0.84
DR-N	0.93	0.92	0.94	0.95	0.94	0.99	0.92	0.76

Table 6: **Accuracy on each of the rendition test sets on class natural without thresholding.** Some classifiers give the illusion of being good but have very low precision or recall(see Sec. D.3).

Model	(Val)	(Test)	IN-R	IN-S	DN-S	DN-Q	DN-P	DN-C	DN-I
DR-R	0.77	0.80	0.93	0.98	0.98	0.96	0.92	0.93	0.88
FT (balanced)	0.78	0.88	0.82	0.94	0.94	0.91	0.80	0.85	0.77
FT	0.76	0.75	0.75	0.91	0.90	0.95	0.73	0.80	0.74
DR-N	0.89	0.92	0.99	0.99	0.99	0.98	0.97	0.97	0.94
FT-R	0.69	0.68	0.69	0.81	0.80	0.79	0.65	0.72	0.67
DR-N (balanced)	0.93	0.94	0.97	0.99	0.99	1.00	0.95	0.94	0.99
FT-R (balanced)	0.86	0.84	0.80	0.92	0.91	0.90	0.75	0.83	0.88
CE	0.61	0.62	0.95	0.90	0.89	0.96	0.95	0.93	0.32
DR-R (balanced)	0.90	0.93	0.99	0.99	0.99	0.99	0.98	0.97	0.96
FT-N	0.84	0.83	0.72	0.83	0.82	0.48	0.63	0.77	0.97
FT-N (balanced)	0.87	0.86	0.75	0.93	0.91	0.96	0.64	0.88	0.98

D.5. Analyzing the Domain Make-Up of Different Data Sets

Both ImageNet and DomainNet are web-scraped datasets that were refined through extensive human annotation. In contrast, LAION-400M is obtained purely through web scraping without subsequent human domain filtering. Since human annotators can make mistakes, and LAION-400M's domain composition is inherently unknown, we use our domain classifiers to understand it.

To this end, we deploy the chosen classifiers from Sec. 3 and label a sample *ambiguous* if the *natural* and *rendition* classifier disagree. We apply the classifiers both with their strict thresholds at $98\,\%$ validation precision which yields a strong lower bound for the number of samples in each domain, as well as with their default thresholds which yields a more rounded estimate. From Tab. 9, it is clear that the LAION-200M contains a considerable portion of strictly stylistic images (with a lower bound of $7.90\,\%$ corresponding to 16 million images), and potentially many more images with some rendition

Table 7: **Domain composition of training sets.** We apply our *natural* and *rendition* domain classifiers with their strict thresholds at 98 % validation precision to get a lower bound of samples from each domain and with their default thresholds to obtain a more balanced estimate. ImageNet-Train has a much smaller fraction of ***rendition*** samples than LAION-200M. We also note that 'combined-pruned', the training set from Mayilvahanan et al. (2024) that corrected for test set contamination still contains a large fraction of renditions.

		Classifier Precision				
Dataset	**# Samples**	*Natural*	*Rendition*	***Natural***	***Ambiguous***	***Rendition***
LAION-200M	199 663 250	0.79	0.77	60.74 %	25.41 %	13.86 %
		0.98	0.98	28.40 %	63.70 %	7.90 %
ImageNet-Train	1 281 167	0.79	0.77	89.20 %	9.62 %	1.18 %
		0.98	0.98	36.00 %	63.60 %	0.40 %
combined-pruned	187 471 515	0.79	0.77	62.98 %	25.18 %	11.83 %
		0.98	0.98	29.58 %	64.02 %	6.40 %

elements are contained in the ambiguous group. In contrast, for ImageNet, we find a much smaller fraction of renditions (at least $0.4\,\%$ of samples). We additionally observe that many evaluation datasets are considerably domain-contaminated (at least $5\,\%$ of samples stem from the opposite domain), especially ImageNet-R, DomainNet-Real, DomainNet-Clipart, DomainNet-Painting, and DomainNet-Infograph (refer to Tab. 8, App. D.6).

We also analyze the domain composition of datasets from Mayilvahanan et al. (2024), who created several subsets of LAION-200M that do not contain samples that are perceptually *highly similar* to ImageNet OOD test sets. These removed images are expected to be (near-) duplicates of test images in terms of both content and style. Their dataset *'combined-pruned'* is a subset of LAION-200M where highly similar images to ImageNet-Sketch, ImageNet-R, ImageNet-Val2, ImageNet-Val, ImageNet-A, and ObjectNet were pruned. In their work, it remained unclear whether pruning also effectively removed all images of the rendition domain, which we can now answer. Tab. 9 reveals that a considerable number of renditions remains in the pruned dataset (at least $6.4\,\%$ corresponding to around 11 million images). These remaining renditions might have played a significant role in the generalization performance of their CLIP models, especially on ImageNet-Sketch and ImageNet-R. As a result, CLIP's domain generalization performance is yet to be evaluated fairly.

D.6. Domain composition at different precision

We provide a detailed overview over the domain composition of datasets at standard precision in Table 8, and over the domain composition of datasets at 98% precision in Table 9.

D.7. On the Domain Composition of (Mayilvahanan et al., 2024)

Please find in Tab. 10 the exact number of rendition examples calculated by deploying our domain classifier on each the 3 datasets (pruned using rendition test sets) from Mayilvahanan et al. (2024). We see that at least 11-13M images are not pruned away from the datasets, therefore explaining the insignificant drop in performance.

D.8. Preparing clean datasets

In Sec. 3.2, we created several train and test sets from LAION-200M and ImageNet / DomainNet shifts respectively, by deploying our classifier at 98% precision. The exact number of samples and the number of (remaining) classes are in Tab. 11.

E. Notes on the CLIP Models

E.1. Resources spent

We train about 28 CLIP ViT-B/32 models on several subsets of LAION-200M. These models took about 8000 A100 GPU hours. We also needed about 18 TB of memory to store these datasets.

Table 8: **Domain composition of datasets at standard precision (without thresholding).** The first three columns show the fraction of samples in the original dataset classified as natural, stylistic, or ambiguous, respectively, while the latter column shows the dataset's total number of samples.

Dataset	Natural [%]	Stylistic [%]	Ambiguous [%]	Total
LAION-200M	60.74	13.86	25.41	199 663 250
ImageNet (Train)	89.2	1.18	9.62	1 281 167
ImageNet (Val)	89.1	1.18	9.72	50 000
ObjectNet	90.22	0.1	9.68	18 574
ImageNet-V2	88.49	1.38	10.13	10000
ImageNet-A	93.79	0.52	5.69	7 500
ImageNet-R	9.75	64.42	25.83	30 000
ImageNet-Sketch	3.69	85.34	10.97	50 889
DomainNet-Real	80.07	7.59	12.34	175 327
DomainNet-Quickdraw	1.35	93.27	5.38	172 500
DomainNet-Clipart	8.28	75.89	15.83	48 833
DomainNet-Painting	13.97	56.33	29.7	75 759
DomainNet-Sketch	3.1	84.18	12.71	70 386
DomainNet-Infograph	11.17	53.41	35.41	53 201

Table 9: **Domain composition of datasets at 98% precision.** The first three columns show the fraction of samples in the original dataset classified as natural, stylistic, or ambiguous, respectively, while the latter column shows the dataset's total number of samples.

Dataset	Natural [%]	Stylistic [%]	Ambiguous [%]	Total
LAION-200M	28.4	7.9	63.7	199 663 250
ImageNet (Train)	36.0	0.4	63.6	1 281 167
ImageNet (Val)	35.73	0.37	63.9	50 000
ObjectNet	50.32	0.0	49.68	18 574
ImageNet-V2	36.04	0.29	63.67	10000
ImageNet-A	43.25	0.16	56.59	7 500
ImageNet-R	3.56	52.82	43.61	30 000
ImageNet-Sketch	1.21	67.92	30.87	50 889
DomainNet-Real	34.31	3.98	61.71	175 327
DomainNet-Quickdraw	0.09	34.41	65.5	172 500
DomainNet-Clipart	3.46	62.53	34.01	48 833
DomainNet-Painting	5.3	47.55	47.15	75 759
DomainNet-Sketch	1.38	69.58	29.04	70 386
DomainNet-Infograph	1.59	28.11	70.3	53 201

E.2. Raw Accuracy Numbers of CLIP Trained on LAION-N vs LAION

In Sec. 4, in Fig. 2, we only reported the relative numbers. Here, in Fig. 8, 10, 9, 11, we report the actual numbers as a function of dataset size.

Table 10: **Number datapoints within the dataset vs number of datapoints pruned away in Mayilvahanan et al. (2024).**

Dataset	Size	Within	Pruned
sketch-pruned	191 481 491	24 016 047	3 654 180
r-pruned	194 088 525	24 304 991	3 365 236
combined-pruned	187 471 515	22 173 006	5 497 221
sketch-pruned (98% precision)	19 1481 491	13 266 999	2 482 751
r-pruned (98% precision)	194 088 525	13 338 759	2 410 991
combined-pruned (98% precision)	187 471 515	11 999 276	3 750 474

Table 11: **Clean datasets composition.** Obtained by deploying the domain classifiers from Sec.D.3 at 98% precision.

Dataset	Classes	Size
LAION-Natural	-	56 685 759
LAION-Stylistic	-	15 749 750
ImageNet-Val	985	17 864
ImageNet-V2	926	3 604
ImageNet-Sketch	991	34 564
ImageNet-R	200	15 847
ImageNet-A	197	3 244
ObjectNet	113	9 347
DomainNet-Real	339	60 148
DomainNet-Quickdraw	344	59 353
DomainNet-Infograph	345	14 957
DomainNet-Clipart	345	30 536
DomainNet-Sketch	344	48 974
DomainNet-Painting	345	36 020

F. Training ResNets on ImageNet

We deploy our natural domain classifier from Sec/ 3 at 90% precision (threshold obtain from LAION 13K Val set) on ImageNet-Train to obtain about 1M datapoints belonging to the natural domain (dubbed ImageNet-N). We create several datasets of smaller sizes subsampling from ImageNet-N. We also create randomly sampled datasets of similar sizes from the original ImageNet. We train ResNet-50 models on all of these datasets. We follow the training recipe A3 of Wightman et al. (2021) and train the models for 200 epochs. We then evaluate these models on standard test sets and clean test sets from Sec.3.2. The accuracies of ResNets trained on subsets of original ImageNet is used for the effective robustness plots in Sec. 4, G. Further, the comparison of accuracies between the models trained on subsets from ImageNet-N and ImageNet is in Fig. 12, 14, 13, 15. As such there is no significant performance difference anywhere, thus indicating that ImageNet does not have substantial domain leakage.

Figure 6: **Labeling setup.** By clicking on the image, the border changes to red, green, or blue, each representing natural, ambiguous, or rendition. By pressing the right or the left button the previous or next set of 25 images are rendered and the labels of the previous images are updated in a json file.

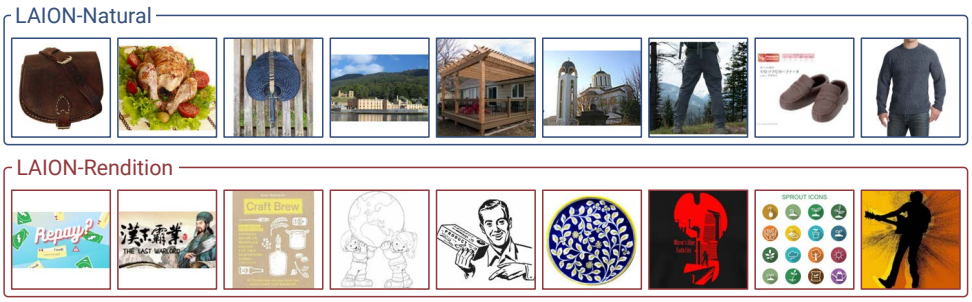

Figure 7: **Random samples from LAION-Natural and LAION-Rendition.**

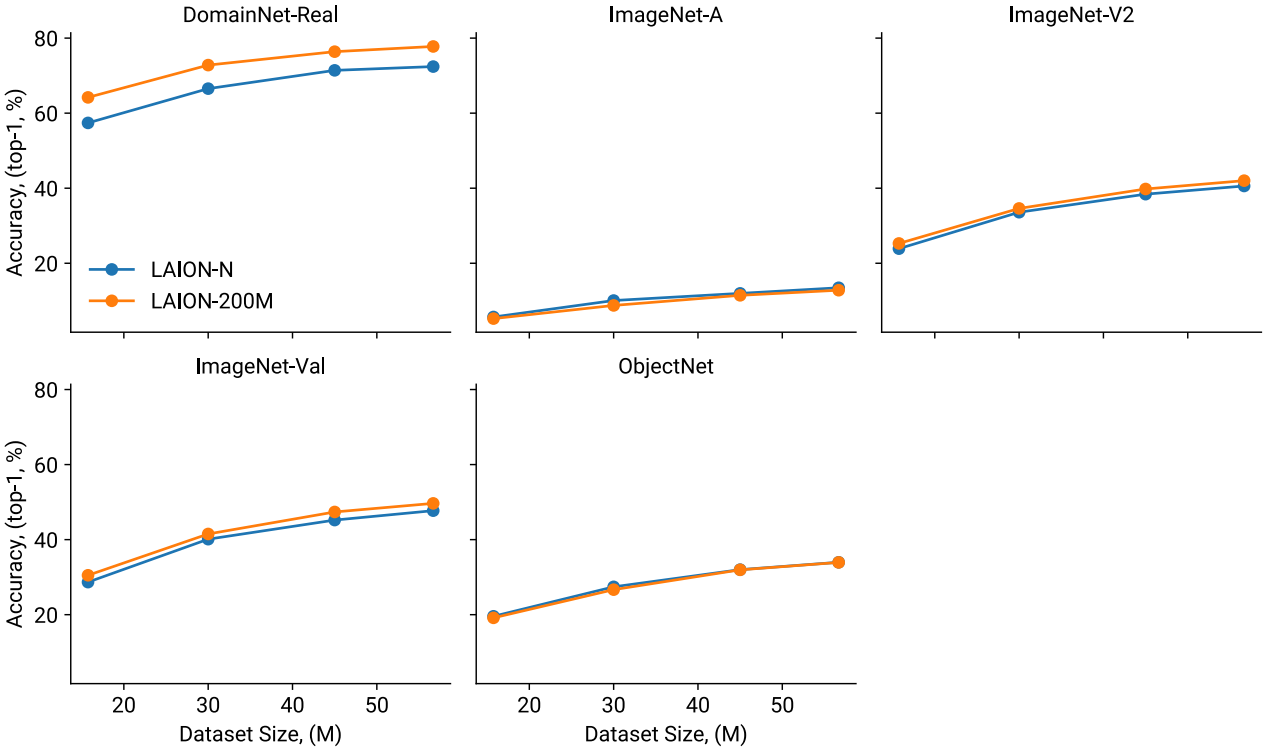

Figure 8: **CLIP trained on LAION v LAION-N performance on standard natural test sets.**

