In Search of Forgotten Domain Generalization

Prasanna Mayilvahanan[1,2,3*] **Roland S. Zimmermann**[1,2,3*] **Thaddäus Wiedemer**[1,2,3]

Evgenia Rusak[1,2,3] **Attila Juhos**[1,2,3] **Matthias Bethge**[1,2] **Wieland Brendel**[2,3,4]

[1]University of Tübingen [2]Tübingen AI Center
[3]Max-Planck-Institute for Intelligent Systems, Tübingen [4]ELLIS Institute Tübingen

prasanna.mayilvahanan@uni-tuebingen.de, research@rzimmermann.com

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 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 domain 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. Training CLIP models on these datasets reveals that a significant portion of their performance is explained by in-domain examples. This indicates that the OOD generalization challenges from the ImageNet era still prevail and that training on web-scale data merely creates the illusion of OOD generalization. Furthermore, through a systematic exploration of combining natural and rendition datasets in varying proportions, we identify optimal mixing ratios for model generalization across these domains. Our datasets and results re-enable meaningful assessment of OOD robustness at scale—a crucial prerequisite for improving model robustness.

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 [27, 37, 30]. One notable example is CLIP [29], a vision-language model pretrained on a vast dataset of image-text pairs, which forms the backbone of numerous other foundation models [30, 20]. CLIP has achieved unprecedented performance across a wide range of benchmarks spanning many domains—a sharp contrast to models from the ImageNet era, which struggled to generalize from a training domain mostly consisting of natural photographs to stylistically different domains such as ImageNet-Sketch [41], ImageNet-R [15], and DomainNet [28].

Domains, while often challenging to quantify in practice [5], emerge from collecting data from specific sources and conditions. Some domains, like *natural images* or *renditions*, are better delineated, allowing the creation of datasets like the ones mentioned above. Out-of-domain (OOD) generalization refers to a model's ability to perform well on data from domains other than its training domain(s) [42]. In this work, we collectively refer to the domain represented by ImageNet-Sketch, ImageNet-R, DomainNet-Painting, DomainNet-Clipart, DomainNet-Sketch, and DomainNet-Quickdraw as the *rendition domain*, since it contains images that are renditions of natural objects and scenes. Generalization to the rendition domain (especially OOD) is crucial for aligning models with human

38th NeurIPS 2024, 2nd Workshop on Attributing Model Behavior at Scale (ATTRIB 2024).

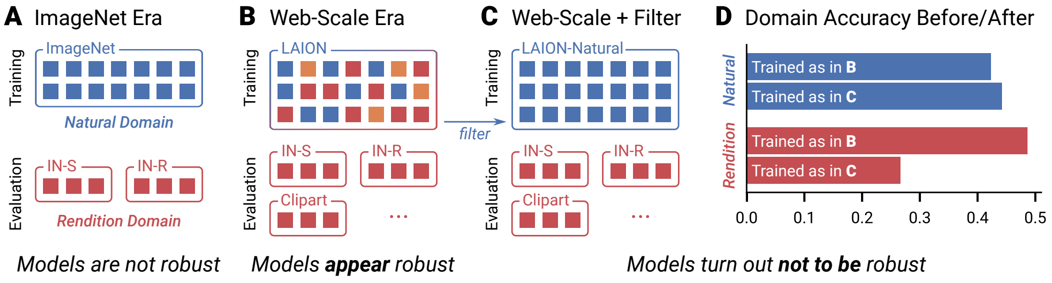

Figure 1: **Evaluated correctly, CLIP's OOD performance on renditions drops significantly**. **A**: Models used to be trained on a single domain like ***natural images*** from ImageNet [33] and evaluated for out-of-domain (OOD) generalization on a different domain like ***renditions*** from test sets such as ImageNet-R [15], ImageNet-Sketch [41]. **B**: Today, large foundation models like CLIP [29] are trained on web-scale datasets such as LAION-400M [36] 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 [1] to obtain LAION-Natural, a web-scale dataset 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, suggesting that its OOD performance has been previously overestimated. The models are evaluated on refined test datasets containing samples only from their intended domains.

perception, as humans can interpret abstract visual renditions, while machines tend to rely heavily on textural cues [15, 13].

CLIP's strong performance in several domains, including renditions, is attributed to its vast training distribution, rather than its contrastive learning objective, language supervision, or dataset size [10]. However, Fang et al. [10] do not specify what characteristics of the training distribution drive this performance. CLIP could be learning more robust representations due to the diversity of natural images in its training set—or it may simply have been exposed to many datapoints from the (assumed to be OOD) test domains during training. Indeed, Mayilvahanan et al. [22] revealed that CLIP's training data contains exact or near duplicates of samples of many OOD datasets. Yet, they showed that CLIP still generalizes well when this *sample contamination* is corrected. However, their analysis failed to account for *domain contamination*.

In contrast to sample contamination, domain contamination does not focus on duplicates of specific datapoints but rather examines whether critical aspects of a test domain are present in the training domain, such as images with different content but similar style to test samples. For example, after the correction by Mayilvahanan et al. [22], many other *rendition* images, while not duplicates, remained in CLIP's training set (refer to Tab. 8). Prior works often assume that CLIP is capable of generalizing OOD [29, 2, 25, 10, 19, 38]; however, it remains unclear whether this is truly the case or if its performance is primarily driven by training on images from the test domain. This leads us to our central question:

To what extent does domain contamination explain CLIP's performance on renditions?

We address the central question with the following contributions:

- **Constructing Clean Single-Domain Datasets**: To rigorously test whether CLIP's success in the rendition domain stems from their exposure during training, we first train a domain classifier to distinguish natural images from renditions (Sec. 3.1). By applying the domain classifier to a deduplicated version of LAION-400M, we create and release two datasets: LAION-Natural contains 57 M natural images; LAION-Rendition consists of 16 M renditions of scenes and objects. Additionally, we refine existing rendition OOD benchmarks (ImageNet-R, ImageNet-Sketch, etc.) by removing samples that do not belong to the corresponding domain (Sec. 3.3).
- **Refining the Evaluation of CLIP's OOD Performance**: Using LAION-Natural, we demonstrate that CLIP trained only on natural images significantly underperforms on rendition domain shifts (Sec. 4). This suggests that its original success stems from domain contamination, not from an intrinsic OOD generalization ability (see Fig. 1 for a summary).

- **Investigating Domain Mixing and Scaling Effects**: Our single-domain datasets enable analyzing the effects of training on controlled mixtures of natural and rendition images across scales (Appx. D). We identify the optimal mixing ratio for the best overall performance and show the degree to which training on one domain enables some generalization to the other.

Through this work, we aim to shed light on the limitations of foundation models like CLIP in handling OOD generalization and provide valuable datasets and tools to the community for further exploration. Fig. 1 illustrates our core methodology.

2 Abridged Related Work

On gauging the OOD generalization performance of CLIP, Mayilvahanan et al. [22] 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. C for a thorough literature review.

3 Distinguishing Image Style Domains

Our work hinges on filtering out datapoints that belong to specific domains from web-scale datasets. As noted above, no precise definition exists 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.

We describe our labeling procedure based on this demarcation in App. F Sec. F.1 and explore different ways to train a domain classifier on the resulting dataset in Sec. 3.1. In Sec. 3.2, 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. 3.3.

3.1 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 containing 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 labeled LAION-200M images, retaining 3000 samples each for a validation and test set. From the domain classification literature discussed in Sec. C, we evaluate four methods with publicly available code that we outline below. All methods build on CLIP ViT-L/14 pre-trained on LAION-2B, which we choose for its balance between accuracy and inference speed.

Contrastive Style Descriptors (CSD) [39] 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 labeled 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 labeled 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), which jointly can be used for our ternary classification.

For further details on Density Ratios, Centroid Embeddings, Fine-Tuning, check out Appx. F.8.

3.2 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-200M's domain

Table 1: **We chose the best *natural* classfier and the best *rendition* classifier** amongst binary classifiers based on Contrastive Style Descriptors (CSD) [39] and Density Ratios (DR) [8] 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

Table 2: **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. [22] 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 %

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. 8, it is clear that LAION-200M contains a considerable portion of strictly rendition images (at least 7.90 % corresponding to 16 million images), and potentially many more images with some rendition 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). Additionally, we 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 (see Tab. 7, Appx. F.5). Both observations suggest that previous domain-generalization performance for models trained or evaluated on those datasets needs to be taken with a grain of salt: It is highly likely that their scores are inflated and the models' true generalization capability is lower.

We also analyze the domain composition of datasets from Mayilvahanan et al. [22] in Appx. F.9.

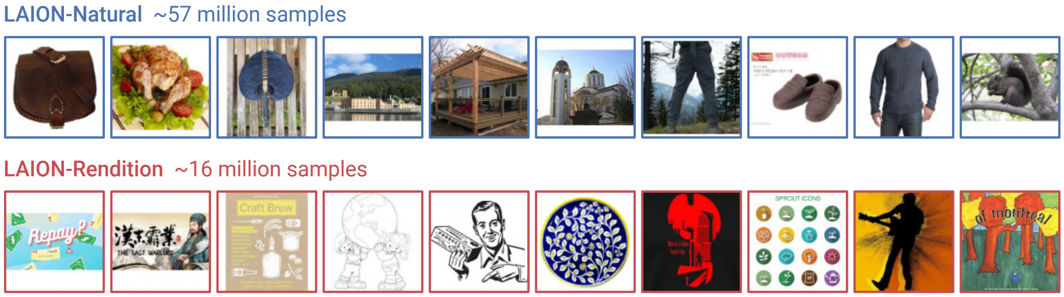
LAION-Natural ~57 million samples
LAION-Rendition ~16 million samples

Figure 2: **Random samples from LAION-Natural and LAION-Rendition**.

Table 3: **Performance on the *rendition* domain is driven by renditions in the training data**. We compare the top-1 accuracy of 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. 8). 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 %	17.81 %
LAION-Mix-13M	37.28 %	40.48 %	38.97 %	40.78 %
LAION-Mix-16M	36.92 %	41.46 %	38.58 %	42.07 %
LAION-Rand-57M	37.62 %	40.66 %	36.99 %	39.58 %

3.3 Creating Single-Domain Datasets

We now use our domain classifiers at 98 % validation-precision to subsample LAION-200M. We obtain LAION-Natural with roughly 57 million samples and LAION-Rendition with roughly 16 million samples. Fig. 2 shows random samples from both datasets, more samples are shown in Fig. 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. 10. These datasets enable us to fairly assess CLIP's domain generalization performance in the following sections.

4 Measuring CLIP's OOD performance

Training Details For all our experiments, we train CLIP ViT-B/32 [9] 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 and hyperparameters provided by Ilharco et al. [17].

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 our cleaned versions of them (see Sec. 3.3). Fig. 3 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 also holds for CLIP models trained on even 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. 3 that adding 13 or

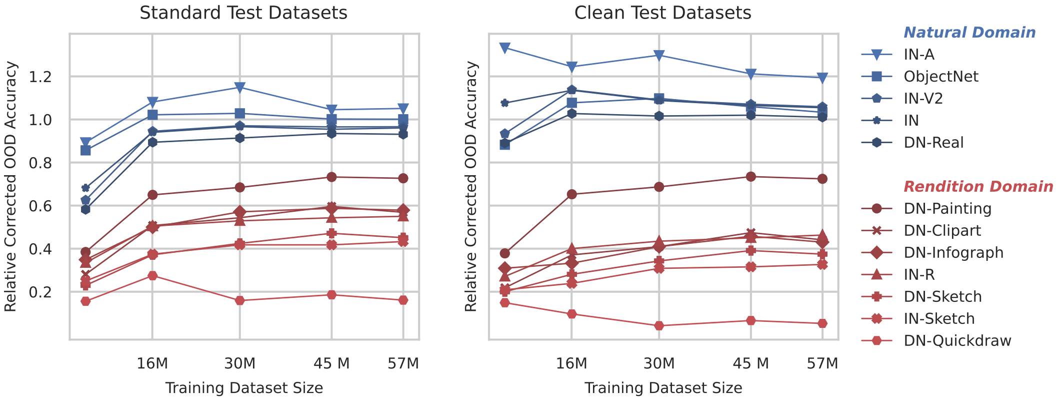

Figure 3: **Across scales, CLIP fails to generalize to unseen domains**. The *relative corrected OOD accuracy* shows performance losses or gains of a CLIP model trained exclusively on the ***natural domain*** via LAION-Natural to a CLIP model trained on a domain-contaminated dataset like LAION-200M. We evaluate models on the standard ImageNet and DomainNet test sets (left) and our cleaned versions of them (right, see Sec. 3.3). When training only on samples from the ***natural domain*** we see a decrease in performance for both standard and cleaned test datasets (i.e., relative performance < 1). This means that without samples from the ***rendition domain***, CLIP's domain generalization ability suffers significantly and consistently across scales.

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 on the *natural* and *rendition* domains. Fig. 4 shows the top-1 classification accuracy of multiple CLIP models trained on LAION-200M, LAION-Natural, LAION-Rendition, LAION-Mix, and ResNets trained on ImageNet (see Appx. H for details). We use the 13M version of LAION-Mix 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 offset of a model to the ImageNet line indicates its effective robustness. While all LAION-trained models achieve a similar effective robustness on the *natural* domain (Fig. 4 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. I. Together, the findings in this section demonstrate that CLIP's unprecedented OOD generalization performance directly results from the domain-contamination of its training distribution. Appx. D contains an analysis on understanding how different ratios of domains in the training data affect downstream performance. We defer a detailed discussion of Comparison of LAION training to ImageNet-training, Short-cut Learning, Domain Classification and Ambiguous Datapoints to Appx. E.

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 affects 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. We release all training set splits as well as pre-trained models and encourage the field to re-consider domain generalization as a central benchmark for future progress on model architectures, inductive biases, and learning objectives.

Acknowledgements

We would like to thank Vishaal Udandarao for helpful discussions and feedback. 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. We thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting PM, TW, RZ, ER, and AJ.

Author Contributions

The project was led and coordinated by PM. The method was jointly developed by PM and RSZ, with insights from WB. TW, ER, AJ, and MB participated in several helpful discussions at various stages of the project. PM, RSZ, ER, and TW contributed to the domain classifier experiments. PM did all the CLIP experiments. TW primarily wrote the manuscript with insights and contributions from PM and RSZ. TW, ER, RSZ, AJ, and PM contributed to the figures and tables.

References

[1] Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint*, 2023. URL `https://arxiv.org/abs/2303.09540`.

[2] Reza Abbasi, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. Deciphering the role of representation disentanglement: Investigating compositional generalization in clip models, 2024. URL `https://arxiv.org/abs/2407.05897`.

[3] Zechen Bai, Yuta Nakashima, and Noa Garcia. Explain me the painting: Multi-topic knowledgeable art description generation. In *CVPR*, 2021.

[4] 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*, 2019.

[5] Shai Ben-David, Koby Crammer John Blitzer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. theory of learning from different domains, 2010. URL `https://doi.org/10.1007/s10994-009-5152-4`.

[6] Wei-Ta Chu and Yi-Ling Wu. Image style classification based on learnt deep correlation features. *IEEE Transactions on Multimedia*, 2018.

