# Visual Reasoning Requires Rethinking Vision-Language Beyond Scaling

**Haider Al-Tahan**[1], **Quentin Garrido**[1,2], **Randall Balestriero**[3],
**Diane Bouchacourt**[1], **Caner Hazirbas**[1], **Mark Ibrahim**[1]
[1]Meta FAIR, [2]Univ Gustave Eiffel, CNRS, LIGM, [3]Brown University

## Abstract

Significant research efforts have been made to scale and improve vision-language model (VLM) training approaches. Yet, with an ever-growing number of benchmarks, researchers are tasked with the heavy burden of implementing each protocol, bearing a non-trivial computational cost, and making sense of how all these benchmarks translate into meaningful axes of progress. To facilitate a systematic evaluation of VLM progress, we utilize `UniBench`: a unified implementation of 50+ VLM benchmarks spanning a comprehensive range of carefully categorized capabilities from object recognition to spatial awareness, counting, and much more. We measure progress by evaluating nearly 60 publicly available vision-language models, trained on scales of up to 12.8B samples. We find that while scaling training data or model size can boost many vision-language model capabilities, scaling offers little benefit for reasoning or relations. Surprisingly, we also discover today's best VLMs struggle on simple digit recognition and counting tasks, e.g. MNIST, that can be solved with much simpler networks. Where scale falls short, we find that more precise interventions, such as data quality or tailored-learning objectives offer more promise. For practitioners, we also offer guidance on selecting a suitable VLM for a given application.

## 1 Introduction

Pre-training visual models with language supervision, as demonstrated by CLIP Radford et al. [2021a], has become a powerful and accessible method for multimodal representation learning. VLMs have shown remarkable flexibility, excelling in zero-shot classification, transfer learning Radford et al. [2021a], text and image retrieval Goel et al. [2022], Cui et al. [2022], Radenovic et al. [2023], robustness Yun et al. [2023], and understanding compositional relationships Yuksekgonul et al. [2023], Thrush et al. [2022], Hsieh et al. [2024]. Despite these successes, the field still lacks a unified dataset to evaluate VLM performance across a wide array of benchmarks and model types. This fragmentation makes it difficult for researchers to draw clear conclusions about the best practices for further advancing VLM capabilities. Therefore, there is a pressing need for a comprehensive evaluation framework to address this gap and guide future research and development in visual representation learning.

To help shed light into the landscape of VLMs, we evaluate nearly 60 openly available vision-language models spanning a range of architectures, model sizes, training dataset scales, and learning objectives with scales of up to 12.8B samples and 1B parameters. Those 59 evaluated on 53 benchmarks, these benchmarks cover a range of VLM capabilities from standard object recognition to spatial understanding, counting, geographic robustness, domain-specific medical and satellite imagery, and many others. With such a comprehensive set of benchmarks, we shine a light on the blind spots in the strengths and weaknesses of the model. Next, to ensure that the research community can translate the many resulting metrics into meaningful axes of progress, we categorize these benchmarks into seven

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

types and seventeen finer-grained capabilities. Researchers can quickly pinpoint model strengths and weaknesses in a comprehensive, apples-to-apples fashion.

We find that scaling, model size, or training data is a powerful lever for many axes of performance, but offers little benefit for visual relations and reasoning. We also find today's best VLMs struggle with simple benchmarks involving numerical comprehension, even with the right training data, on tasks such as character recognition or counting—including decades old benchmarks such as MNIST and SVHN [LeCun et al., 1998, Netzer et al., 2011]. Where scale falls short, we find tailored learning objectives and training data quality are promising levers for relations and reasoning. Finally, we provide practical recommendations on which models practitioners should select. For example, we find large open models such as Eva ViT-E/14 to be a good choice for a general-purpose VLM while models such as NegCLIP excel at specialized tasks such as visual relations.

# 2 Evaluation Setup

Here we describe the benchmarks, protocols, and axes of progress that comprise `UniBench` as well as the VLMs evaluated.

## 2.1 VLMs Considered

We evaluate 59 openly available VLMs across a range of model sizes, pre-training dataset sizes, learning objectives, and architectures (full list in Appendix Table 5). For traning dataset size, we include models trained and/or fine-tuned with datasets ranging from 13 million to 12.8 billion samples; including DataComp [Gadre et al., 2023] (small, medium, large, and extra-large), LIAON [Schuhmann et al., 2022] (400M, 2B, 5B), MetaCLIP [Xu et al., 2023] (400M and 2.5B), Flickr [Young et al., 2014], PMD [Singh et al., 2022], and COCO [Lin et al., 2015]. For model size and architecture, we categorize models based on the number of parameters and whether these models are convolutional or transformer-based models, ranging from ResNet50 [He et al., 2016] with 38 million parameters to EVA02 ViT E [Fang et al., 2023b] with 4.3 billion parameters.

**Evaluation Procedure** We evaluate performance of zero-shot classification benchmarks similar to [Radford et al., 2021b], by contrasting the representations of class labels (averaged across prompts as defined by Cherti et al. [2022]) with the image representations and using the class with the highest probability as the predicted class. For relation benchmarks, we follow the standard protocol of contrasting correct and incorrect captions with image representations.