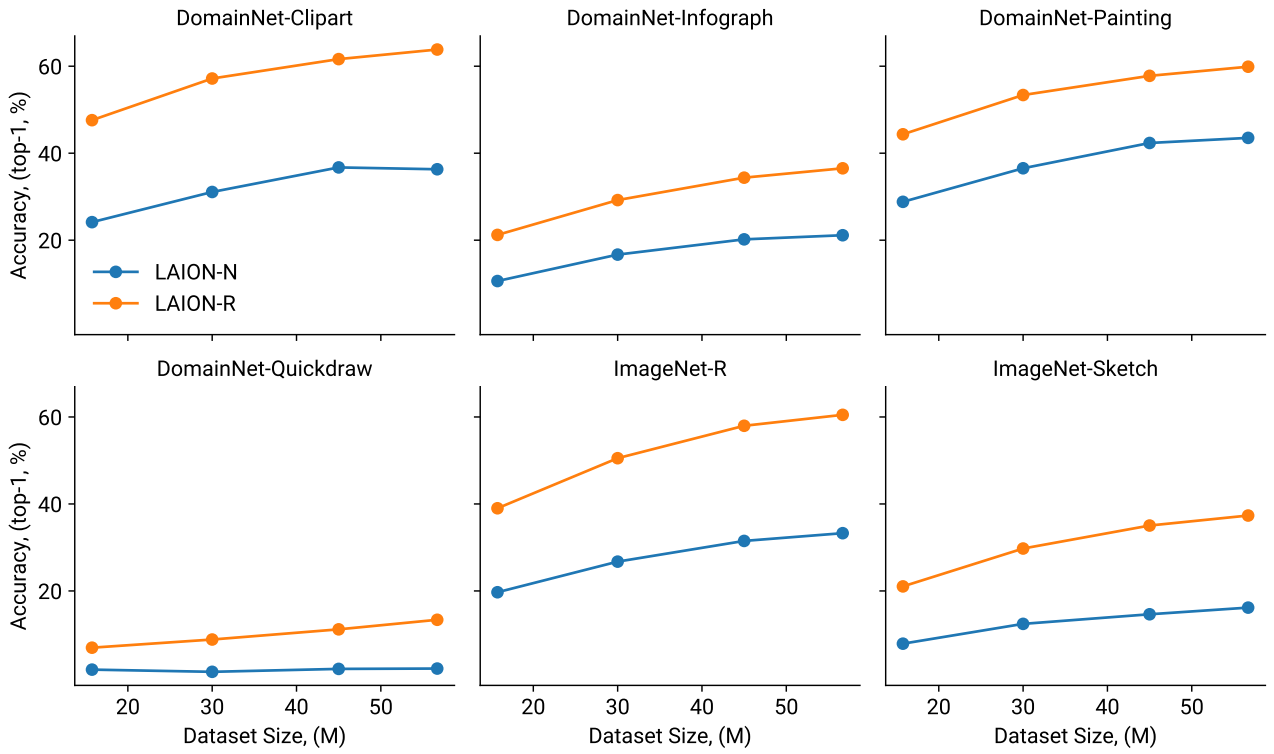

Figure 9: **CLIP trained on LAION v LAION-N performance on standard rendition test sets.**

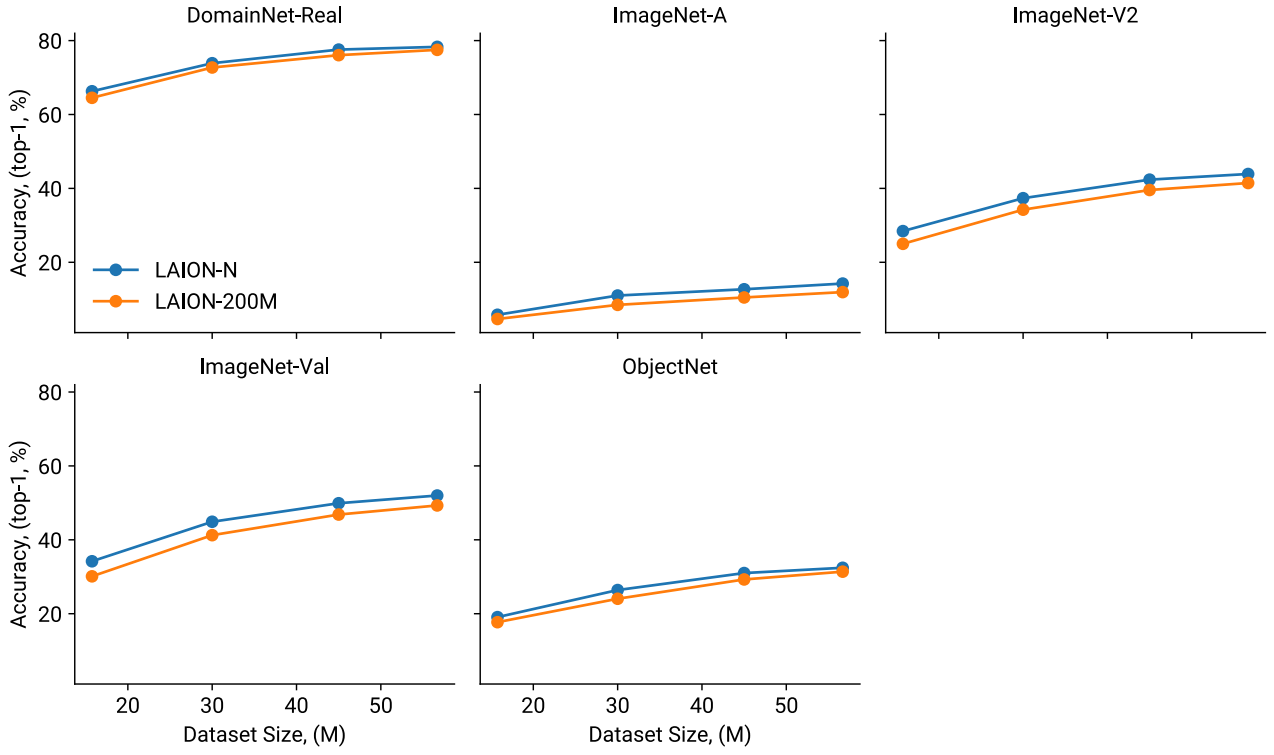

Figure 10: **CLIP trained on LAION v LAION-N performance on clean natural test sets.**

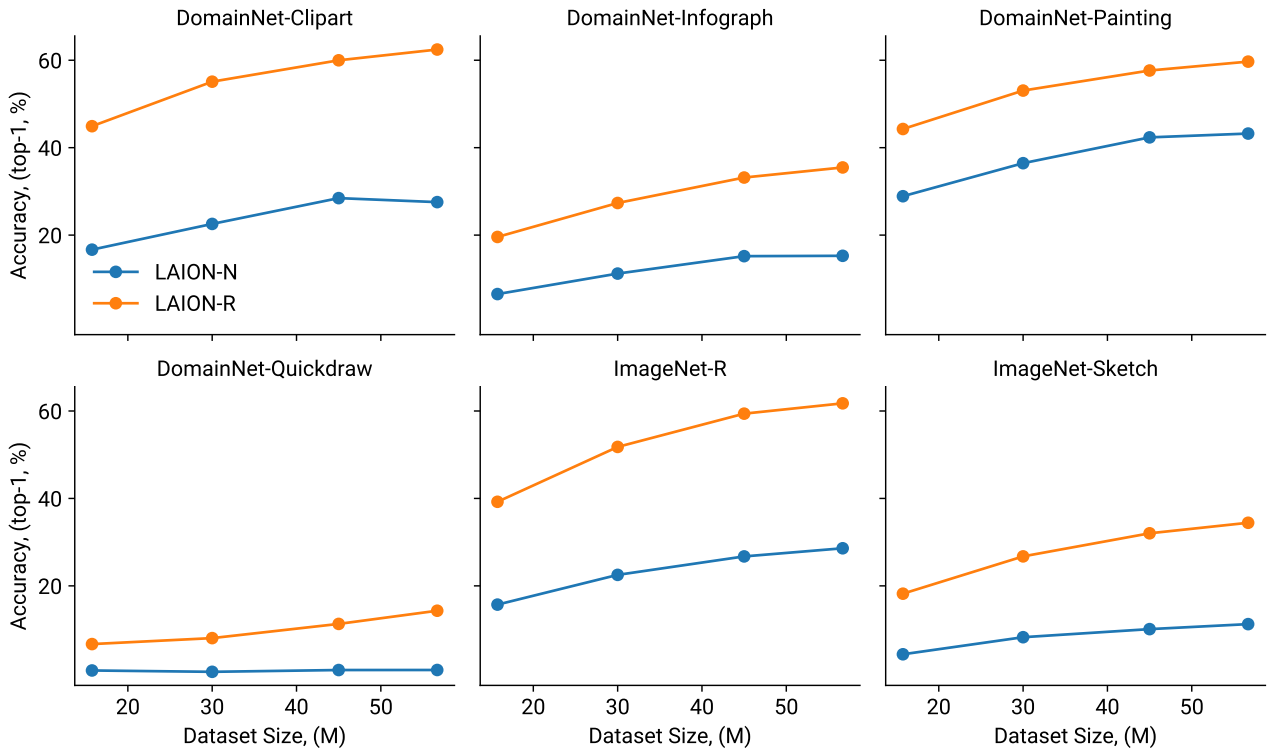

Figure 11: **CLIP trained on LAION v LAION-N performance on clean rendition test sets.**

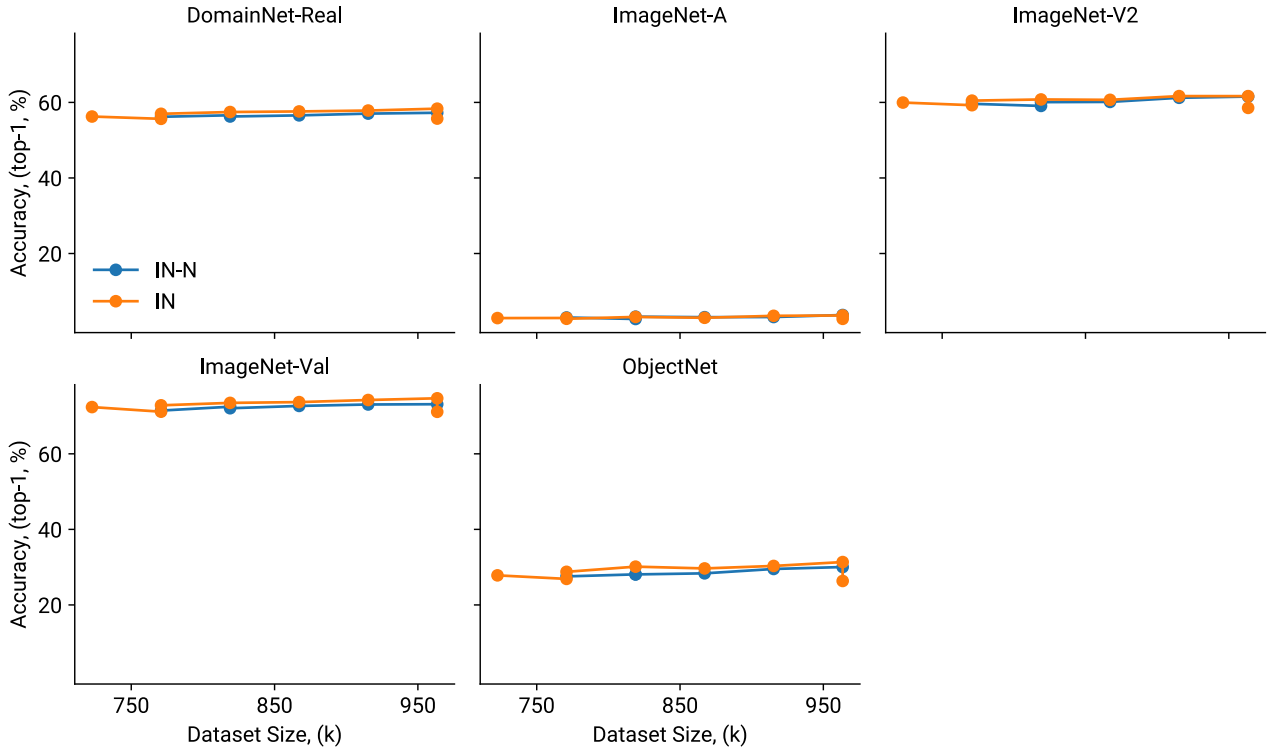

Figure 12: **Resnets trained on ImageNet v ImageNet-N performance on standard natural test sets.**