[7] Benjamin Cohen-Wang, Joshua Vendrow, and Aleksander Madry. Ask your distribution shift if pre-training is right for you. *arXiv preprint*, 2024. URL `https://arxiv.org/abs/2403.00194`.

[8] Benjamin Cohen-Wang, Joshua Vendrow, and Aleksander Madry. Ask your distribution shift if pre-training is right for you. *arXiv preprint*, 2024. URL `https://arxiv.org/abs/2403.00194`.

[9] 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*, 2020. URL `https://arxiv.org/abs/2010.11929`.

[10] 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 *ICML*, 2022.

[11] Noa Garcia and George Vogiatzis. How to read paintings: semantic art understanding with multi-modal retrieval. In *ECCV*, 2018.

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

[13] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2019.

[14] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *ICLR*, 2021.

[15] 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. In *ICCV*, 2021.

[16] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021.

[17] 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.

[18] Akshay Joshi, Ankit Agrawal, and Sushmita Nair. Art style classification with self-trained ensemble of autoencoding transformations. *arXiv preprint*, 2020. URL `https://arxiv.org/abs/2012.03377`.

[19] Xuanlin Li, Yunhao Fang, Minghua Liu, Zhan Ling, Zhuowen Tu, and Hao Su. Distilling large vision-language model with out-of-distribution generalizability, 2023. URL `https://arxiv.org/abs/2307.03135`.

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[21] Zhuang Liu and Kaiming He. A decade's battle on dataset bias: Are we there yet? *arXiv preprint*, 2024. URL `https://arxiv.org/abs/2403.08632`.

[22] Prasanna Mayilvahanan, Thaddäus Wiedemer, Evgenia Rusak, Matthias Bethge, and Wieland Brendel. Does clip's generalization performance mainly stem from high train-test similarity? *arXiv preprint*, 2023. URL `https://arxiv.org/abs/2310.09562`.

[23] Orfeas Menis-Mastromichalakis, Natasa Sofou, and Giorgos Stamou. Deep ensemble art style recognition. In *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020.

[24] John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *ICML*, 2021.

[25] Bac Nguyen, Stefan Uhlich, Fabien Cardinaux, Lukas Mauch, Marzieh Edraki, and Aaron Courville. Saft: Towards out-of-distribution generalization in fine-tuning, 2024. URL `https://arxiv.org/abs/2407.03036`.

[26] Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. Quality not quantity: On the interaction between dataset design and robustness of clip. *NeurIPS*, 2022.

[27] OpenAI. Gpt-4 technical report. *arXiv preprint*, 2023. URL `https://arxiv.org/abs/2303.08774`.

[28] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, 2019.

[29] 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 *ICML*, 2021.

[30] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint*, 2022. URL `https://arxiv.org/abs/2204.06125`.

[31] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019.

[32] 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`.

[33] 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*, 2015.

[34] Catherine Sandoval, Elena Pirogova, and Margaret Lech. Two-stage deep learning approach to the classification of fine-art paintings. *IEEE Access*, 2019.

[35] 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)*, 2018.

[36] 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*, 2021. URL `https://arxiv.org/abs/2111.02114`.

[37] John Schulman, Barret Zoph, Christina Kim, Jacob Hilton, Jacob Menick, Jiayi Weng, Juan Felipe Ceron Uribe, Liam Fedus, Luke Metz, Michael Pokorny, Rapha Gontijo Lopes, Shengjia Zhao, Arun Vijayvergiya, Eric Sigler, Adam Perelman, Chelsea Voss, Mike Heaton, Joel Parish, Dave Cummings, Rajeev Nayak, Valerie Balcom, David Schnurr, Tomer Kaftan, Chris Hallacy, Nicholas Turley, Noah Deutsch, Vik Goel, Jonathan Ward, Aris Konstantinidis, Wojciech Zaremba, Long Ouyang, Leonard Bogdonoff, Joshua Gross, David Medina, Sarah Yoo, Teddy Lee, Ryan Lowe, Dan Mossing, Joost Huizinga, Roger Jiang, Carroll Wainwright, Diogo Almeida, Steph Lin, Marvin Zhang, Kai Xiao, Katarina Slama, Steven Bills, Alex Gray, Jan Leike, Jakub Pachocki, Phil Tillet, Shantanu Jain, Greg Brockman, Nick Ryder, Alex Paino, Qiming Yuan, Clemens Winter, Ben Wang, Mo Bavarian, Igor Babuschkin, Szymon Sidor, Ingmar Kanitscheider, Mikhail Pavlov, Matthias Plappert, Nik Tezak, Heewoo Jun, William Zhuk, Vitchyr Pong, Lukasz Kaiser, Jerry Tworek, Andrew Carr, Lilian Weng, Sandhini Agarwal, Karl Cobbe, Vineet Kosaraju, Alethea Power, Stanislas Polu, Jesse Han, Raul Puri, Shawn Jain, Benjamin Chess, Christian Gibson, Oleg Boiko, Emy Parparita, Amin Tootoonchian, Kyle Kosic, and Christopher Hesse. Chatgpt. `https://openai.com/blog/chatgpt`, 2022. Accessed: 2023-05-13.

[38] Yang Shu, Xingzhuo Guo, Jialong Wu, Ximei Wang, Jianmin Wang, and Mingsheng Long. Clipood: Generalizing clip to out-of-distributions, 2023. URL `https://arxiv.org/abs/2302.00864`.

[39] 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*, 2024. URL `https://arxiv.org/abs/2404.01292`.

[40] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *NeurIPS*, 2020.

[41] Haohan Wang, Songwei Ge, Zachary C. Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019.

[42] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, and Tao Qin. Generalizing to unseen domains: A survey on domain generalization. In *IJCAI*, 2021.

[43] Sheng-Yu Wang, Alexei A Efros, Jun-Yan Zhu, and Richard Zhang. Evaluating data attribution for text-to-image models. In *CVPR*, 2023.

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

[45] Yihao Xue, Siddharth Joshi, Dang Nguyen, and Baharan Mirzasoleiman. Understanding the robustness of multi-modal contrastive learning to distribution shift. In *ICLR*, 2024.

A Appendix

B Measuring the effective robustness

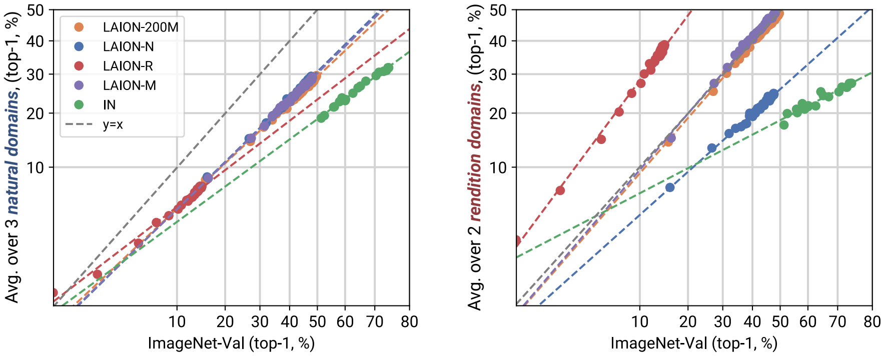

Figure 4: **CLIP's effective robustness to renditions is driven by domain-contamination**. We evaluate effective robustness [10, 40] for models trained on different LAION-200M subsets. Most notably, CLIP trained on LAION-Natural matches the effective robustness of a LAION-200M-trained CLIP on the ***natural*** 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.

C 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 [45], Fang et al. [10] 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. [22] 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. [26] discover that a model's effective robustness [10, 40] 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 [14], 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. [12] proposed to measure using Gram Matrices and which has been widely explored since then [34, 23, 35, 18, 11, 6, 3]. More recently, Cohen-Wang et al. [7] use a fine-tuned CLIP model from OpenCLIP [17] to distinguish between ImageNet and domain-shifted versions of ImageNet, such as ImageNet-Sketch, ImageNet-R, and ImageNet-V2 [31]. Wang et al. [43] and Somepalli et al. [39] develop a dataset classifier using a backbone trained by self-supervised learning and classification through retrieval via a database. Liu and He [21] report high performance when training image classifiers to distinguish between different large-scale and diverse datasets.

D Understanding Domain Mixtures

We now expand on the experiment from Tab. 3 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. 5, 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, starting from purely rendition/natural datasets, the performance steeply increases on natural/renditions shifts while remaining stable on the other domain as we slowly increase the fraction of natural/renditions samples.

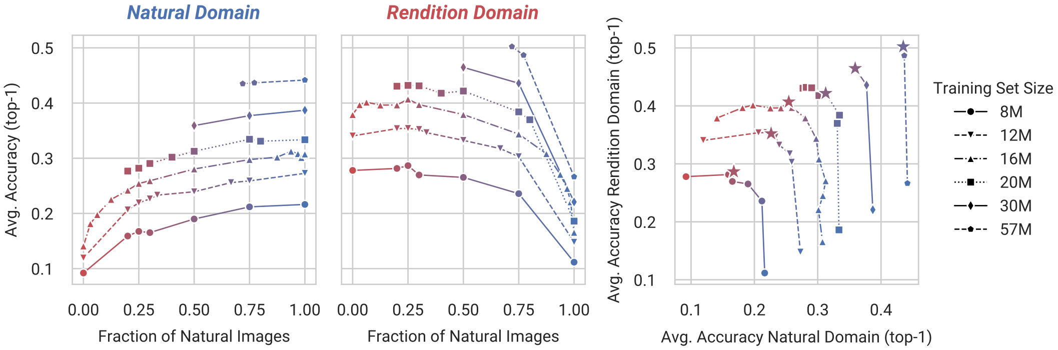

Figure 5: **Optimal data mixture transfers across scales**. We show the average accuracy on the ***natural*** and ***rendition*** domains for models trained with LAION-Mix of different absolute sizes and ratios. As expected, performance on each domain increases with the number of samples from that domain (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).

E 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 our classifier's accuracy and recall are high, it should be noted that we did no further controls of potential biases (like favoring 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.

Our work examines changes in style as one specific type of distribution shift between train and test sets. Other distribution shifts exist that are harder to define and, thus, harder to annotate and might not be easily captured by a domain classifier. Despite these challenges, we expect our main conclusion—that domain contamination explains much of the OOD performance but is generally not controlled for—to hold.

While the scale of our training sets is limited by the amount of natural/rendition samples we can find, we expect our insights drawn from over one order of magnitude of training data size to scale to even larger datasets [24, 10, 22], especially since domain contamination is a problem for rigorous evaluation independent of scale.

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.

F More Details on the Domain Classifier

We describe our labeling procedure based on this demarcation in Sec. F.1 and explore different ways to train a domain classifier on the resulting dataset in Sec. 3.1. In Sec. 3.2, 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. 3.3.

LAION-200M. 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. [1]. Both Abbas et al. [1] and Mayilvahanan et al. [22] demonstrate that CLIP trained on LAION-200M obtains comparable downstream performance while greatly reducing the computational burden of analyzing the dataset and training models from scratch.

F.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 from the *natural* or *rendition* domain. We also add an *ambiguous* class for images with elements of both domains and edge cases.

We manually label images based on a codebook derived from analyzing the existing OOD test sets, which we outline in Appx. F.2. In general, we adopt a *texture*-centric approach to distinguish renditions of a scene or object 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. 6 illustrates this demarcation on samples from LAION-200M, ImageNet test sets and DomainNet test sets.

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 [4], ImageNet-V2, ImageNet-A [16], 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 before they are annotated by a human. The annotator verifies and potentially updates pre-labels for 25 images from the same group at a time (see Fig. 7).

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. 3.2. Tab. 4 contains a detailed breakdown of labels for each data set. We show more samples grouped by domain for each data set in Fig. 22 and 33.

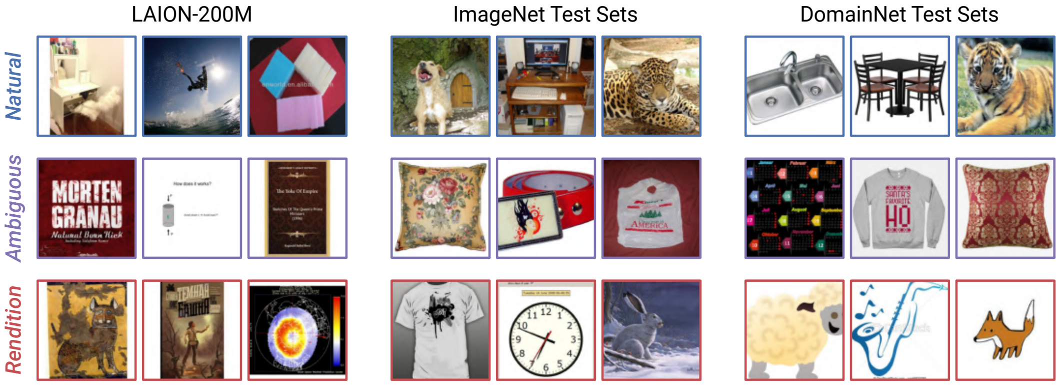

Figure 6: **Labeled *Natural*, *ambiguous*, and *rendition* samples from different data sets**. *Natural* images are photos or high-quality renders with minor filters that preserve *fine-grained textures*, while *renditions* are typically sketches, paintings, or graphics with *flat or simplified textures*. Images with elements of both, such as collages or natural images with large stylized elements, and images mainly containing text are labeled as *ambiguous*.

F.2 Labeling

As mentioned in Sec. F.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**.

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:

F.3 Training Details for the Domain Classifiers

As mentioned in Sec.3.1, we train several domain classifiers with several different training procedures. For the baselines [8, 39], we simply use the training code detailed in their works and their public code. For the FT (Finetuning) model, as mentioned in Sec. 3.1, 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.

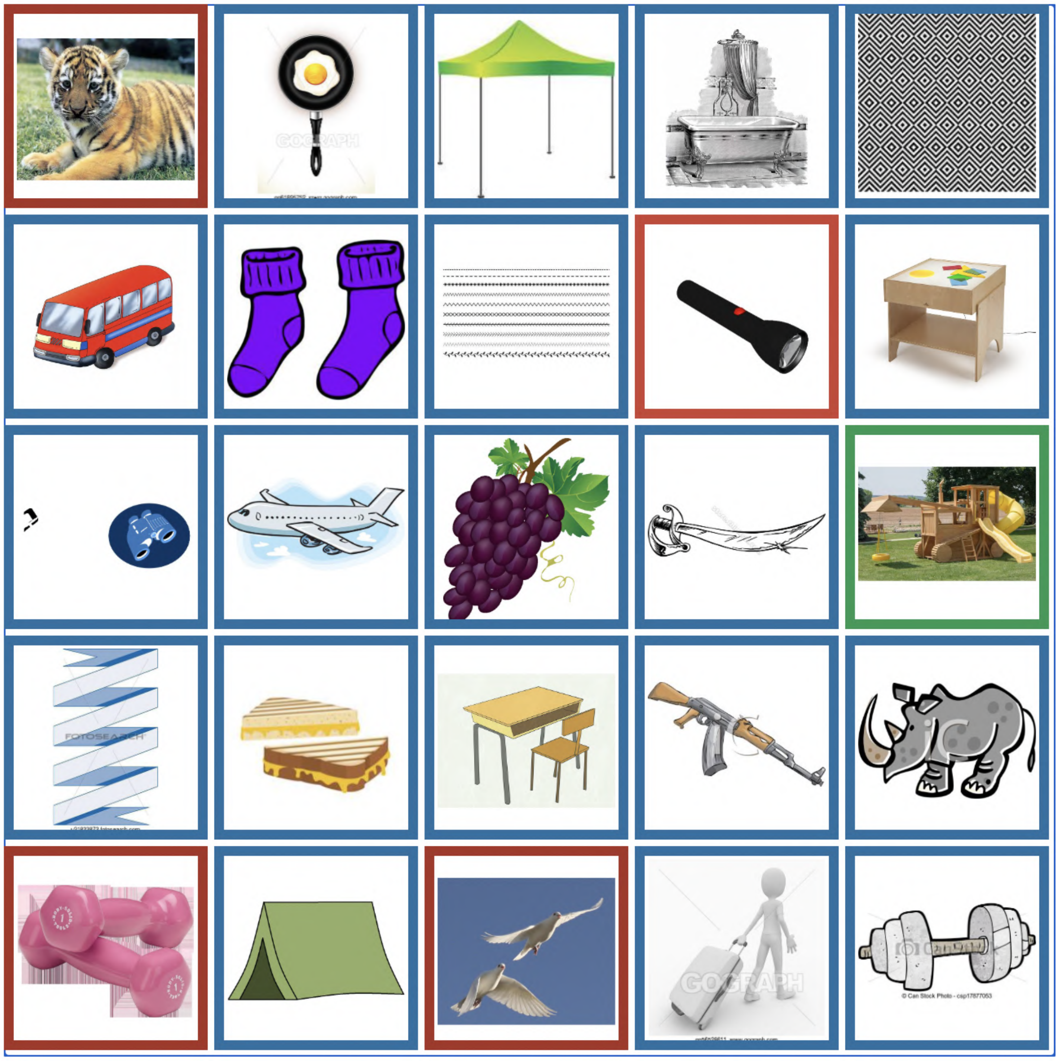

Figure 7: **Labeling setup.** By clicking on the image, the border changes to red, green, or blue, each representing natural, ambiguous, or rendition. By pressing the right or the left button the previous or next set of 25 images are rendered and the labels of the previous images are updated in a json file.

We train these models on the 13K LAION domain dataset or subsets of it with 2 or 3 classes. To compare with the models from Cohen-Wang et al. [8], 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.

F.4 Raw Domain Classifier Performance on labeled sets

In the main text in Sec.3.1 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.

Table 4: **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

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. 3.1).

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. 3.1).

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

F.5 Domain composition at different precision

We provide a detailed overview over the domain composition of datasets at standard precision in Table 7, and over the domain composition of datasets at 98% precision in Table 8.

Table 7: **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 8: **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

F.6 On the Domain Composition of [22]

Please find in Tab. 9 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. [22]. We see that at least 11-13M images are not pruned away from the datasets, therefore explaining the insignificant drop in performance.

Table 9: **Number datapoints within the dataset vs number of datapoints pruned away in Mayilvahanan et al. [22].**

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

F.7 Preparing clean datasets

In Sec. 3.3, 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. 10.

Table 10: **Clean datasets composition.** Obtained by deploying the domain classifiers from Sec.3.1 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.8 Details on Training and Choosing the Domain Classifier

Density Ratios Cohen-Wang et al. [8] 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. Again, we obtain two binary classifiers, DR-N and DR-R, that distinguish natural from non-natural samples and renditions from non-renditions, respectively.

Centroid Embeddings Inspired by the baselines used by Somepalli et al. [39], we implement a simple model (embedding model plus linear readout). Here, we take the pre-trained CLIP ViT-L/14 as the embedding model and create a linear readout by comparing embeddings to the centroid embedding of each domain. We use this as a ternary untrained nearest-neighbor classifier, dubbed CE.

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

We use the validation set to determine the two best domain 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 choose FT, the fine-tuned ternary classifier, and DR-R, the binary classifier using density ratios, to detect natural and rendition images, respectively. We use these classifiers for all subsequent experiments. Tab. 1 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 Tab. 5 and 6 in Appx. F.4.

F.9 Analysis of domain composition from Mayilvahanan et al. [22]

We also analyze the domain composition of datasets from Mayilvahanan et al. [22], 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. 8 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.

G Notes on the CLIP Models

G.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.

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

In Sec. 4, in Fig. 3, we only reported the relative numbers. Here, in Fig. 8, 10, 9, 11, we report the actual numbers as a function of dataset size.

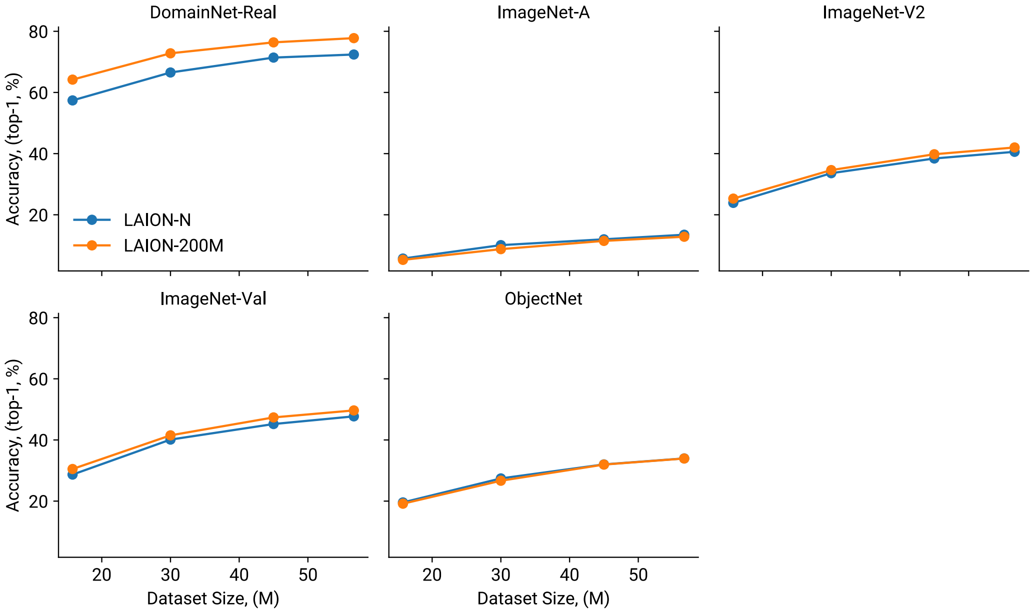

Figure 8: **CLIP trained on LAION v LAION-N performance on standard natural test sets.**

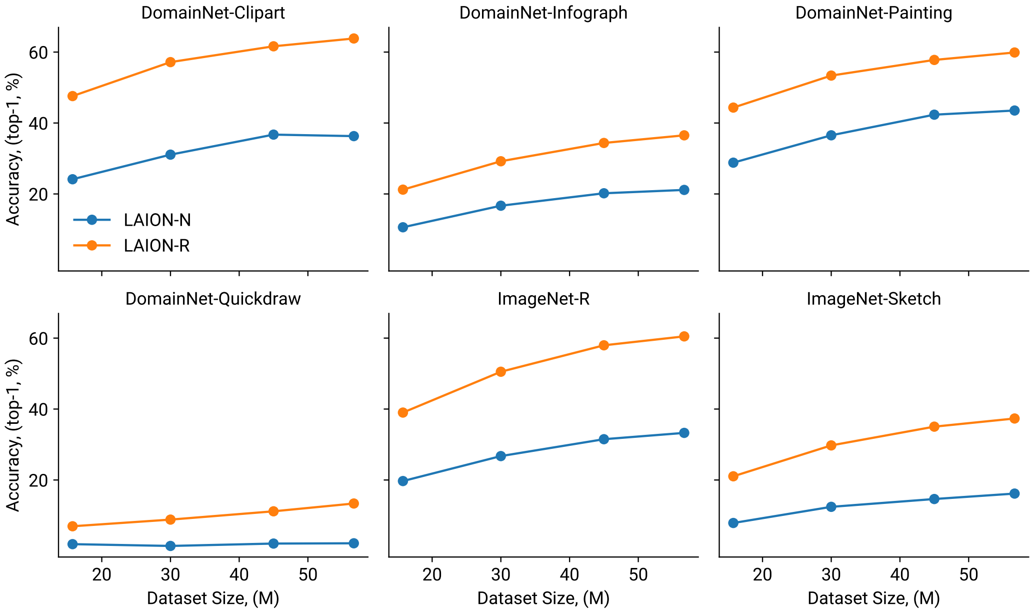

Figure 9: **CLIP trained on LAION v LAION-N performance on standard rendition test sets.**

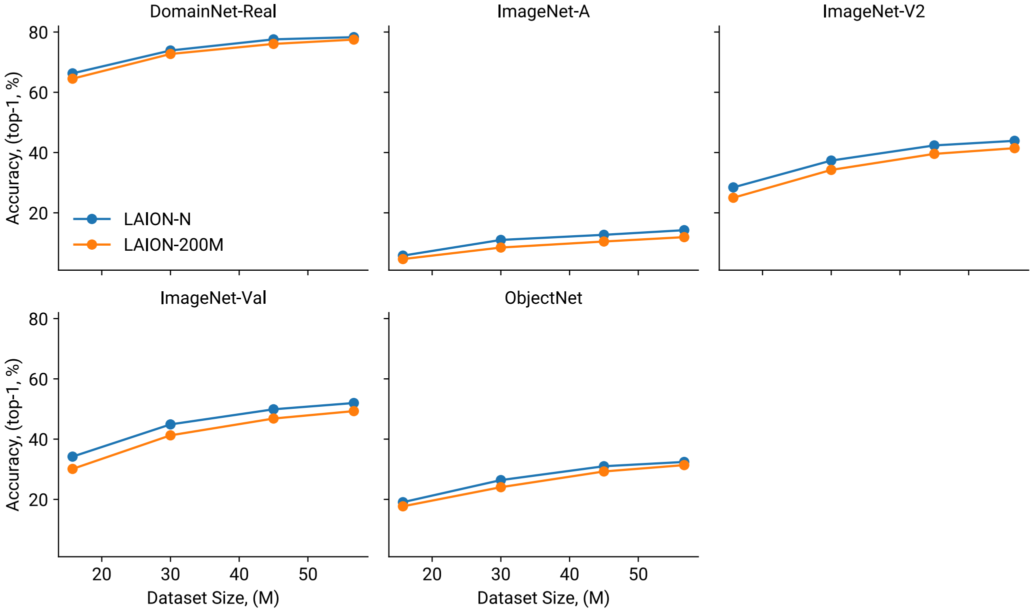

Figure 10: **CLIP trained on LAION v LAION-N performance on clean natural test sets.**

H 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. [44] and train the models for 200 epochs. We then evaluate these models on standard test sets and clean test sets from Sec.3.3. The accuracies of ResNets trained on subsets of original ImageNet is used for

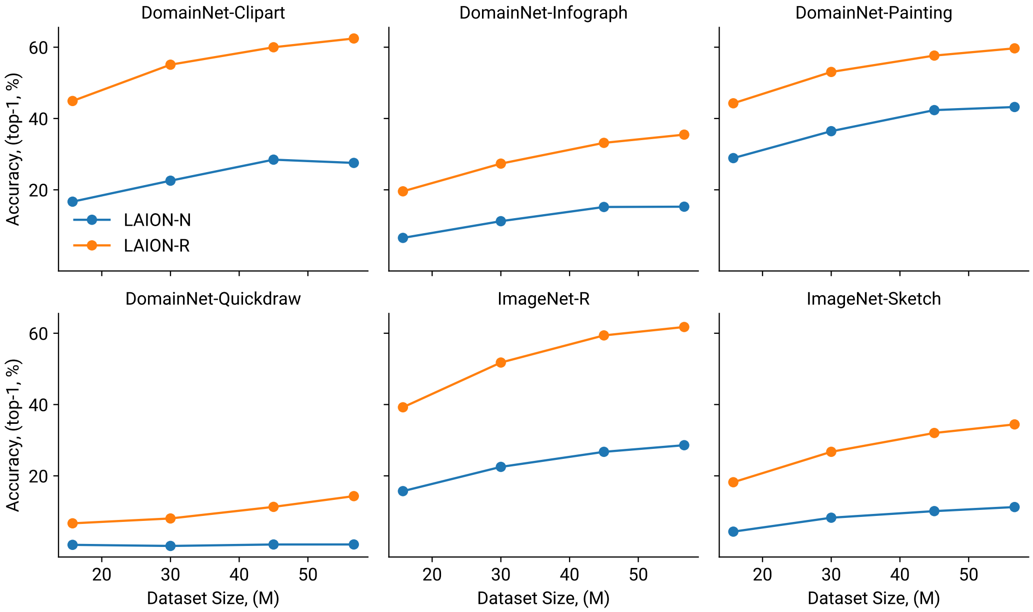

Figure 11: **CLIP trained on LAION v LAION-N performance on clean rendition test sets.**

the effective robustness plots in Sec. 4, I. 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.

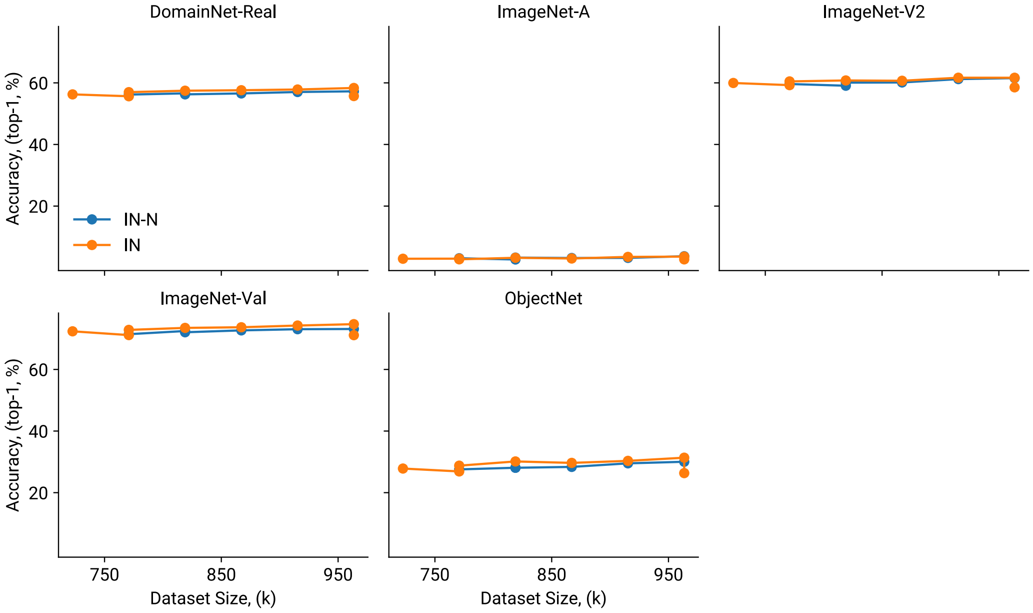

Figure 12: **Resnets trained on ImageNet v ImageNet-N performance on standard natural test sets.**