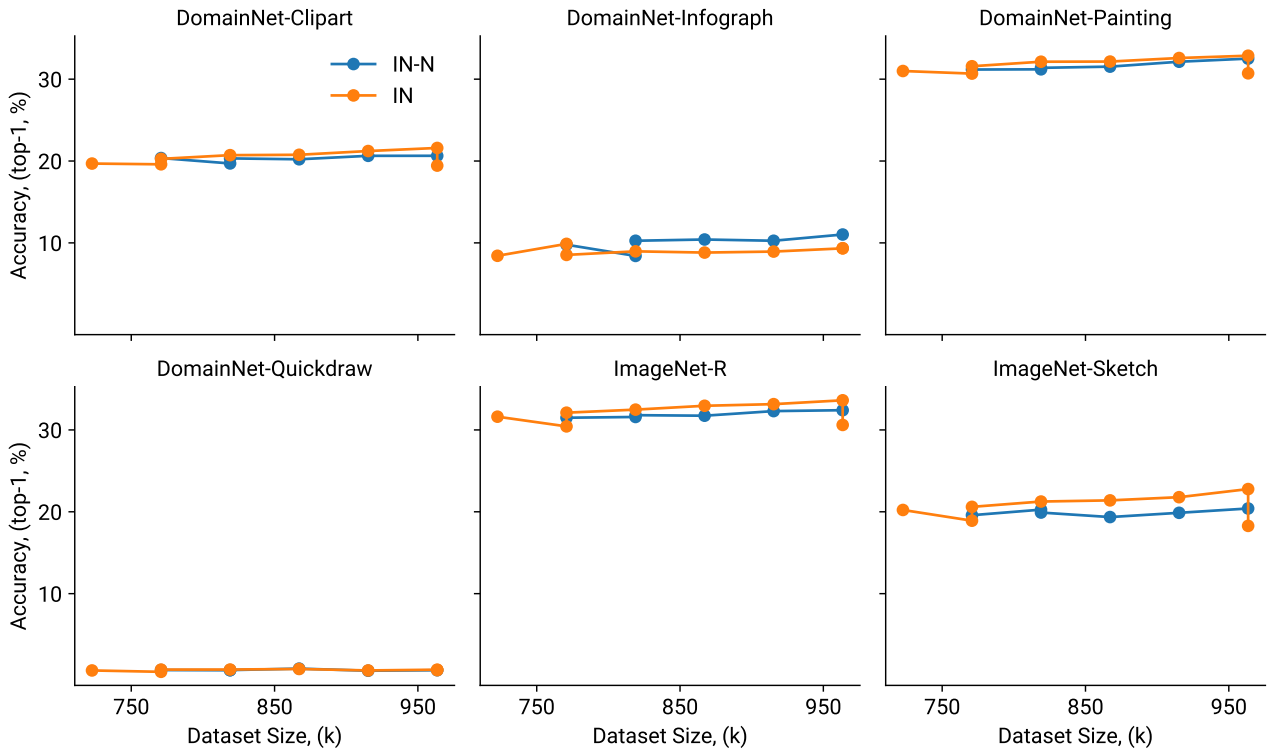

Figure 13: **Resnets trained on ImageNet v ImageNet-N performance on standard rendition test sets.**

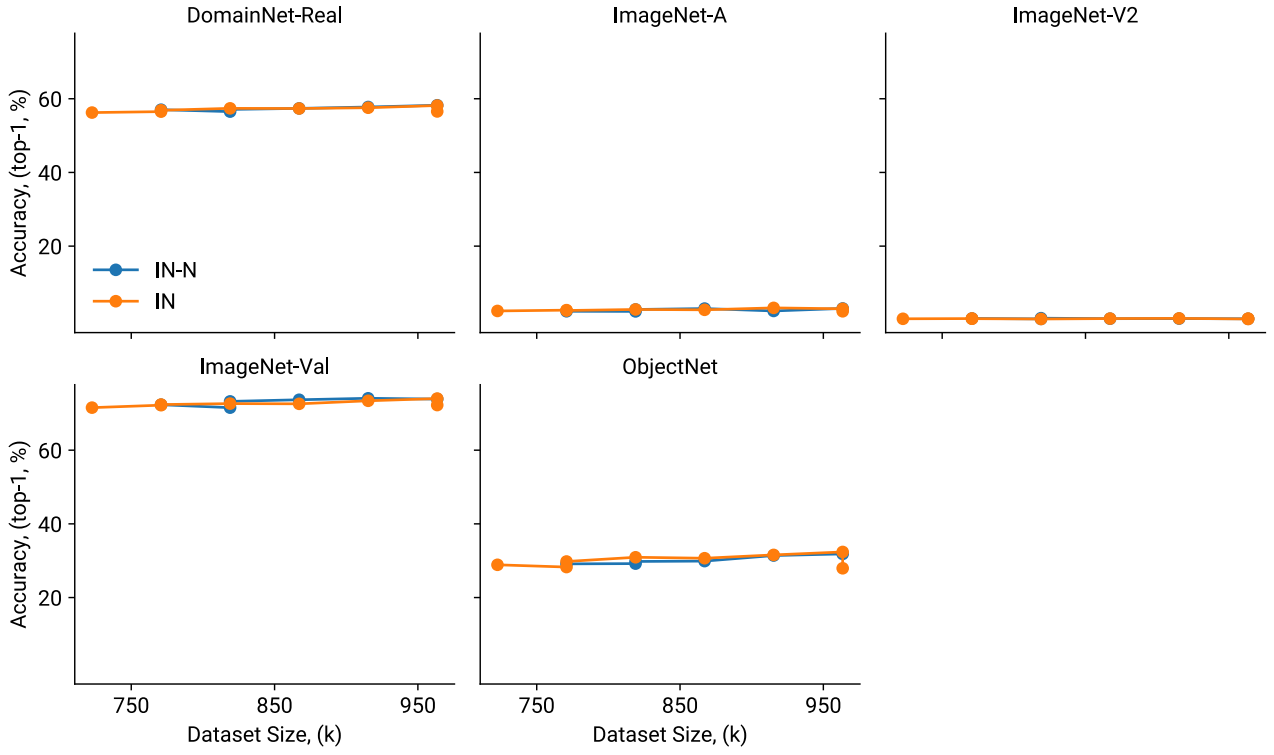

Figure 14: **Resnets trained on ImageNet v ImageNet-N performance on clean natural test sets.**

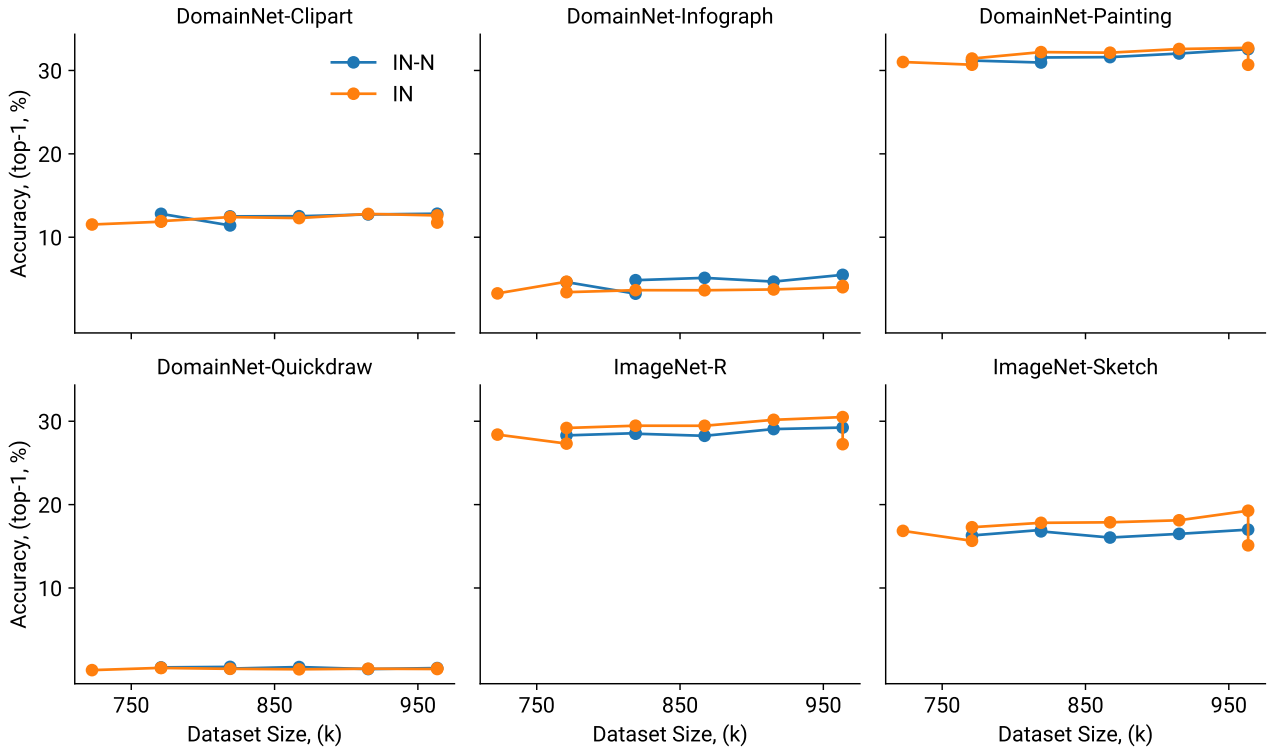

Figure 15: **Resnets trained on ImageNet v ImageNet-N performance on clean rendition test sets.**

G. Detailed Effective Robustness plots on individual shifts

In Fig. 3 in the main manuscript, we report aggregated results where we average over natural and stylistic ImageNet distribution shifts. We display the results on the individual distribution shifts in Fig. 16. On ImageNet-R and ImageNet-Sketch (bottom row), we observe that the effective robustness of the CLIP models can be modulated by training it on the different dataset splits, i.e. LAION-Natural, LAION-Rendition, LAION-Mix. The model trained on LAION-Natural is much closer to the ImageNet trained model in terms of effective robustness compared to the model trained on LAION-Rendition. In contrast, effective robustness is barely affected on the natural splits (top row). This can be explained by the final data distributions of the different training splits: Our filtering procedure does not affect natural images which are most responsible for the performance on natural datasets which explains the consistency in performance.

We also investigate effective robustness on the DomainNet shifts in Fig. 17. We note that the ImageNet model's accuracy numbers on DomainNet are not comparable to the CLIP models because the ImageNet model has been evaluated on a subset of DomainNet (ImageNet-D, Rusak et al., 2022) which is compatible with ImageNet classes. DomainNet has many classes which are not present in ImageNet, such as for example "The Great Wall of China" or "paper clip" which have been removed in ImageNet-D to enable evaluating ImageNet trained models without the need for training an additional readout layer. In contrast, we evaluate the CLIP trained models on the full DomainNet splits following standard zero-shot evaluation procedure. We will add a Figure where we control for the missing classes and evaluate the CLIP models on ImageNet-D in the next version of the manuscript.

On DomainNet, we similarly observe strong changes in effective robustness of the CLIP trained models when evaluating on the stylistic domains (all domains except for DomainNet-Real), and barely any changes when evaluating on the DomainNet-Real domain.