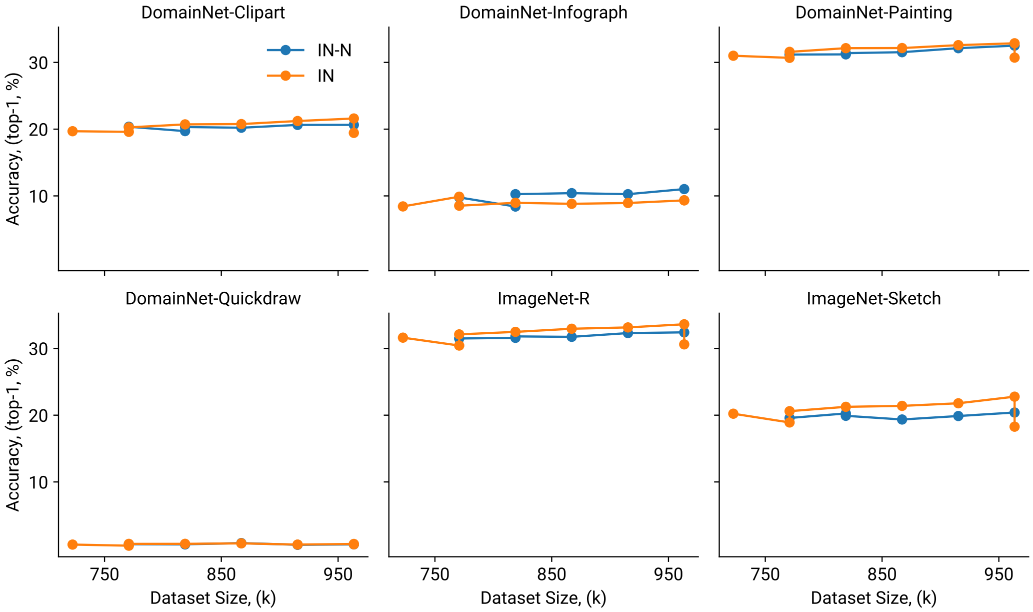

Figure 13: **Resnets trained on ImageNet v ImageNet-N performance on standard rendition test sets.**

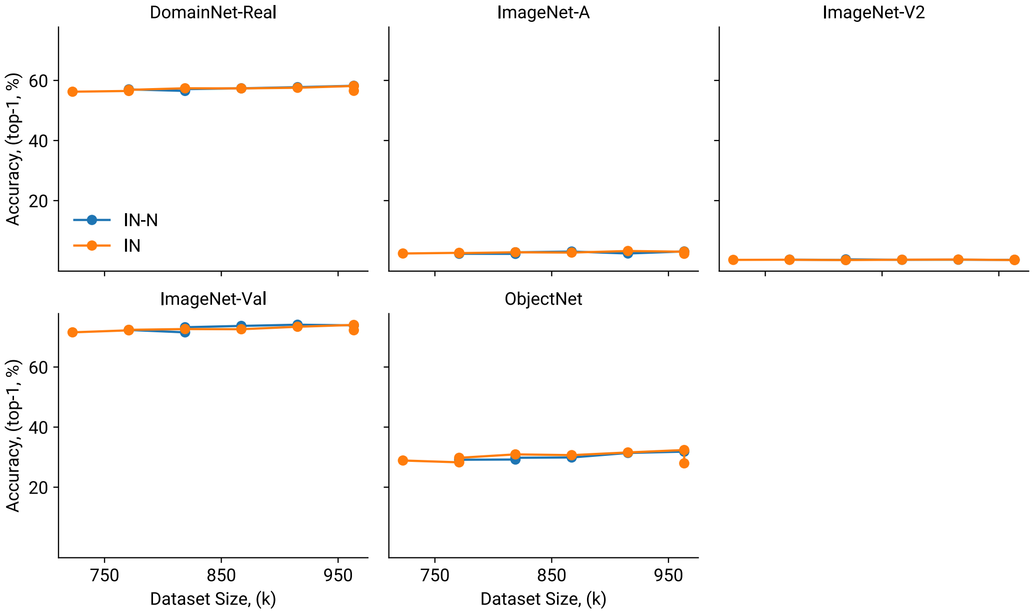

Figure 14: **Resnets trained on ImageNet v ImageNet-N performance on clean natural test sets.**

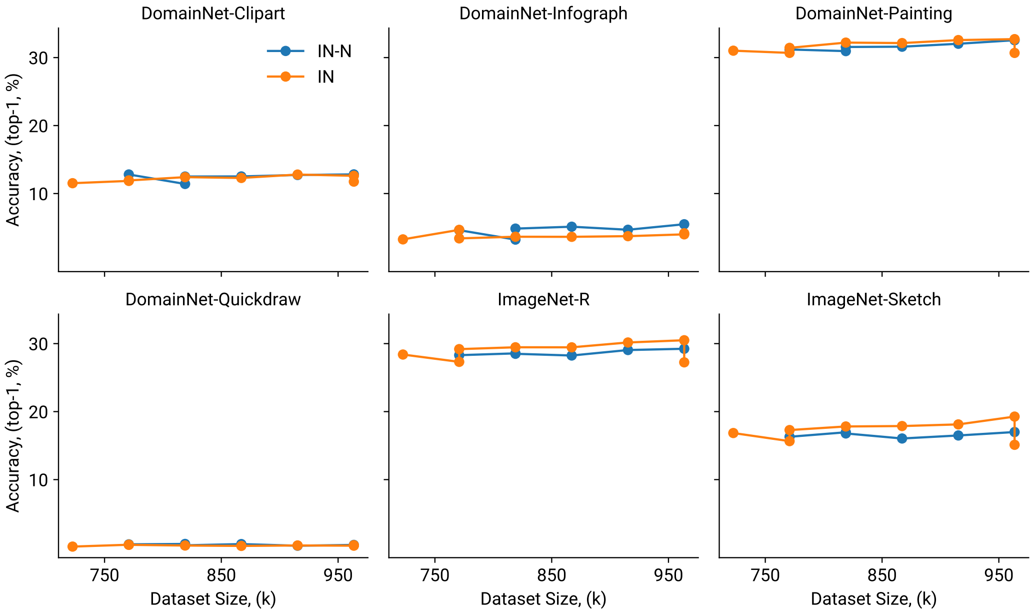

Figure 15: **Resnets trained on ImageNet v ImageNet-N performance on clean rendition test sets.**

I Detailed Effective Robustness plots on individual shifts

In Fig. 4 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, 32) 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.

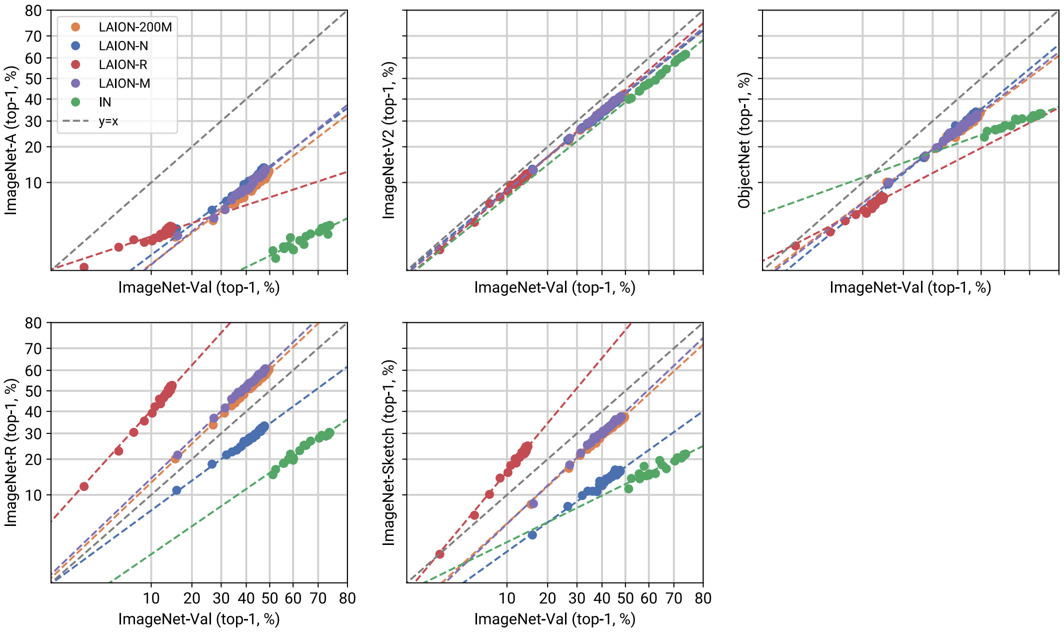

Figure 16: **Effective Robustness of different models on different ImageNet distribution shifts.** On ImageNet-R and ImageNet-Sketch (bottom row), we observe that the effective robustness of the CLIP models can be modulated by training it on the different dataset splits, i.e. LAION-Natural, LAION-Rendition,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

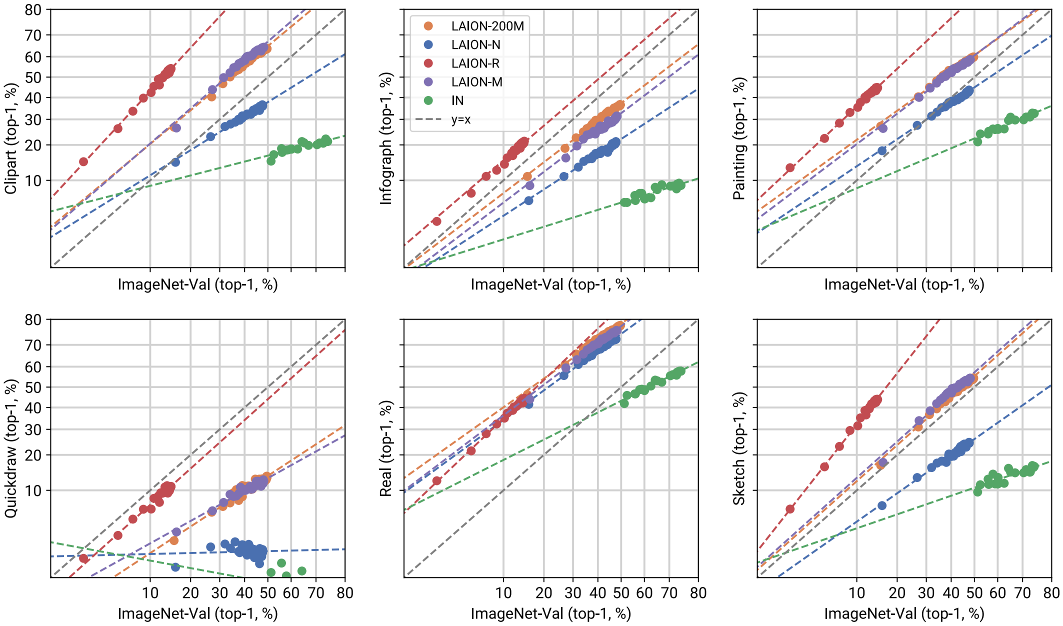

Figure 17: **Effective Robustness of different models on different DomainNet distribution shifts.** On the stylistic domains, we observe that the effective robustness of the CLIP models can be modulated by training it on the different dataset splits, i.e. LAION-Natural, LAION-Rendition, LAION-Mix. Effective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

J 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.

K Visualization of samples from the LAION dataset

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

L 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.

M 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.

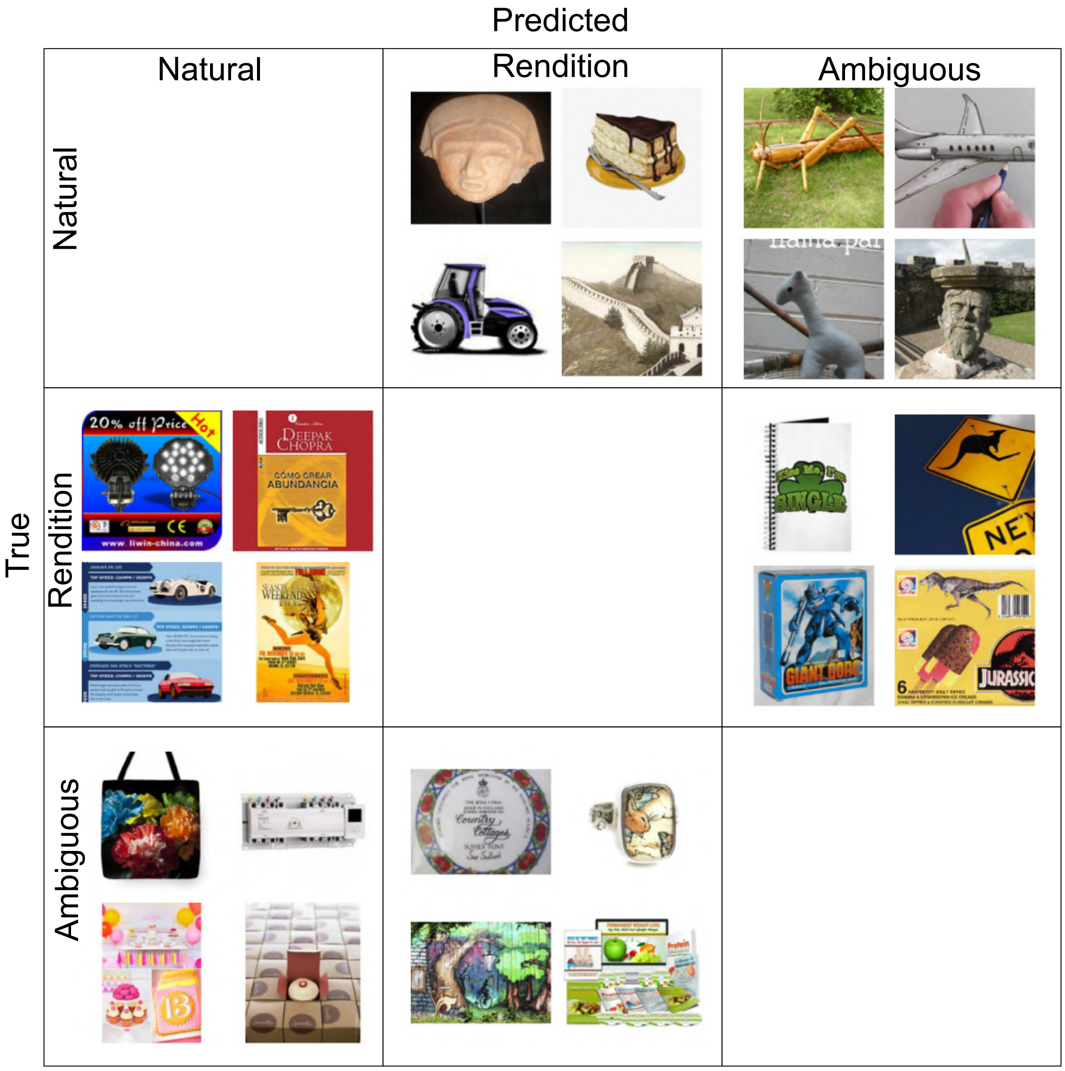

Figure 18: **Confusion matrix of example images which have been misclassified by our domain classifier.**