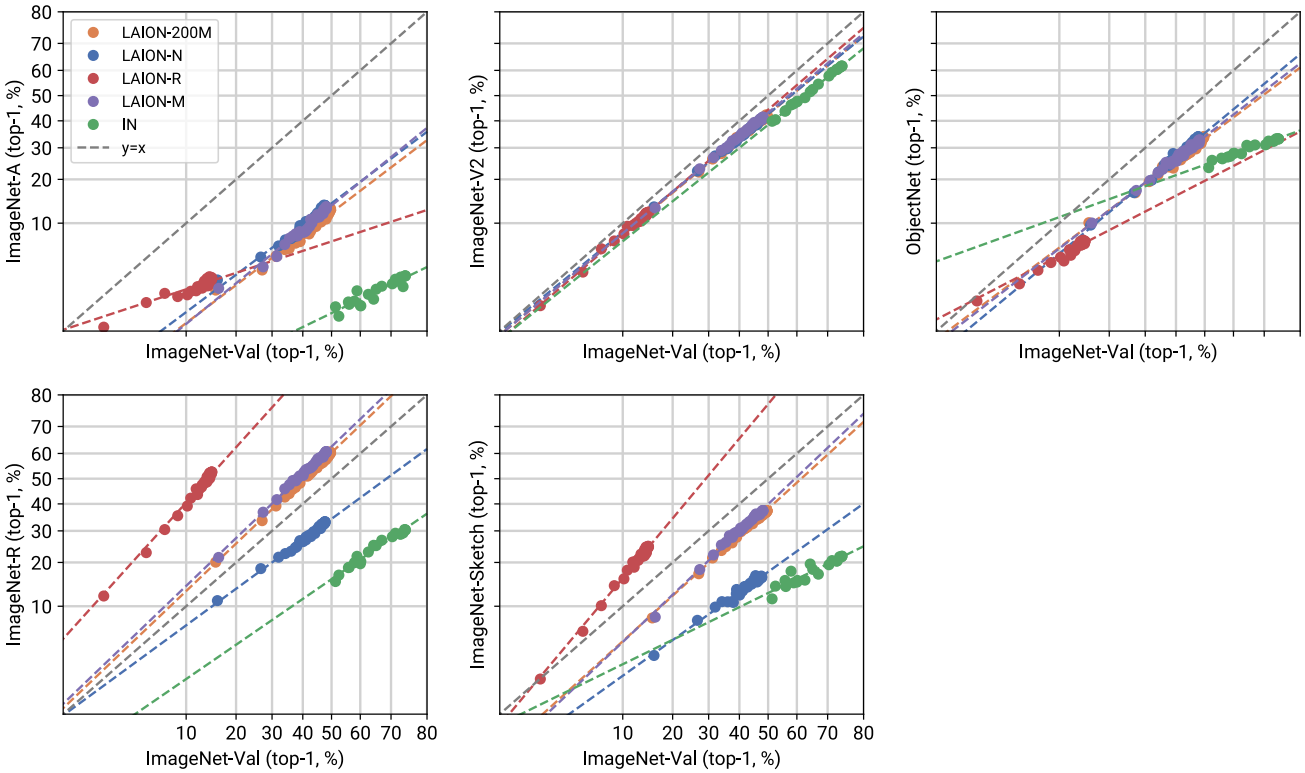

Figure 16: **Effective Robustness of different models on different ImageNet distribution shifts.** On ImageNet-R and ImageNet-Sketch (bottom row), we observe that the effective robustness of the CLIP models can be modulated by training it on the different dataset splits, i.e. LAION-Natural, LAION-Rendition, LAION-Mix. The model trained on LAION-Natural is much closer to the ImageNet trained model in terms of effective robustness compared to the LAION-Rendition model.

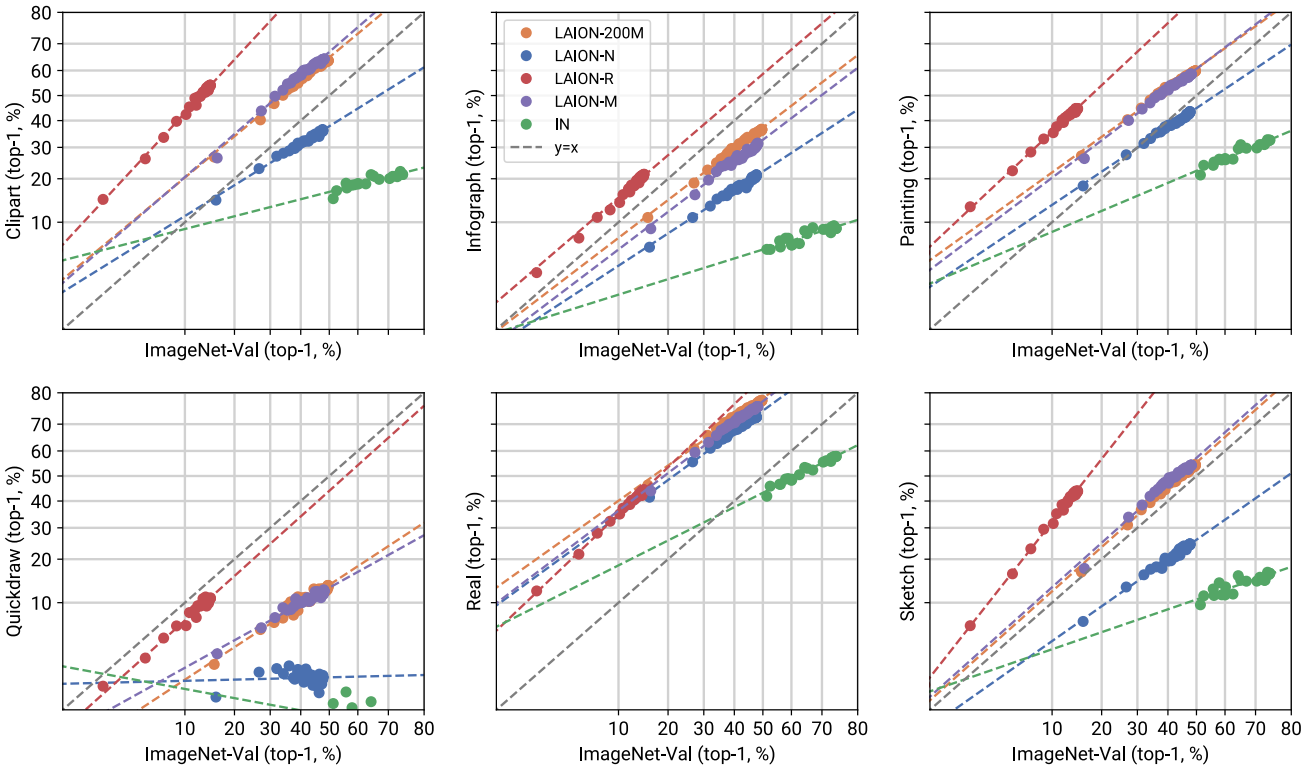

Figure 17: **Effective Robustness of different models on different DomainNet distribution shifts.** On the stylistic domains, we observe that the effective robustness of the CLIP models can be modulated by training it on the different dataset splits, i.e. LAION-Natural, LAION-Rendition, LAION-Mix. Effective robustness barely changes when evaluating different CLIP models on DomainNet-Real.

H. Visualization of Errors made by the domain classifier

We show images which have been misclassified by our domain classifier Fig. 18. We observe that the errors are interpretable. For example, the "natural" images which have been classified as "ambiguous" are indeed ambiguous: We see a sculpture in one image, a large woodwork of an ant in another and a pencil drawing of an airplane with a partly visible human hand drawing it in a third image.

I. Visualization of samples from the LAION dataset

We visualize random examples from the "Natural", "Rendition" and "Ambiguous" domains from LAION in Figs. 19-21.

J. Visualizations of ImageNet Distribution Shifts

We visualize random examples from the "Natural", "Rendition" and "Ambiguous" domains from the considered ImageNet shifts datasets in Figs. 22-27. We show 20 images per split; occasionally, there are fewer than 20 images in some of these splits, such as e.g. there are very few renditions in ImageNet-A. In that case, we plot all images from that split and leave the remaining subplots blank.

K. Visualizations of DomainNet Distribution Shifts

We visualize random examples from the "Natural", "Rendition" and "Ambiguous" domains from different DomainNet datasets in Figs. 28-33. We show 20 images per split; occasionally, there are fewer than 20 images in some of these splits, such as e.g. no natural images in the Quickdraw domain. In that case, we plot all images from that split and leave the remaining subplots blank.

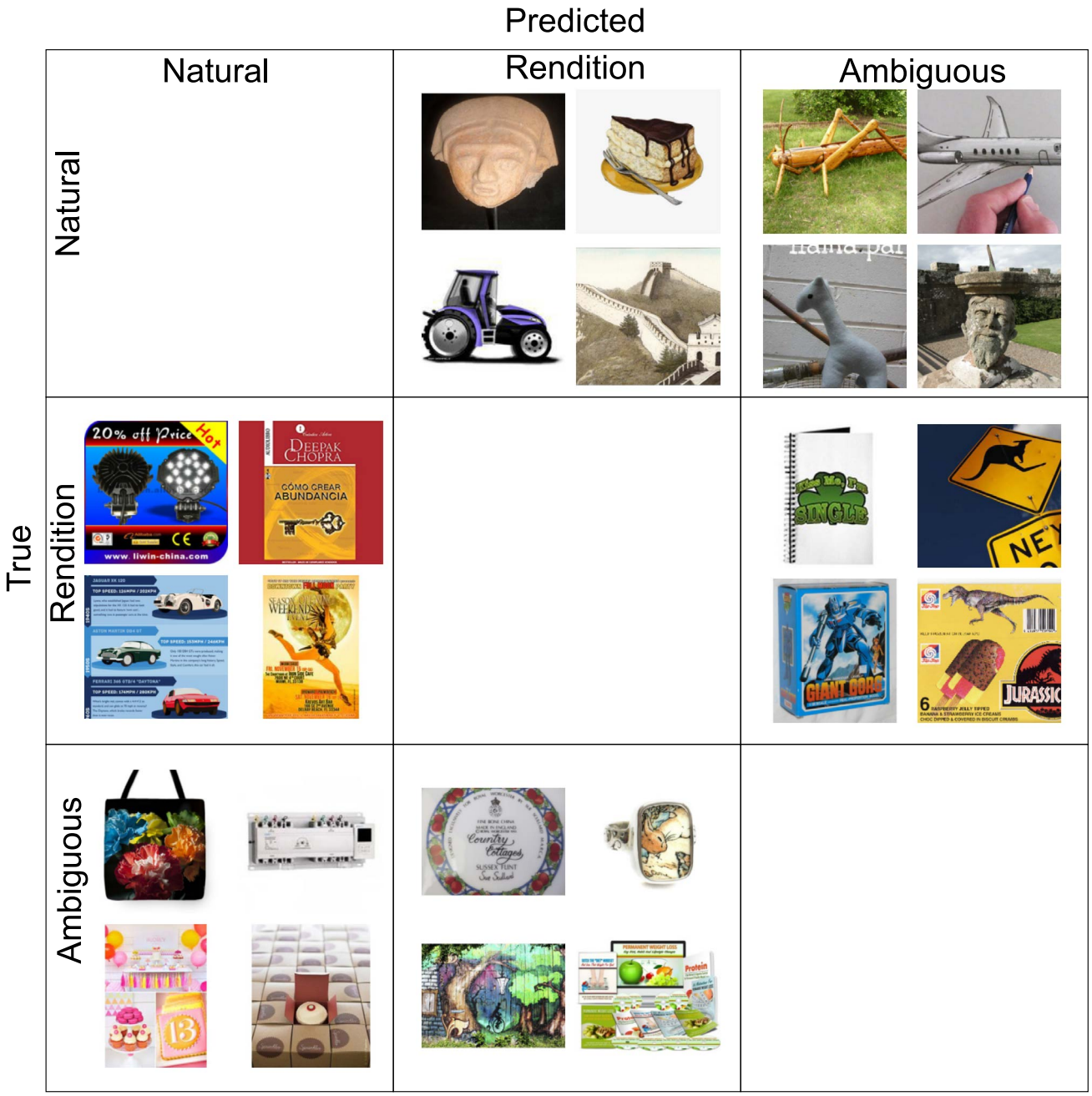

Figure 18: **Confusion matrix of example images which have been misclassified by our domain classifier.**