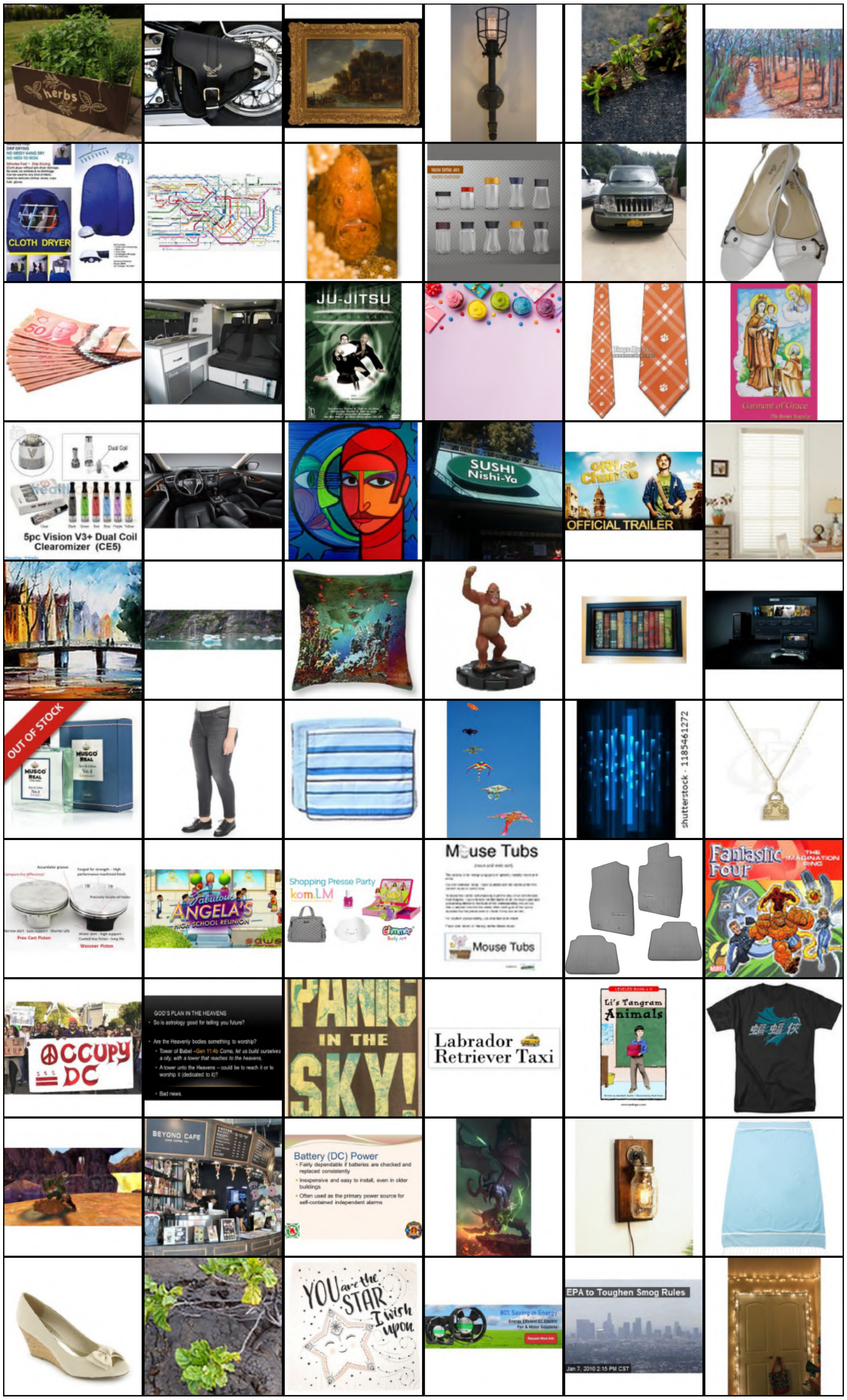

Figure 19: **Random samples from LAION-200M**. We omit NSFW images and images of humans.

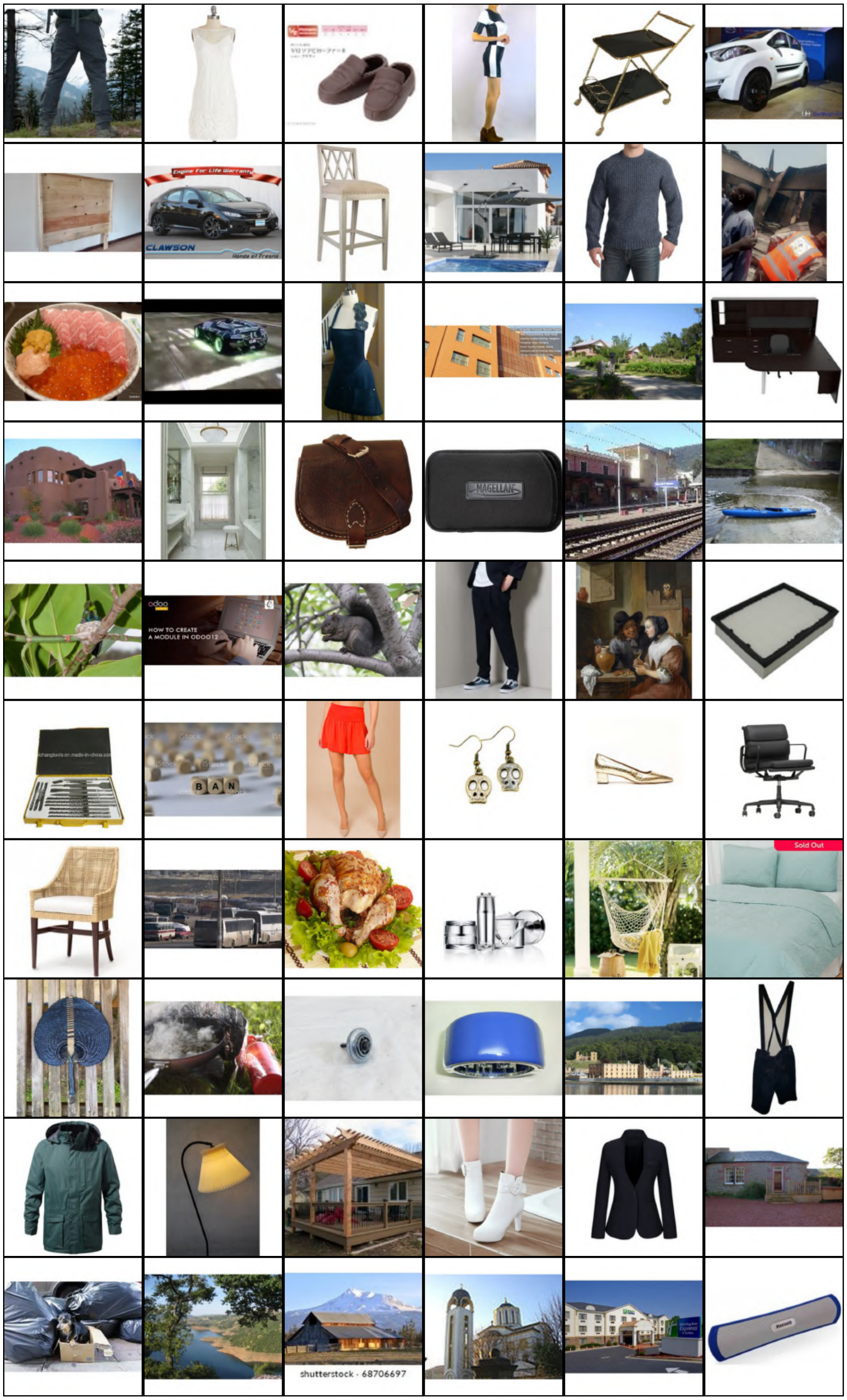

Figure 20: **Random samples from LAION-Natural**. We omit NSFW images and images of humans.

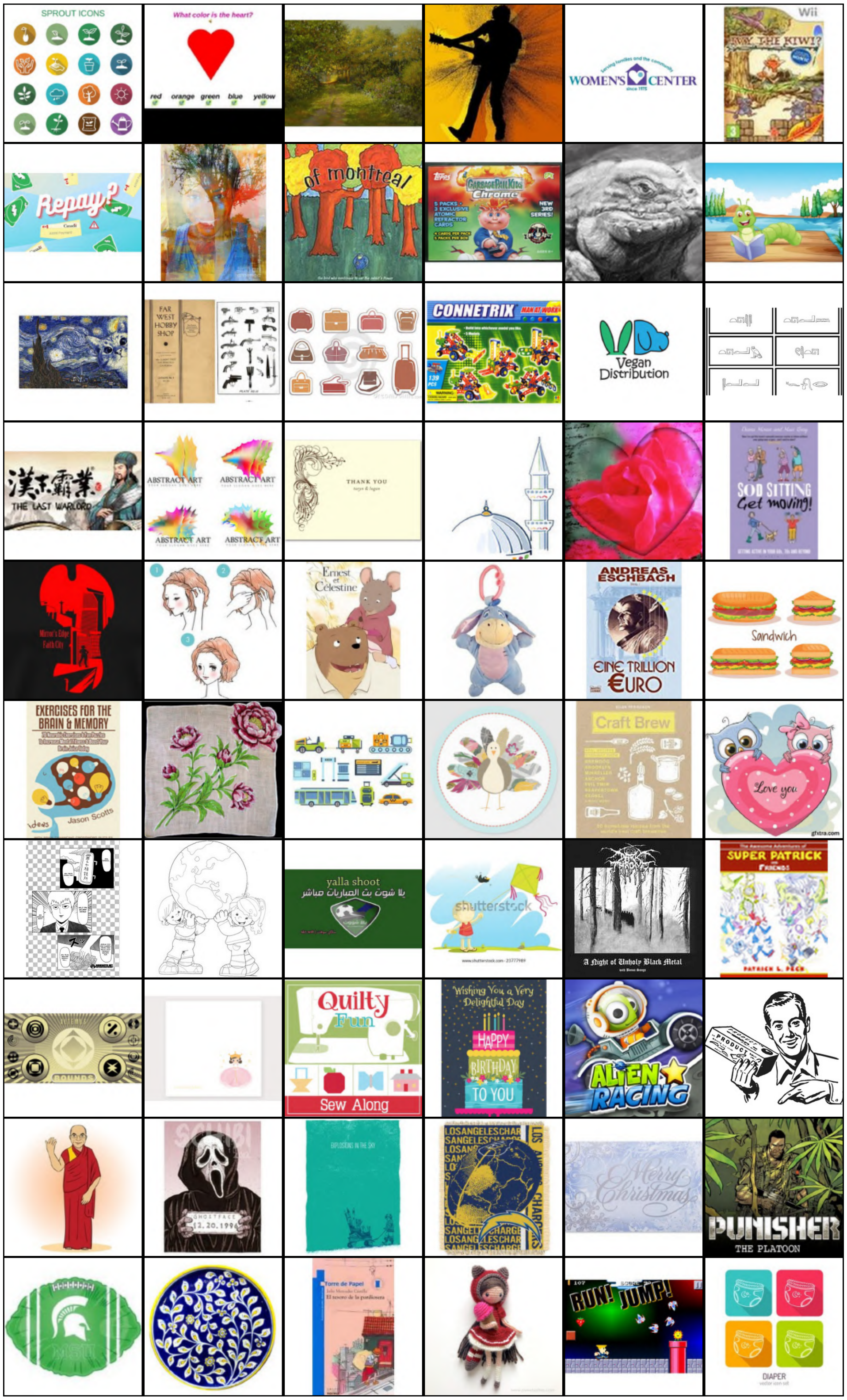

Figure 21: **Random samples from LAION-Rendition**. We omit NSFW images and images of humans.

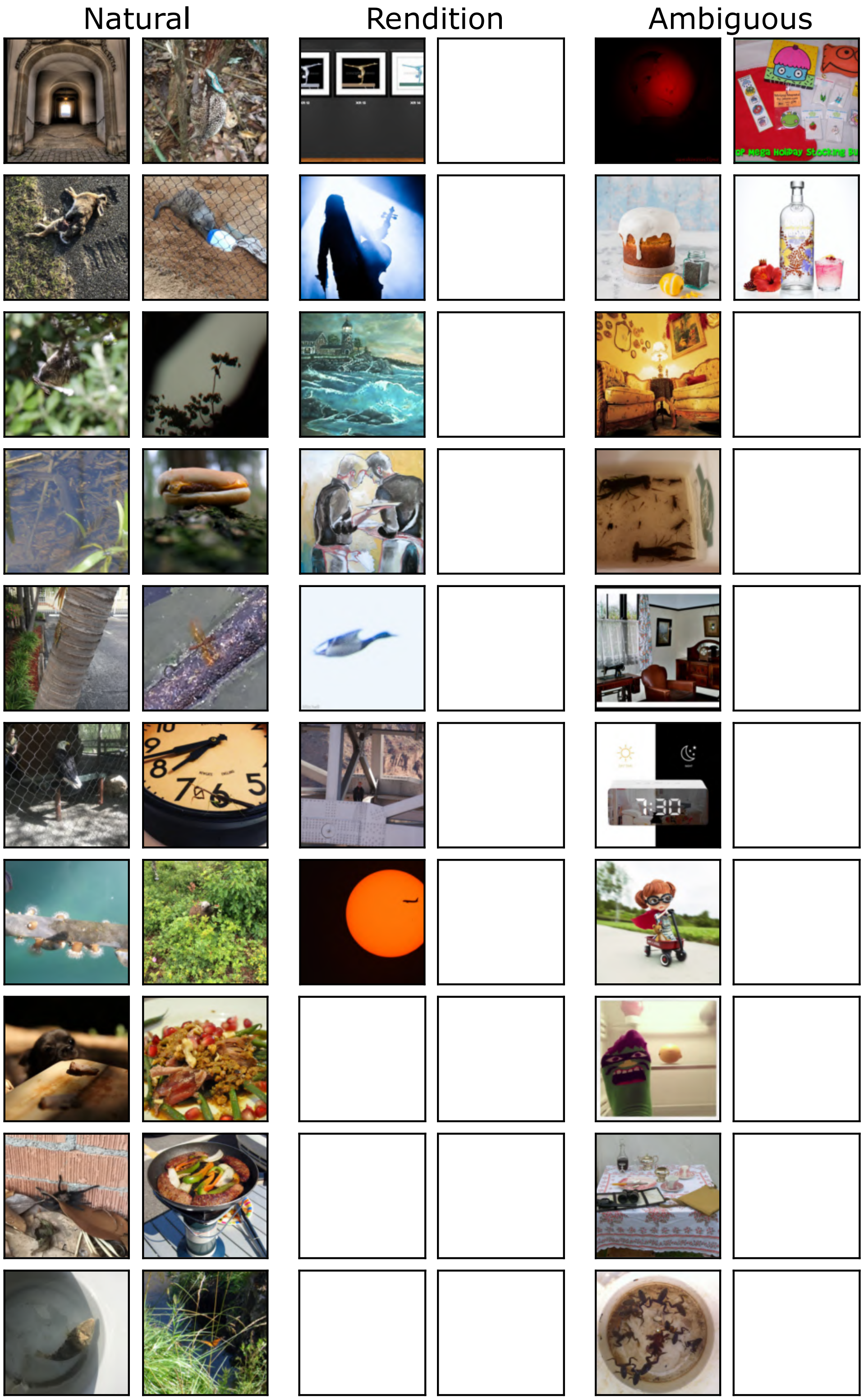

Figure 22: **Random samples of ImageNet-A grouped by domain.** We omit NSFW images and images of humans.