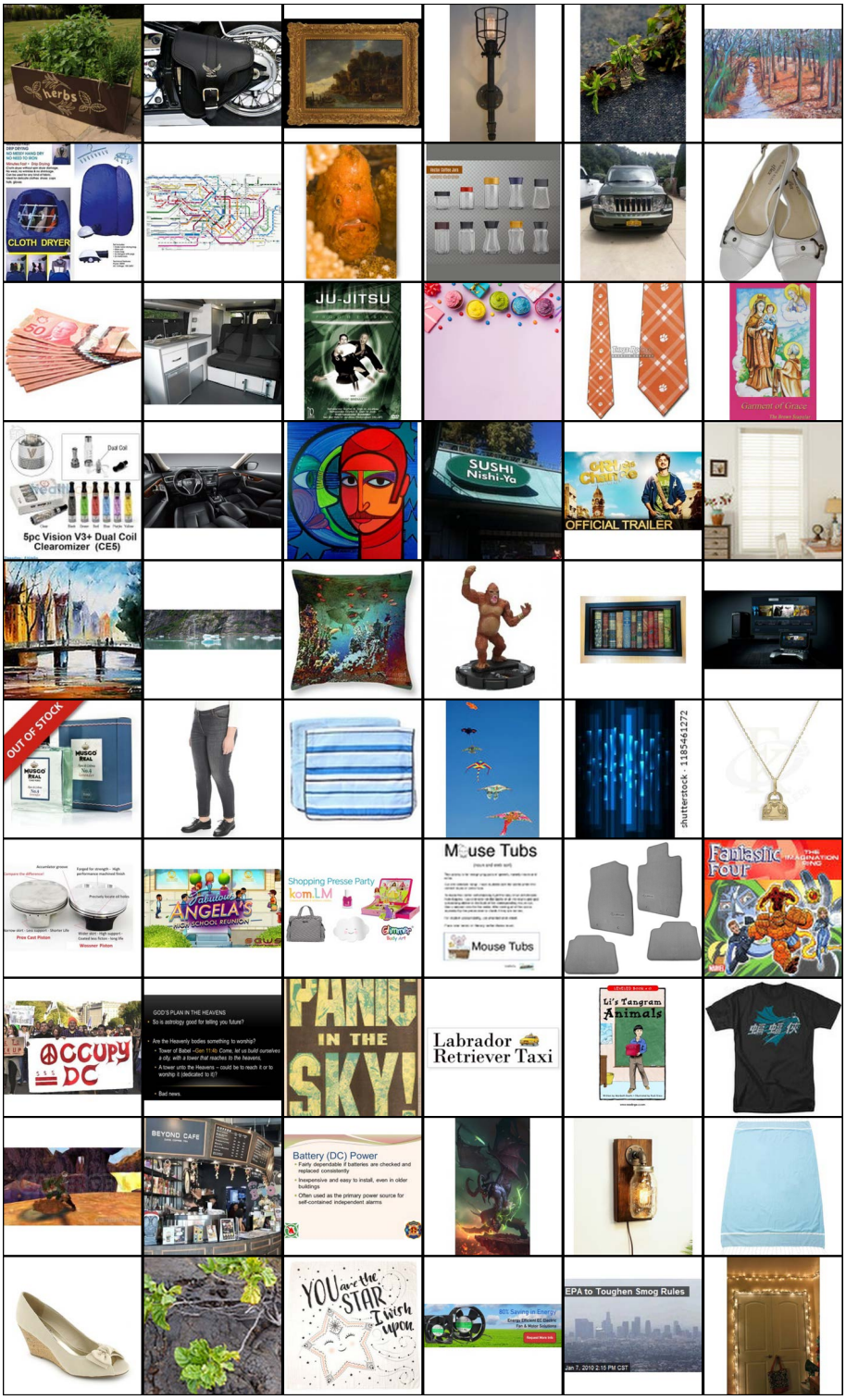

Figure 19: **Random samples from LAION-200M**. We omit NSFW images and images of humans.

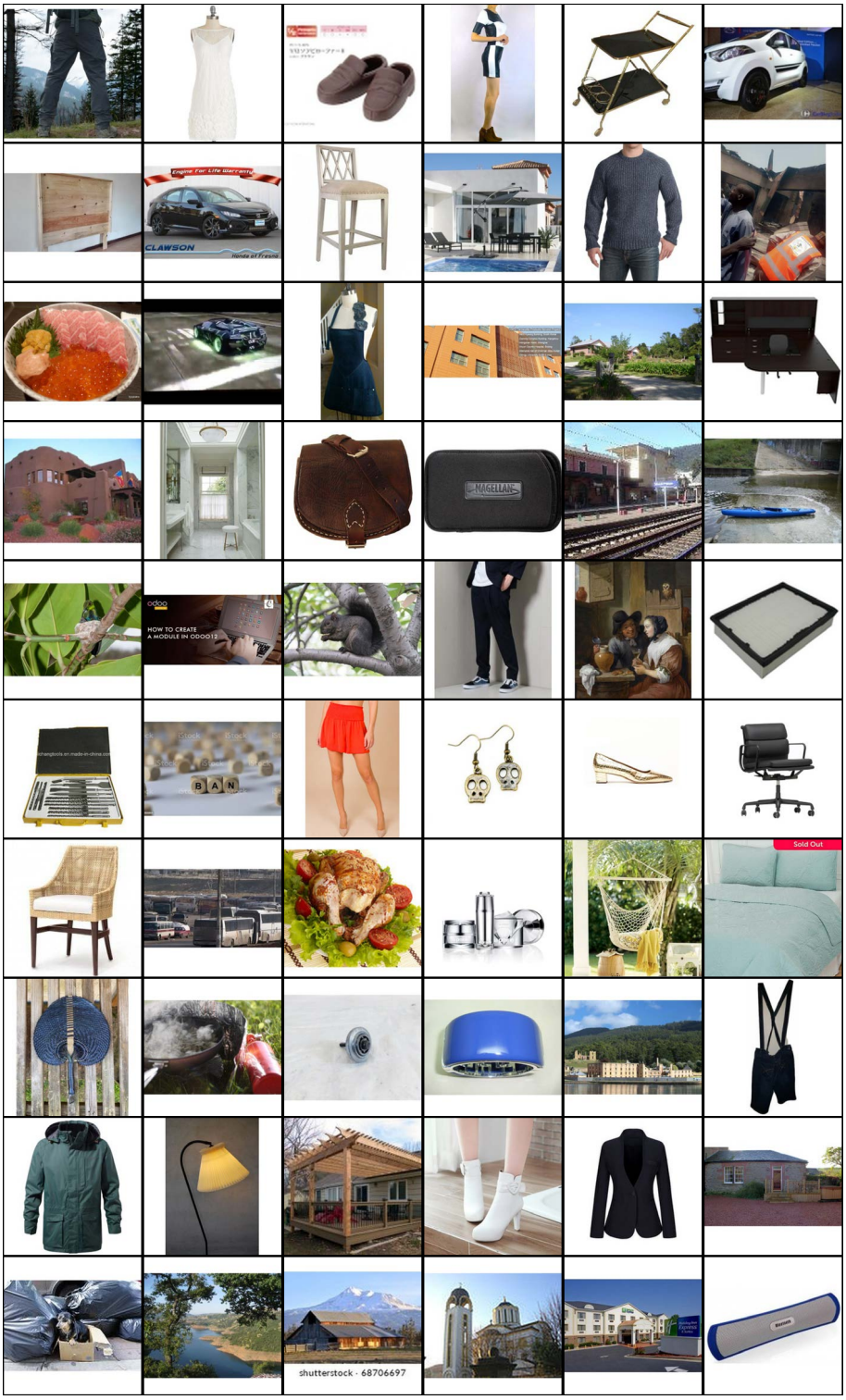

Figure 20: **Random samples from LAION-Natural**. We omit NSFW images and images of humans.

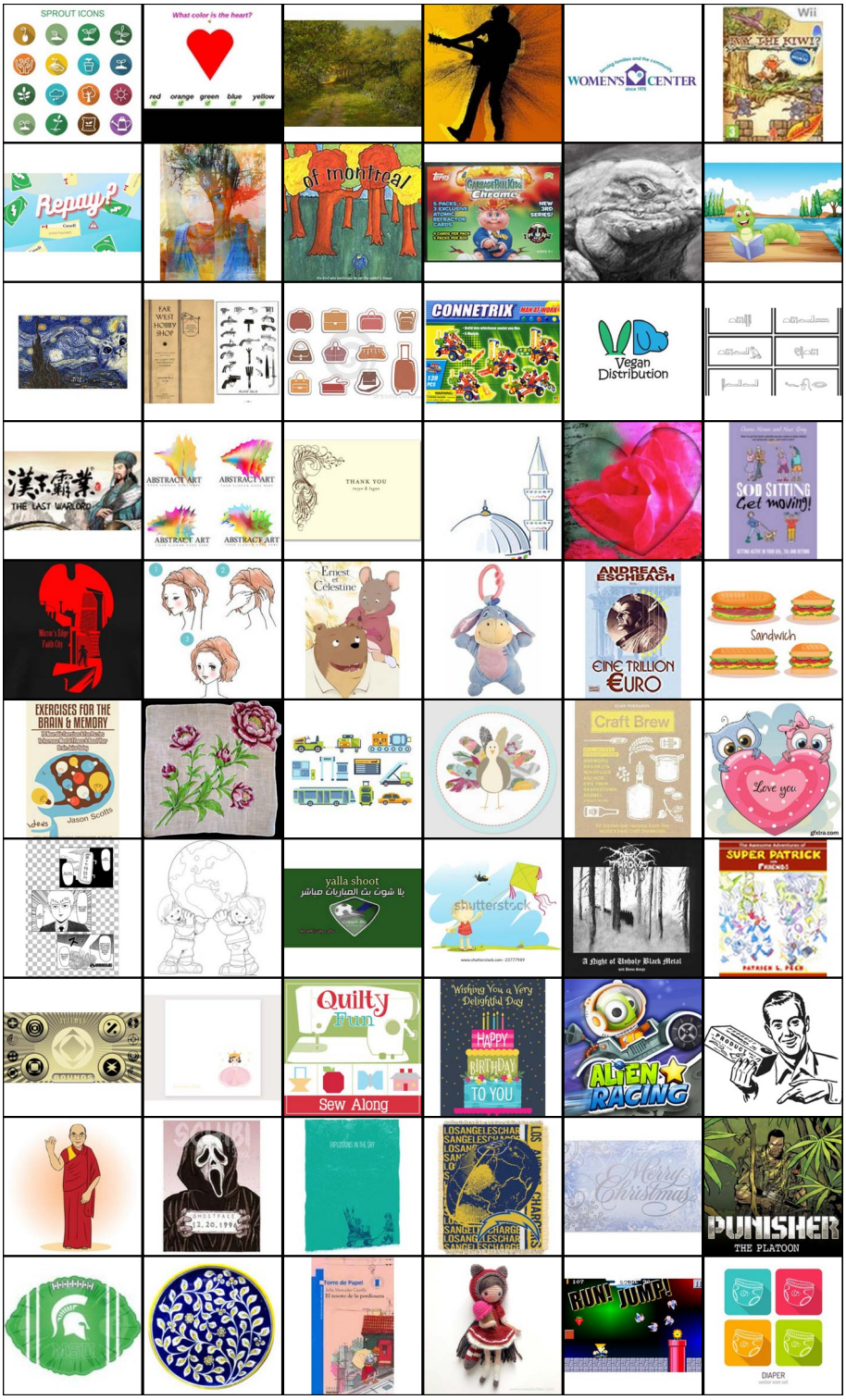

Figure 21: **Random samples from LAION-Rendition**. We omit NSFW images and images of humans.

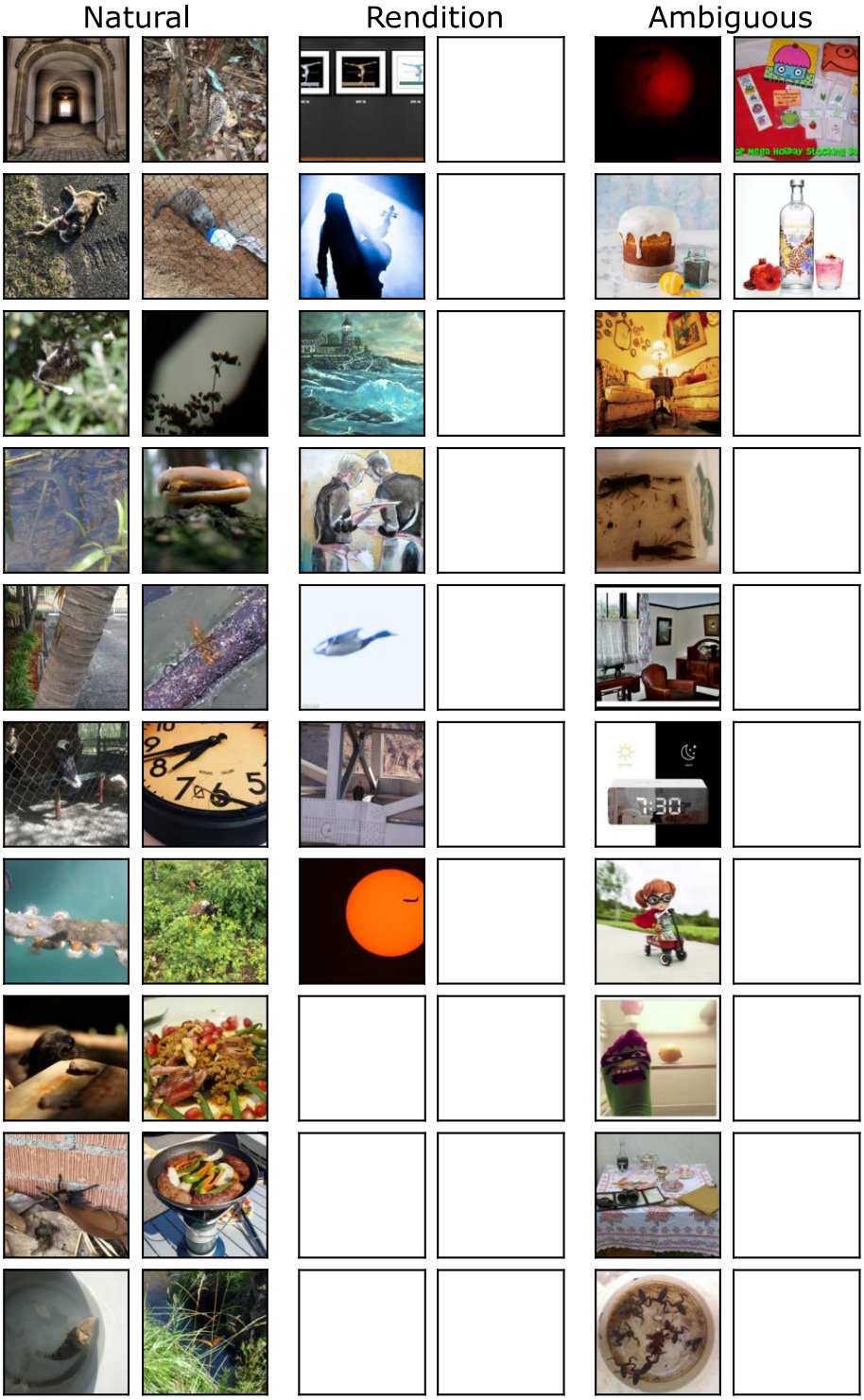

Figure 22: **Random samples of ImageNet-A grouped by domain.** We omit NSFW images and images of humans.