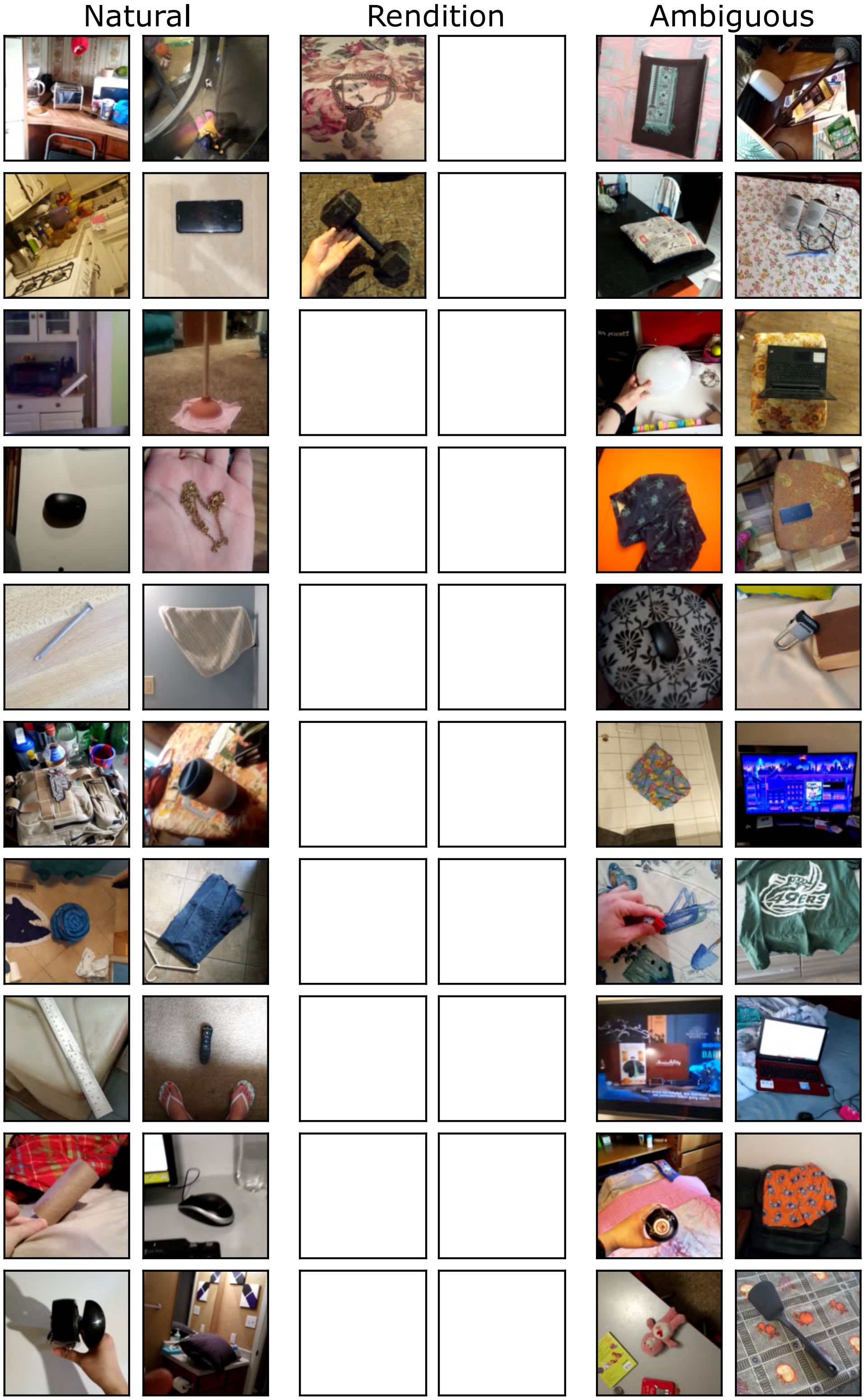

Figure 23: Random samples of ObjectNet grouped by domain. We omit NSFW images and images of humans.

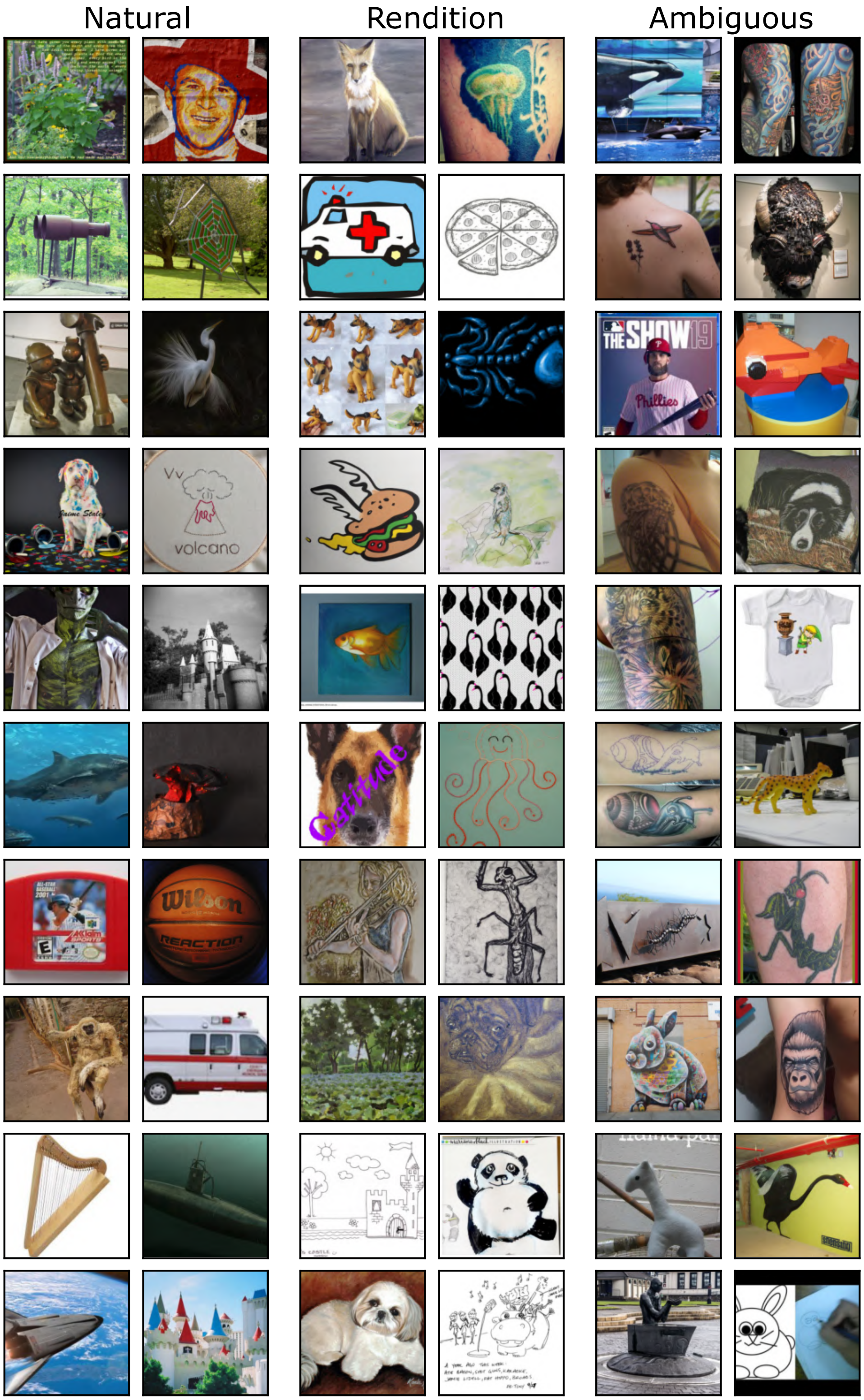

Figure 24: **Random samples of ImageNet-R grouped by domain.** We omit NSFW images and images of humans.

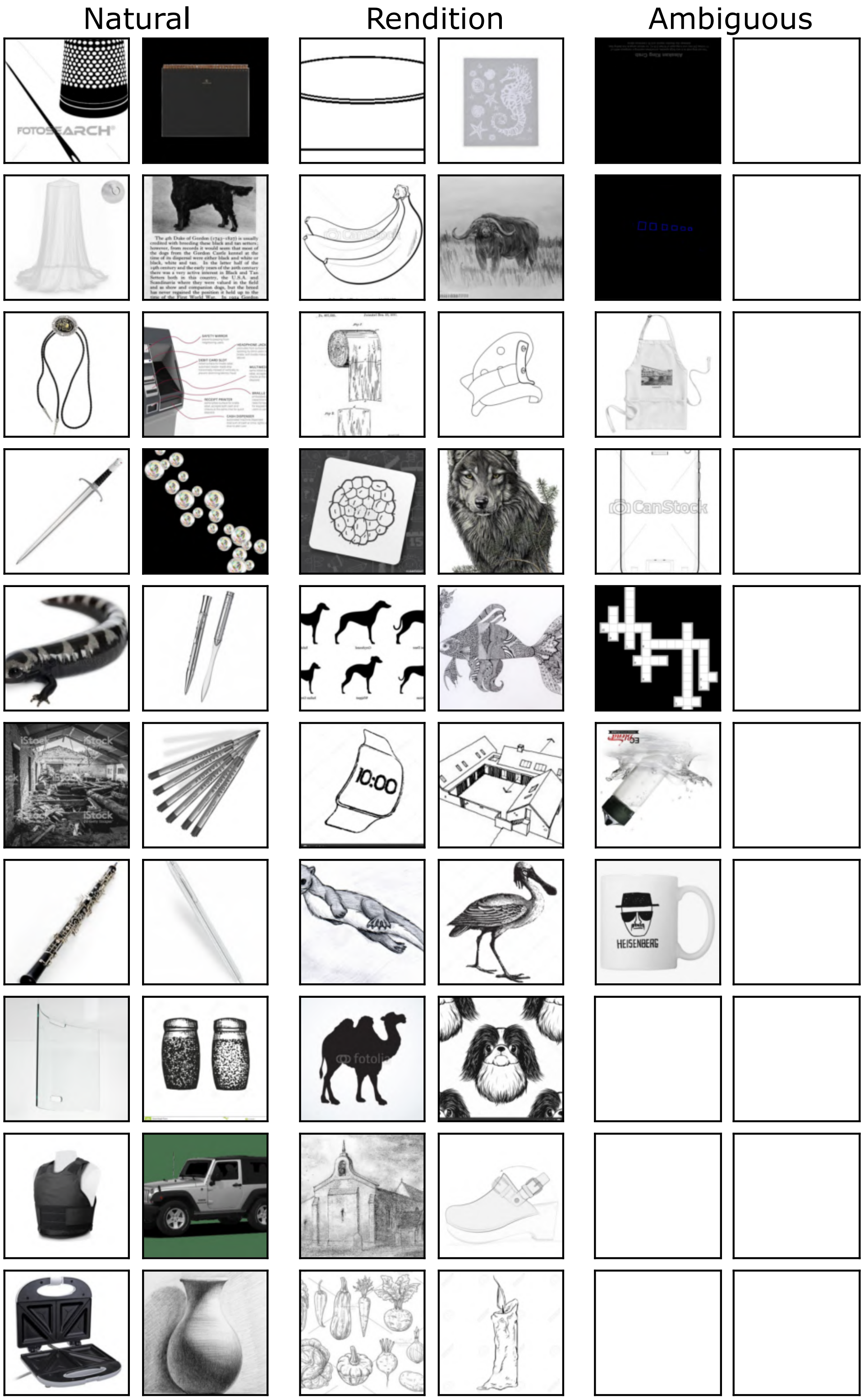

Figure 25: **Random samples of ImageNet-Sketch grouped by domain.** We omit NSFW images and images of humans.

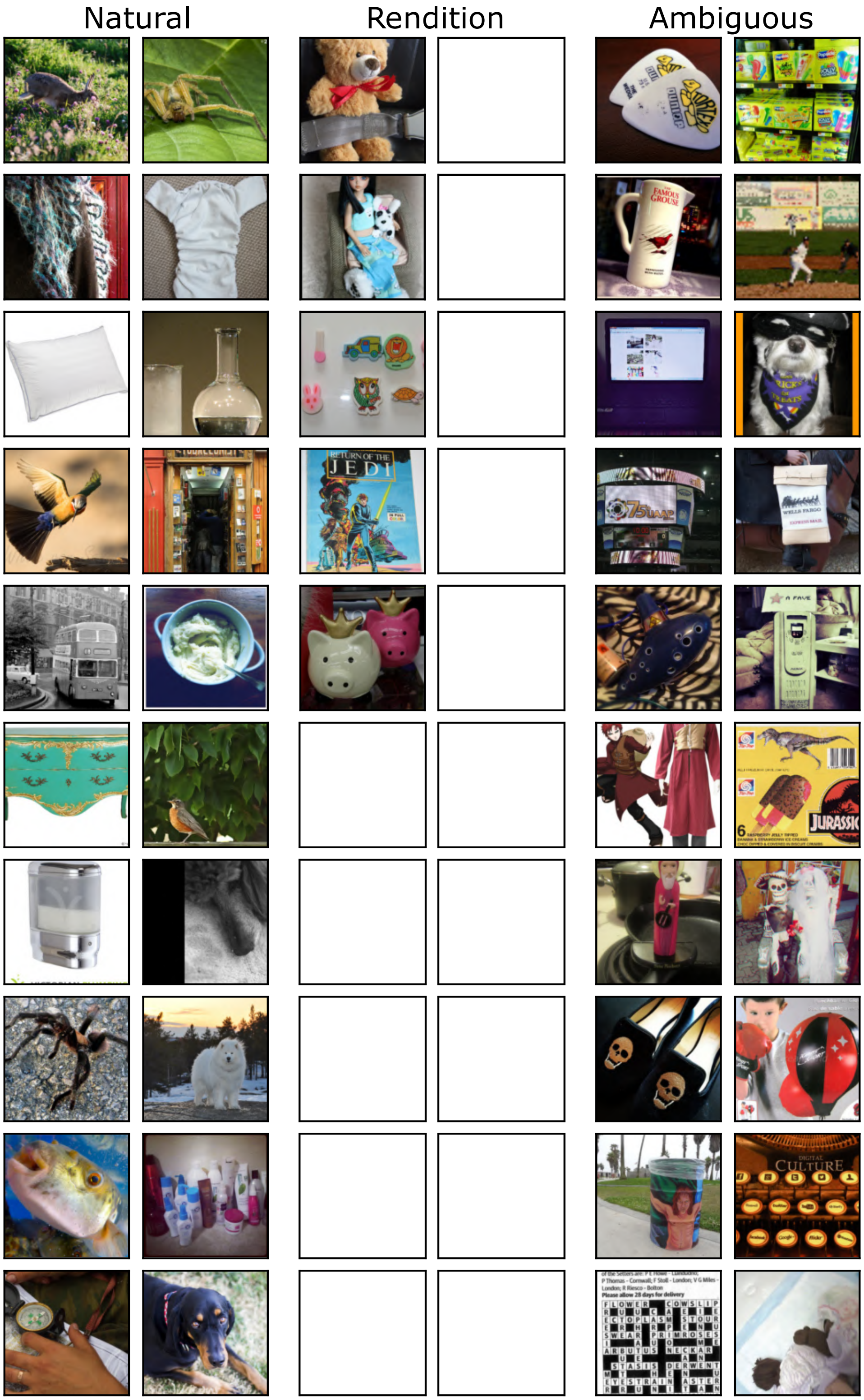

Figure 26: **Random samples of ImageNet-V2 grouped by domain.** We omit NSFW images and images of humans.