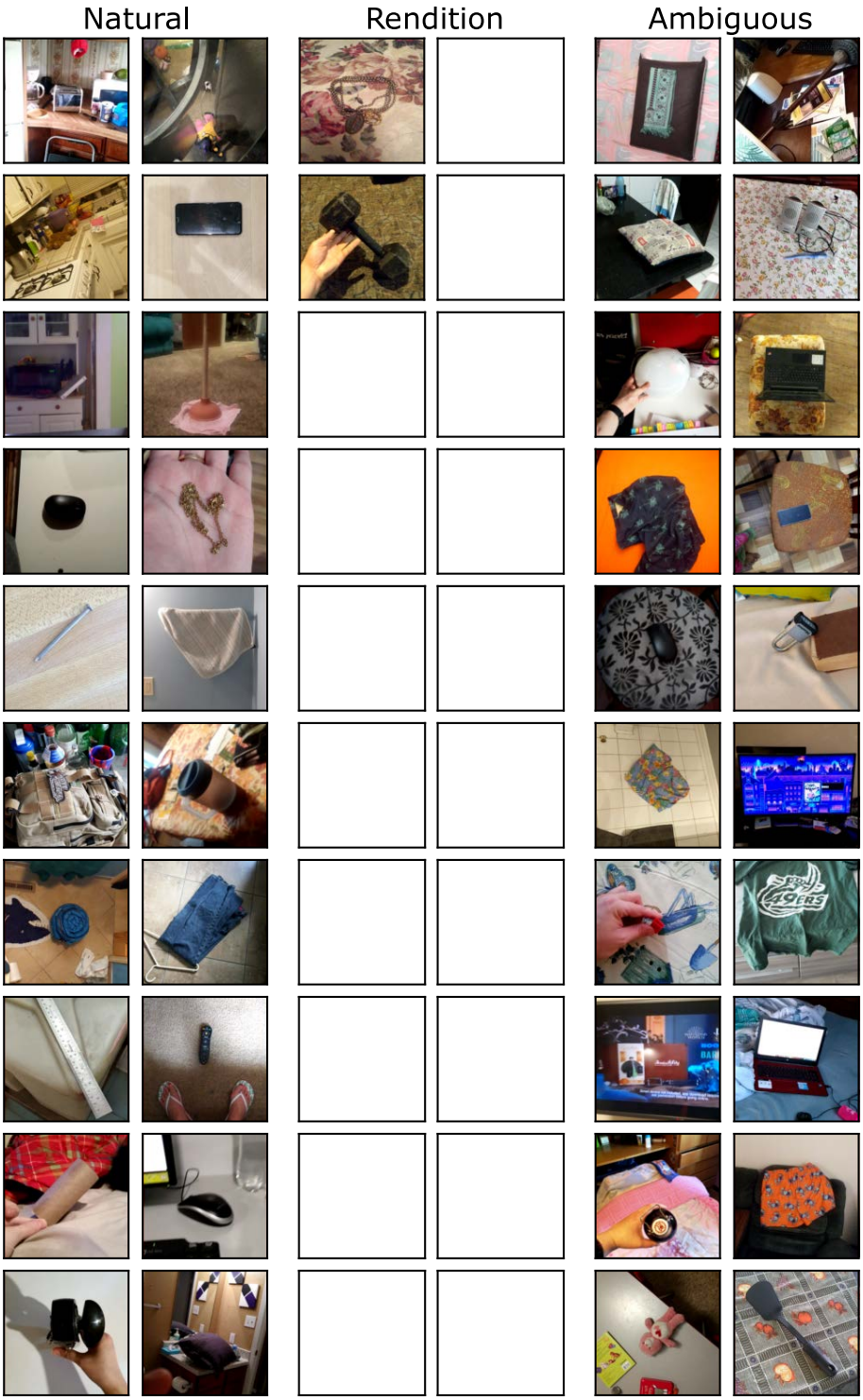

Figure 23: Random samples of ObjectNet grouped by domain. We omit NSFW images and images of humans.

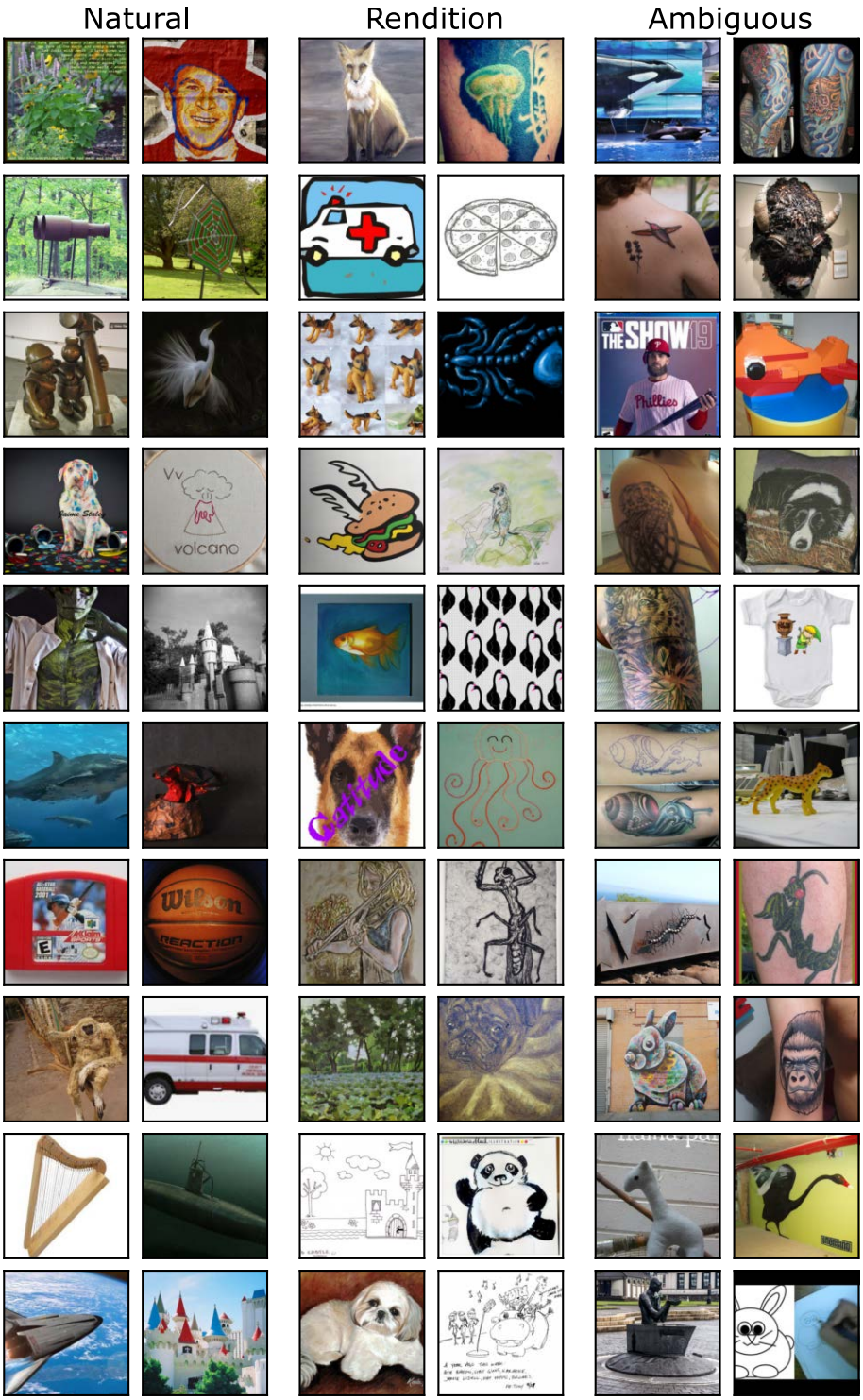

Figure 24: **Random samples of ImageNet-R grouped by domain.** We omit NSFW images and images of humans.

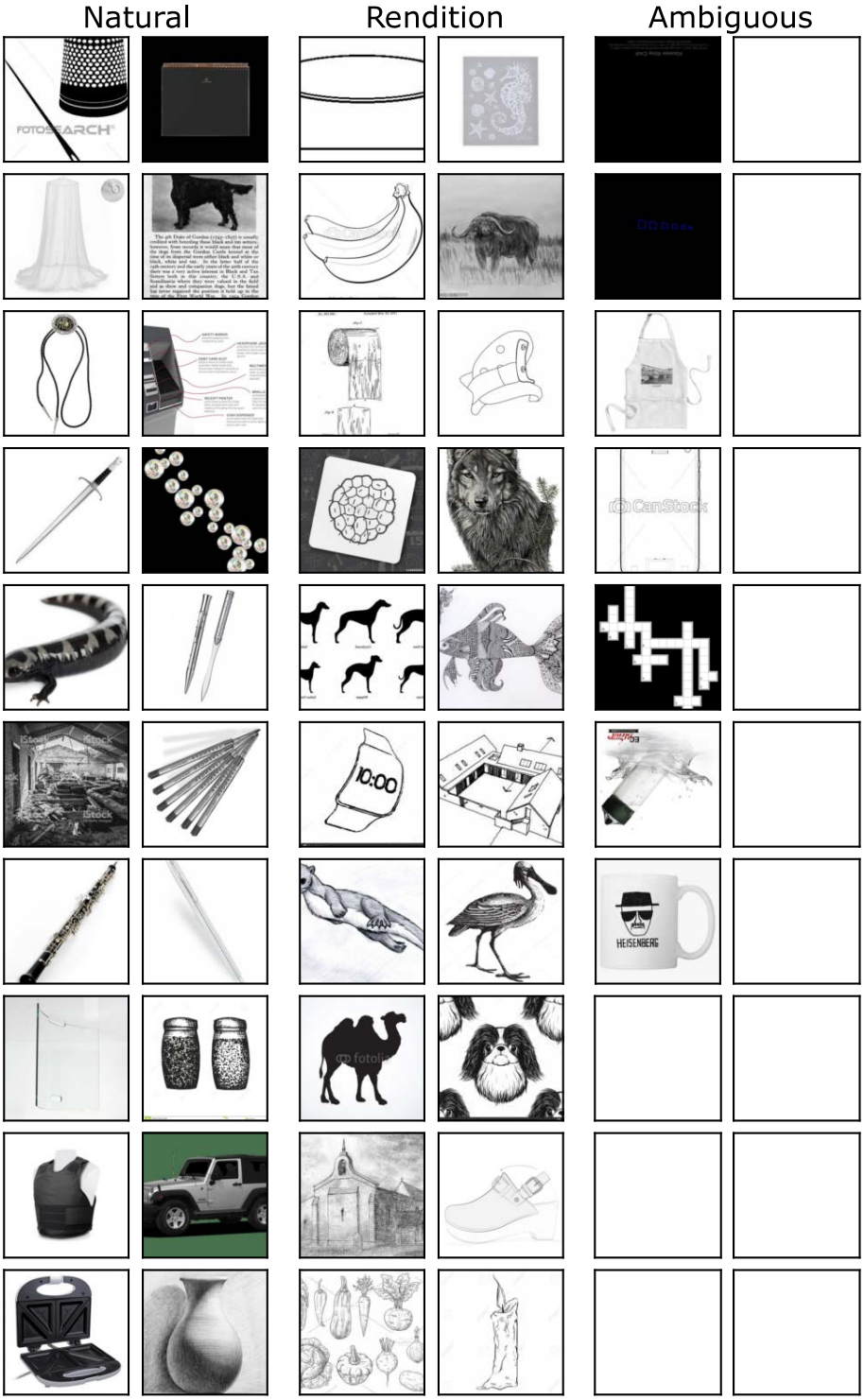

Figure 25: **Random samples of ImageNet-Sketch grouped by domain.** We omit NSFW images and images of humans.