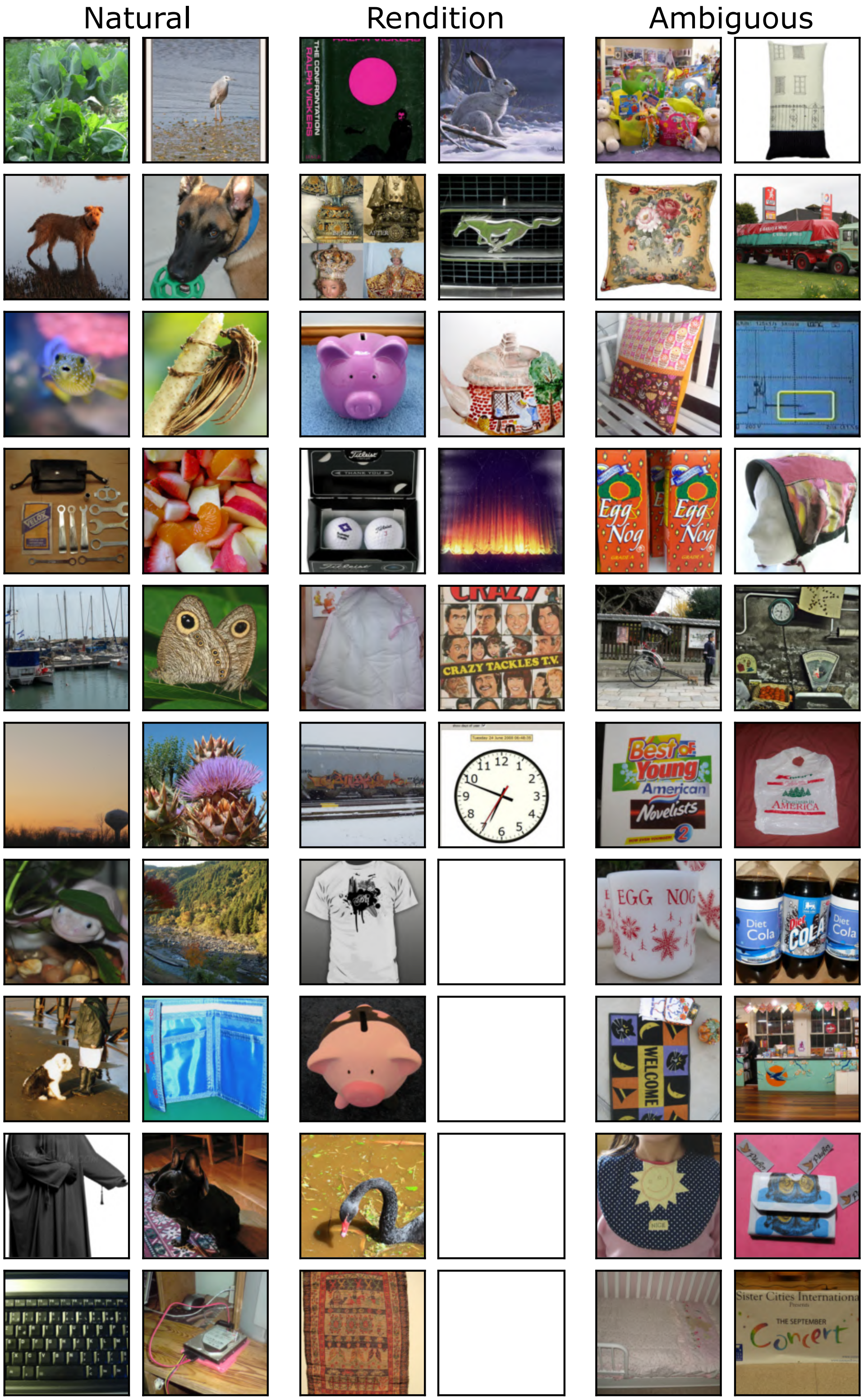

Figure 27: **Random samples of ImageNet-Val grouped by domain.** We omit NSFW images and images of humans.

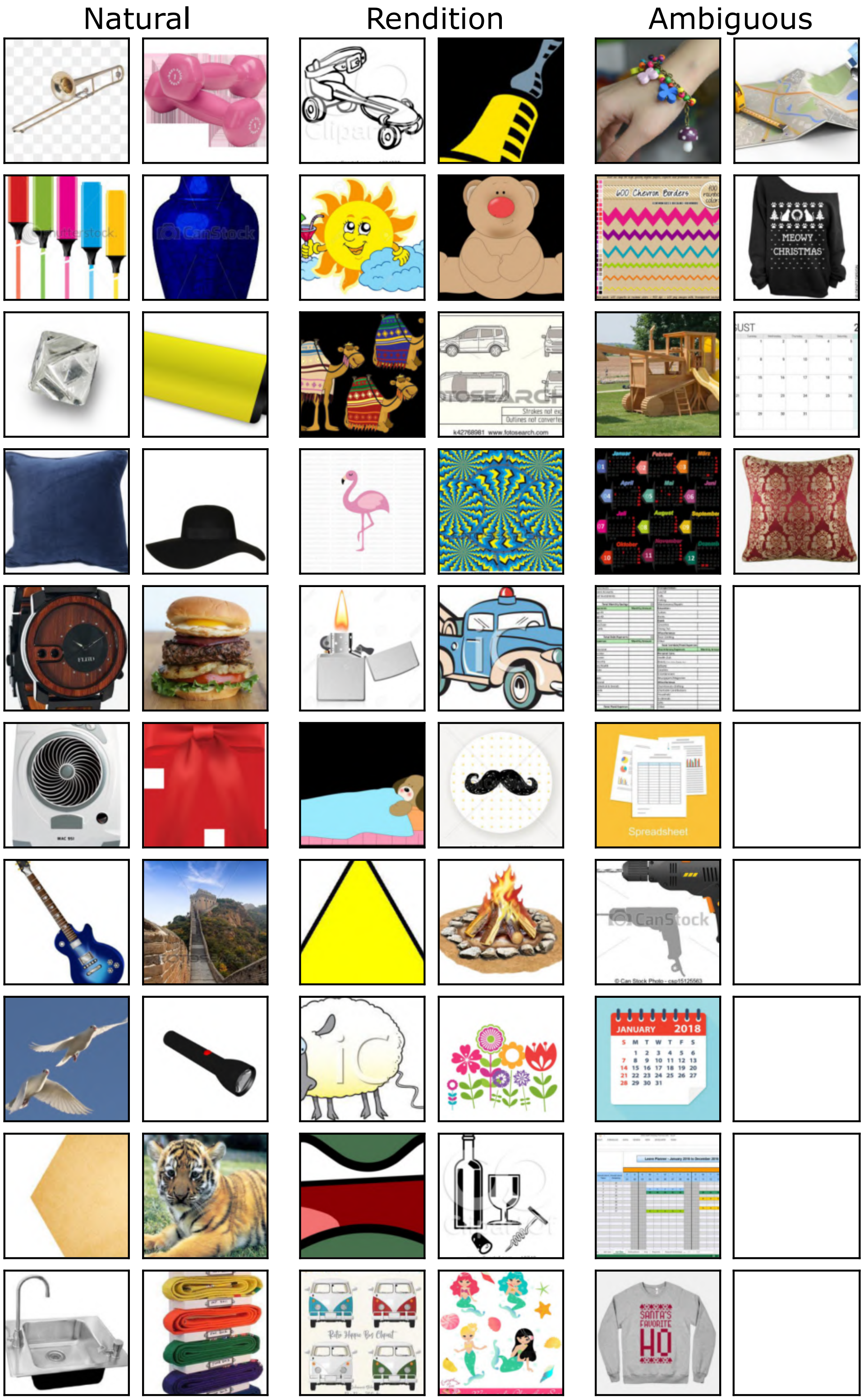

Figure 28: **Random samples of DomainNet-Clipart grouped by domain.** We omit NSFW images and images of humans.

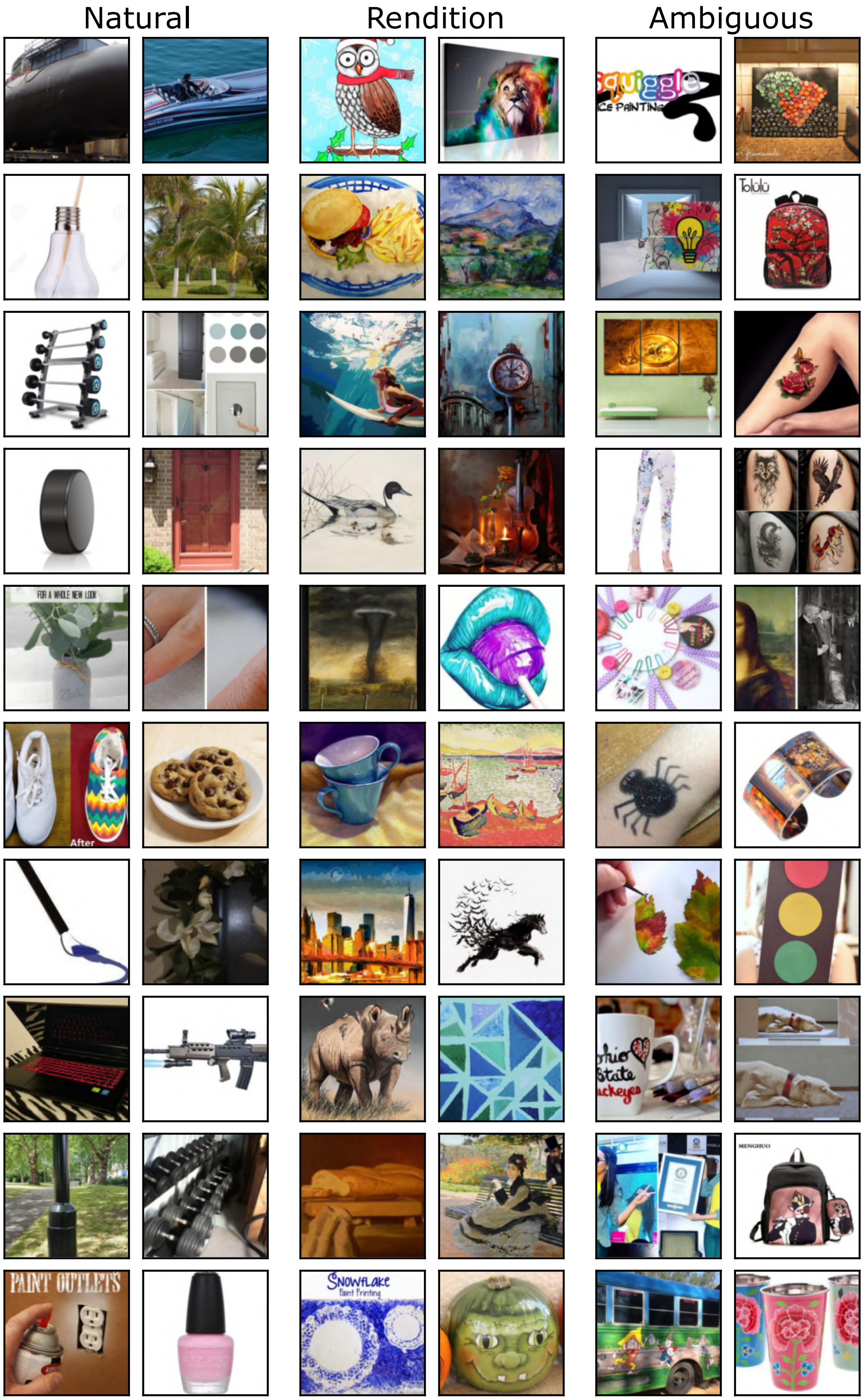

Figure 29: **Random samples of DomainNet-Painting grouped by domain.** We omit NSFW images and images of humans.

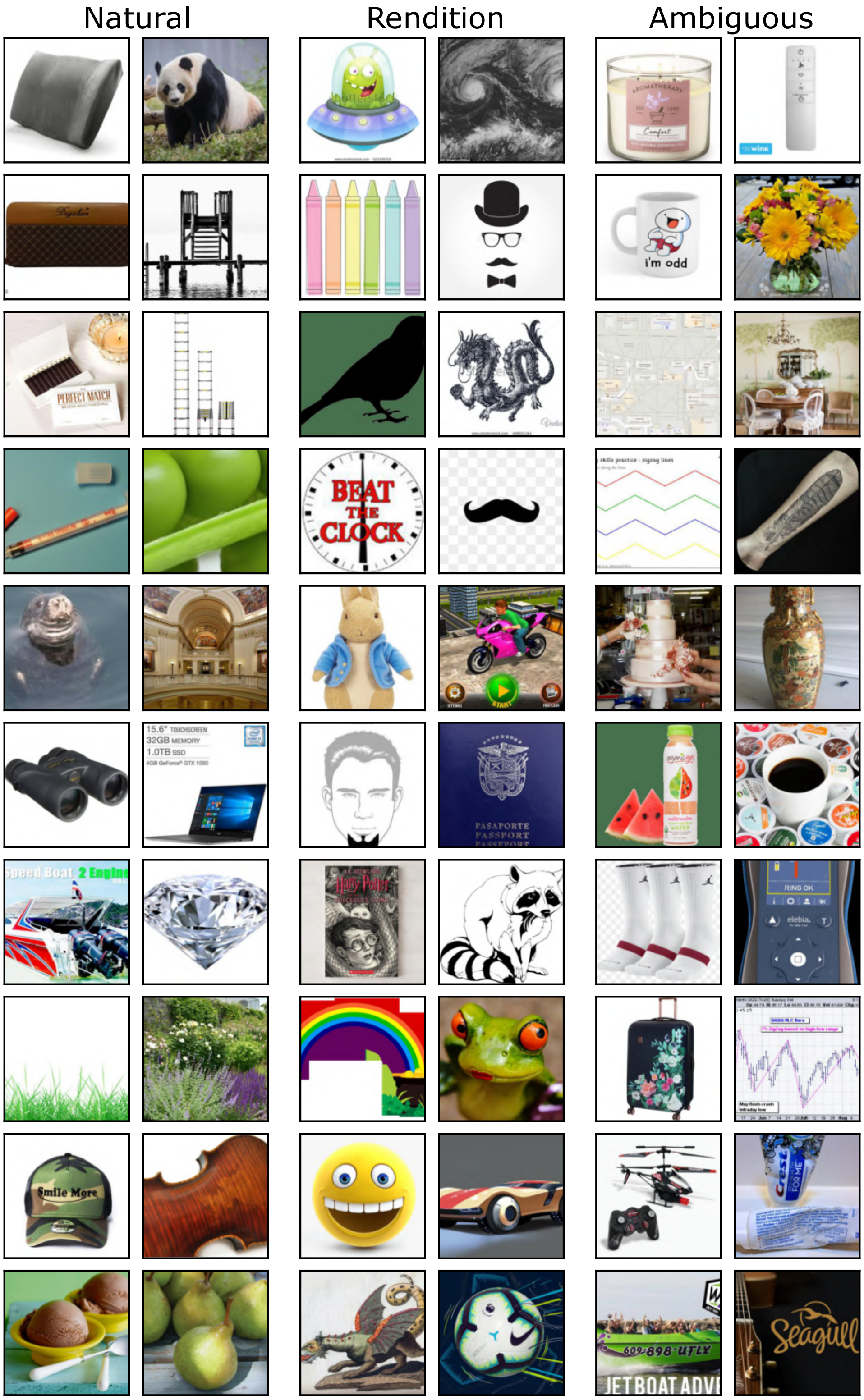

Figure 30: **Random samples of DomainNet-Real grouped by domain.** We omit NSFW images and images of humans.

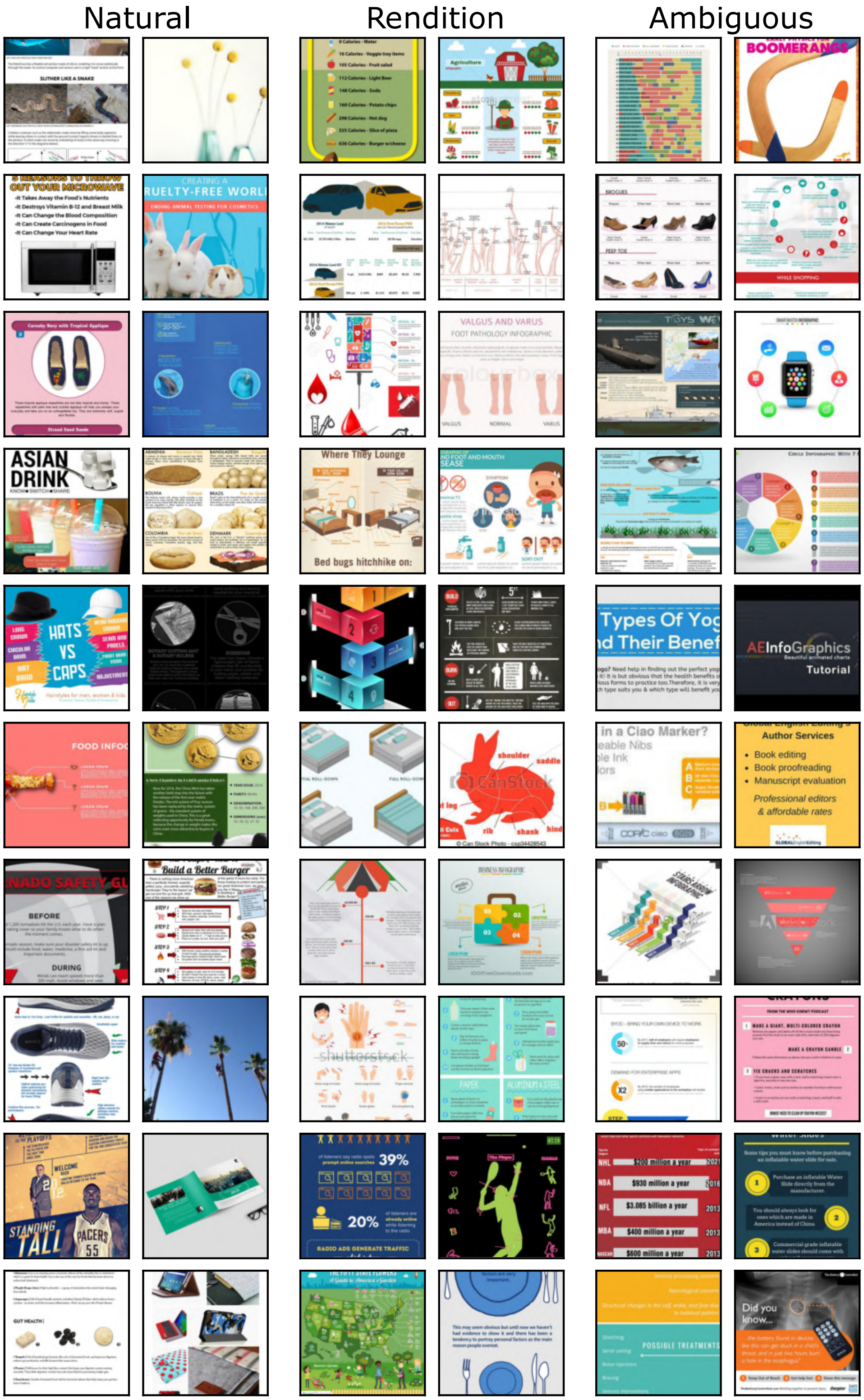

Figure 31: **Random samples of DomainNet-Infograph grouped by domain.** We omit NSFW images and images of humans.

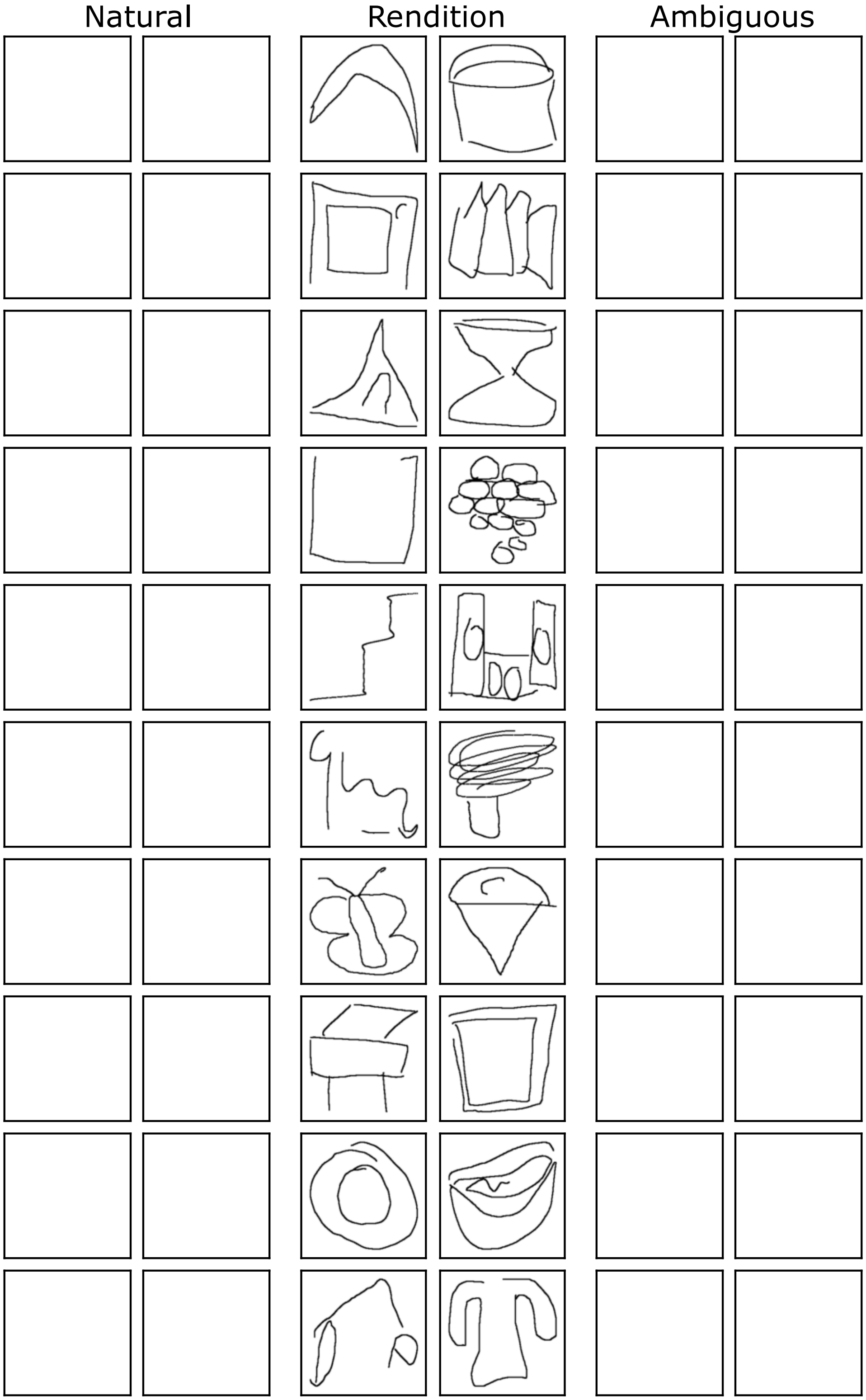

Figure 32: **Random samples of DomainNet-Quickdraw grouped by domain.** We omit NSFW images and images of humans.

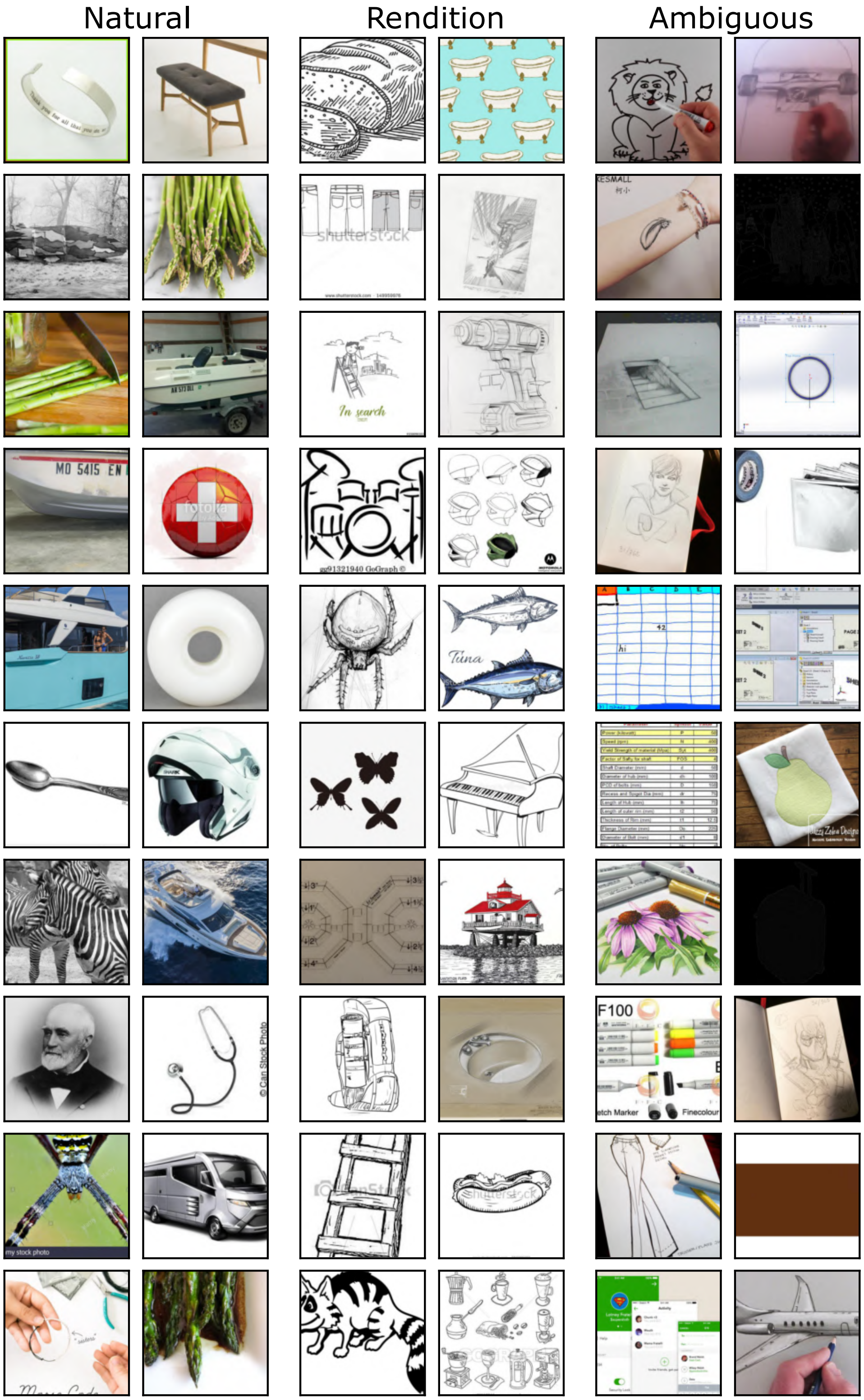

Figure 33: **Random samples of DomainNet-Sketch grouped by domain.** We omit NSFW images and images of humans.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main goals of our work as stated in the abstract or introduction is: 1. create web-scale datasets and test datasets of single domains, 2. train and test CLIP models on them and show that they do not generalize well to the test datasets, and 3. analyze the performance on mixtures of the created datasets. In Sec.3 we train, test, and deploy domain classifiers to create large datasets of several millions. In Sec. 4, we train these models on the created datasets, especially the natural version and show that CLIP model struggle to generalize. In Sec. D, we train CLIP models on mixtures of the natural and rendition datasets and state some observations. Therefore, we believe that the claims stated in the abstract and introduction are justified.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Section **??**, we identify several limitations, including potential biases in our domain classifiers, the impact of class data distribution, and the influence of ambiguous data points. These limitations are acknowledged, but they do not invalidate our core conclusions, which we believe remain significant.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not include any theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe the methodology to create all of the datasets we use in Sec. 3.3, F.1, F.2. We also sketch the training details of all our models in Sec. 3.1,4, H, F.3. This should be sufficient to reproduce all our experimental results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We have attached the code in the supplementary material and the training details are in Sec. 3.1,4, H, F.3. These together should enable anyone to reproduce all our experimental results.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The training and evaluation details of all our models is described in Sec. 3.1,4, H, F.3. We describe the methodology to create all of the used train and test datasets we use are in Sec. 3.3, F.1, F.2. This should be sufficient to understand all our experimental results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The models we train are significantly expensive, therefore we are unable to train several of them and report error bars.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The computation and memory resources spent for each of the training experiments are in Sec.F.3,G.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We uniformly conform to the Code of Ethics and, in particular, all data-related concerns about our datasets curated from LAION. We will communicate the details of the curated datasets with a license upon release, allow access to research artifacts, make our work reproducible, carefully consider all societal impacts and harmful consequences of our research output. Note that we use the LAION-400M dataset. LAION-400M has not been shown to contain any harmful child-sexual abuse material (CSAM).

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Our work is aimed to promote better understanding of the generalization capabilities of foundation models. We claim that performance on downstream tasks is directly related to (distributional) similarity between the task and the training data. We think that better understanding of whether, when and why our models generalize can enable us to build more reliable and fair models. However, of course, we cannot exclude the possibility of dual-use where malicious agents would train models on even more biased and unfair datasets to further increase the model's harmful behavior.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We train CLIP models on subsets of the LAION-400M dataset. We do not consider our models to be more unsafe than the open-source OpenCLIP models which are publicly available.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite OpenCLIP following the guidelines on their website. We cite all baselines for training the domain classifier which we have compared our domain classifier against.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: New assets generated by this paper are the datasets LAION-N, LAION-S and LAION-Mix, as well as the CLIP models trained on these datasets. We describe in detail how the dataset splits have been created and will release the code as well as the trained checkpoints upon acceptance of this paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

Justification: Our work does not paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our work does not paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.