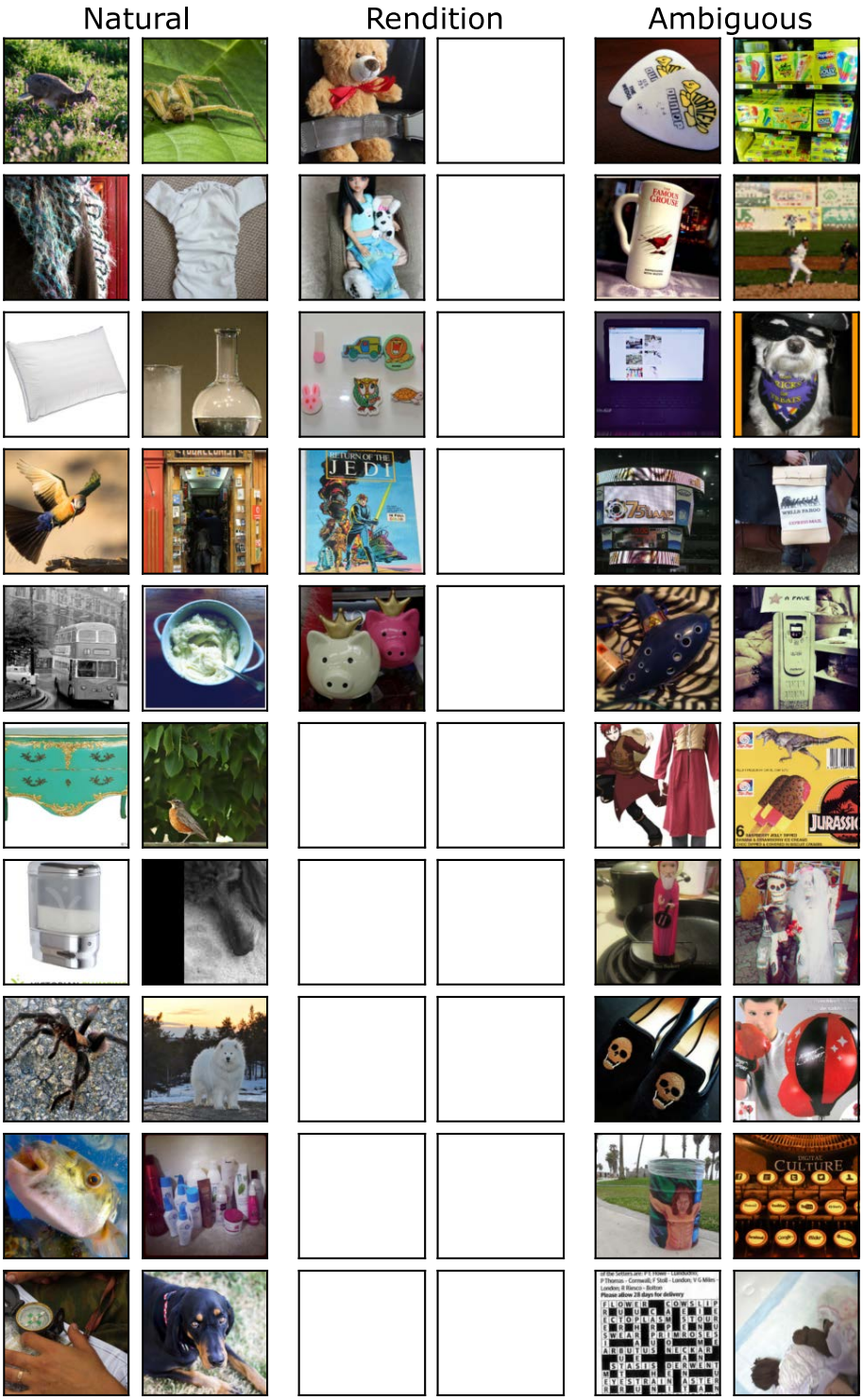

Figure 26: **Random samples of ImageNet-V2 grouped by domain.** We omit NSFW images and images of humans.

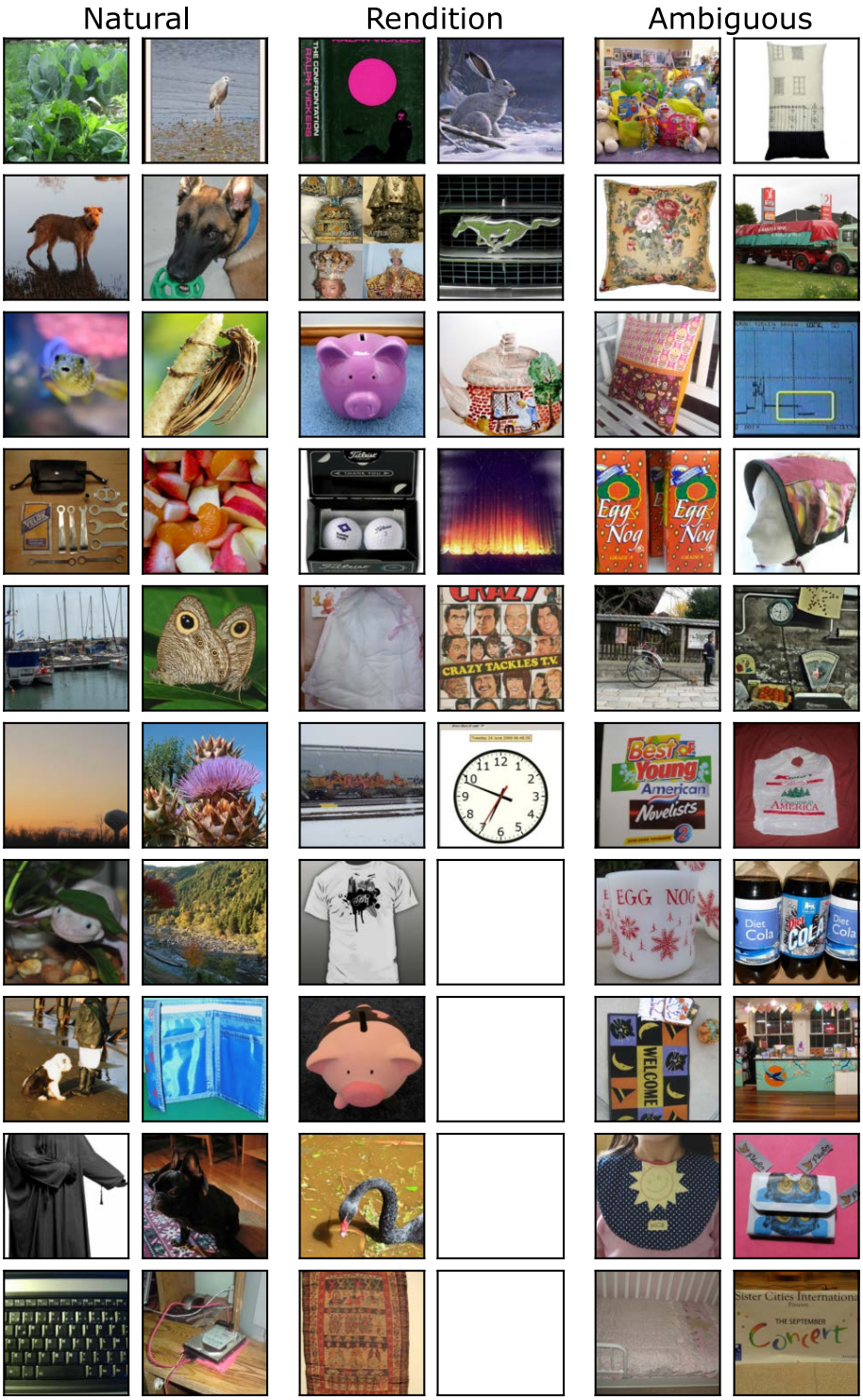

Figure 27: **Random samples of ImageNet-Val grouped by domain.** We omit NSFW images and images of humans.

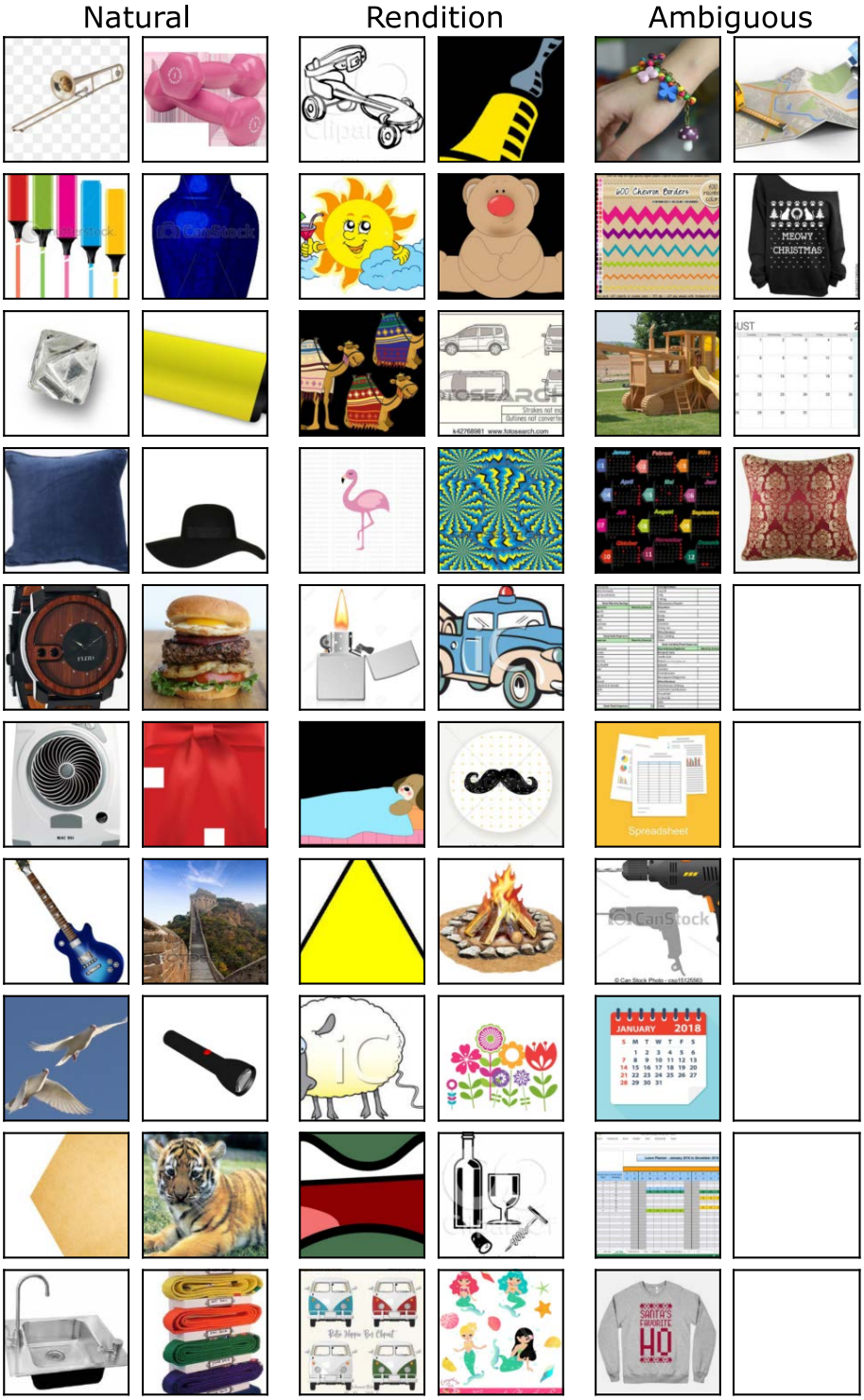

Figure 28: **Random samples of DomainNet-Clipart grouped by domain.** We omit NSFW images and images of humans.

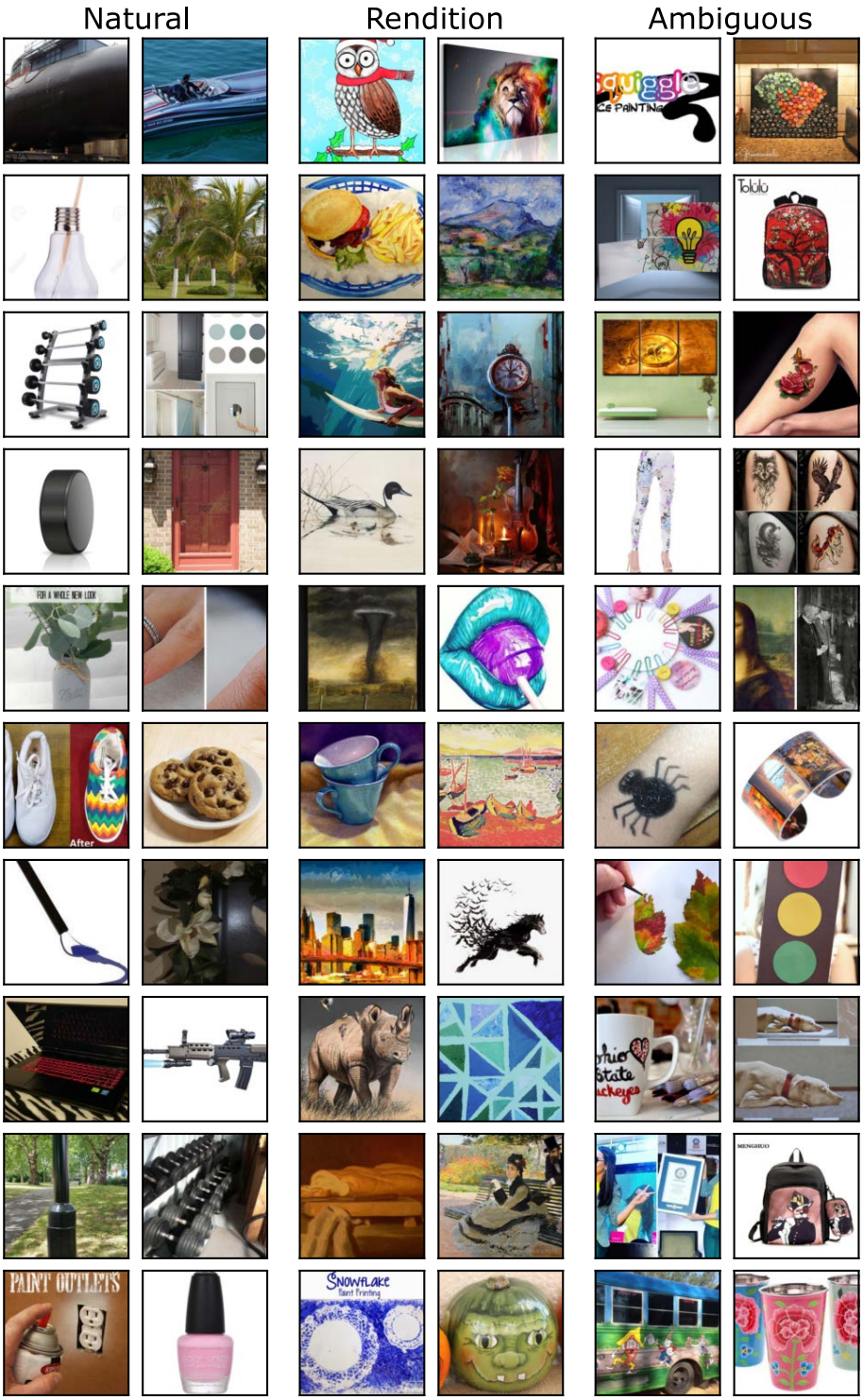

Figure 29: **Random samples of DomainNet-Painting grouped by domain.** We omit NSFW images and images of humans.

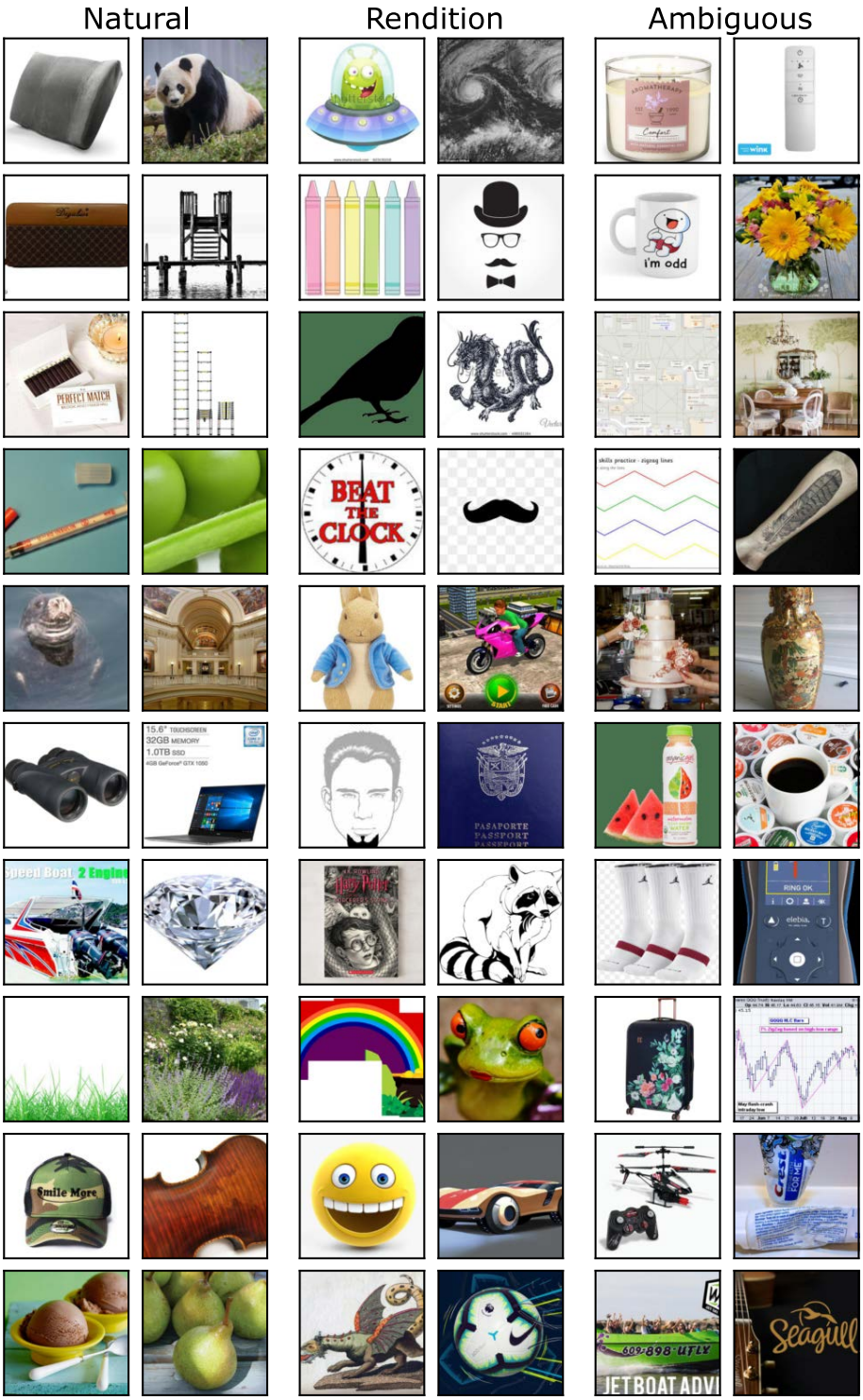

Figure 30: **Random samples of DomainNet-Real grouped by domain.** We omit NSFW images and images of humans.

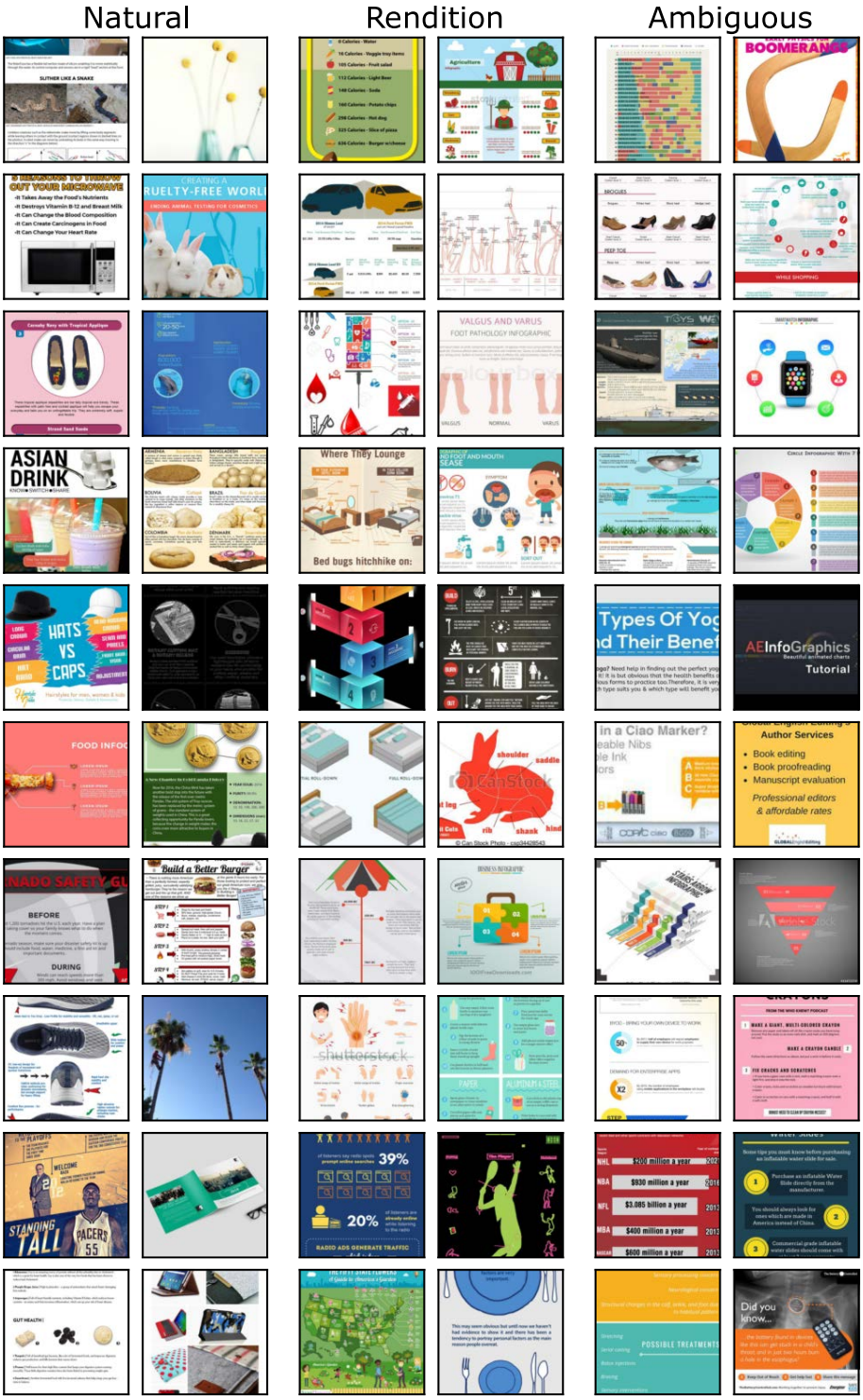

Figure 31: **Random samples of DomainNet-Infograph grouped by domain.** We omit NSFW images and images of humans.

Figure 32: **Random samples of DomainNet-Quickdraw grouped by domain.** We omit NSFW images and images of humans.

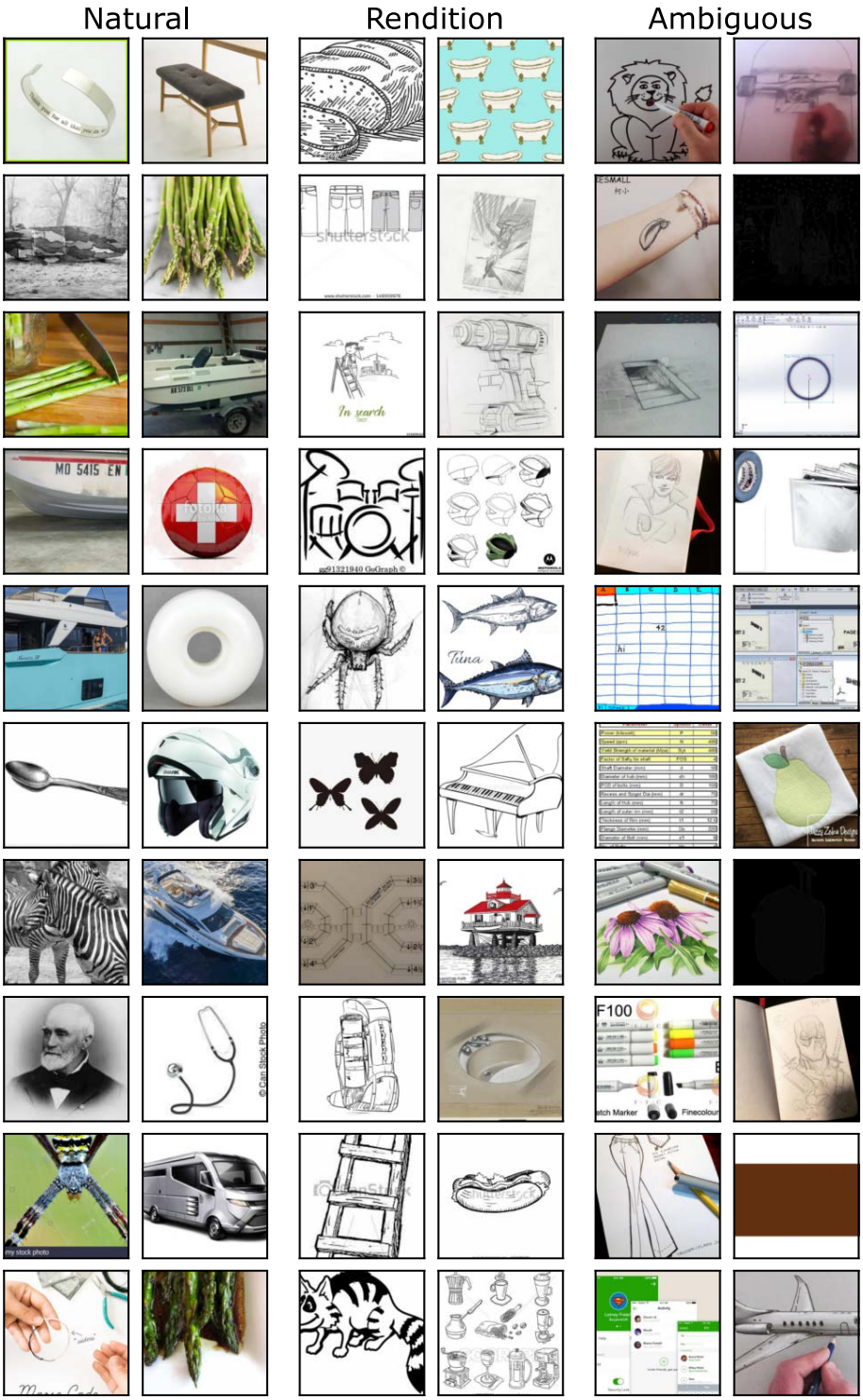

Figure 33: **Random samples of DomainNet-Sketch grouped by domain.** We omit NSFW images and images of humans.