# 3 Gauging progress in Vision Language Modeling

We show the overall median performance of the nearly 60 VLMs we examined on 53 benchmarks in Figure 1 ranked by their zero-shot classification performance. The results suggest that, while VLMs perform remarkably well on many tasks, for others, VLM performance is near or below random chance level. These results highlight the need for a unifying pipeline to systematically surface model limitations.

## 3.1 Scaling improves many benchmarks, but offers little benefit for reasoning and relations

**Scaling training dataset size hardly helps for reasoning and relations.** While scaling training dataset size improves performance across many tasks, this trend does not hold for benchmarks assessing relation understanding and reasoning capabilities. To control for other confounding factors, we fix the model and only vary the training data size in Figure 2. The results suggest despite increasing the training dataset size by a factor of $1000\times$, relational and reasoning benchmarks performance is fairly flat compared to the significant boost in performance on other tasks. We observe a similar trend overall when we include all 59 models in Appendix Figure 6. We specifically pinpoint capabilities such as Depth Estimation, Spatial Understanding, Counting, Scene and Text Recognition, as the underlying capabilities where scale does not lead to improvements as shown in Figure 3.

**Scaling model size also offers little to no benefit for reasoning or relations.** When we scale models' size from 86 million parameters to 1 billion parameters, we also find that models struggle to

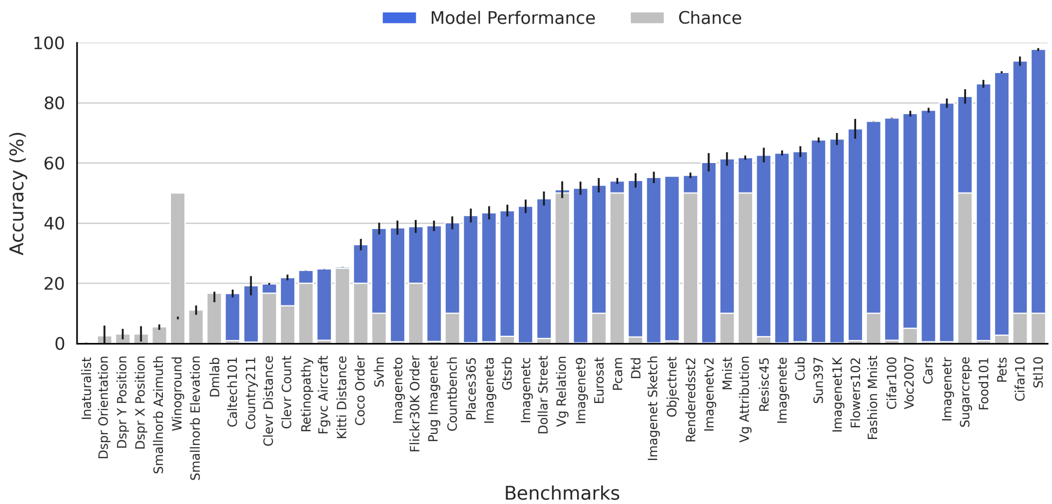


Figure 1: **Median performance of all 59 VLMs on 53 benchmarks, illustrating that despite advancements, VLMs still struggle on several benchmarks.** Benchmarks that barely exceed chance-level performance include Winoground, iNaturalist, DSPR, Small Norb, dmlab, Clevr, PCam, Renderedssst2, and Kitti. Blue bars represent the median zero-shot performance of the models, while grey bars indicate the chance-level for each benchmark.

scale on similar proportions on relation and reasoning tasks as shown in Figure 2. While for other benchmark types including object recognition, robustness, etc. performance improves by 17.4% as model size scales by $11\times$, relations and reasoning improve by a modest 3.41% with a fairly flat scaling curve. Similar to scaling training dataset size, scaling model size also offers little benefit for capabilities such as Depth Estimation, Spatial Understanding, Counting, Scene and Text Recognition as shown in Figure 3.

### 3.2 A Case Study: Digit Recognition and Counting are notable limitations for VLMs even with the right training data

A surprising aspect of VLMs is their poor performance on benchmarks that are traditionally considered straightforward, such as MNIST, CIFAR-10, and CIFAR-100, as shown in Figure 1. For example, a simple 2-layer MLP achieves 99% accuracy on MNIST [Wan et al., 2013] significantly outperforming all 59 VLMs we studied. To delve deeper into this unexpected result, we controlled for several variables:

1. **VLM confusions go beyond top-1:** To further understand the performance results on MNIST, we compute more generous top-2,-3,-4, and -5 accuracy measures to understand whether models confuse similar digits. We show in Appendix Figure 8 that even when we compute top-5 accuracy (with 50% being chance), VLMs barely reach 90% accuracy suggesting poor performance is not due to minor confusions among digits.
2. **Prompt engineering isn't enough for good performance:** To ensure that the poor performance was not an artifact of the prompts used, we tested multiple hand-crafted prompts that included detailed descriptions of the image characteristics Appendix Figure 5. Despite these tailored prompts, which explicitly described the black-and-white nature and content of the images, the performance still lagged significantly behind simpler models.
3. **Training data contains ample samples with digit concept:** We investigated whether the subpar performance could be attributed to a lack of training images containing digit concepts by analyzing the popular LAION 400M dataset. Our findings reveal a substantial number of captions with both word digits (100k-2M) and integer digits (15M-48M) in the training captions, suggesting that the poor performance is not merely due to insufficient training data (see Figure 9 for exact counts by digit).

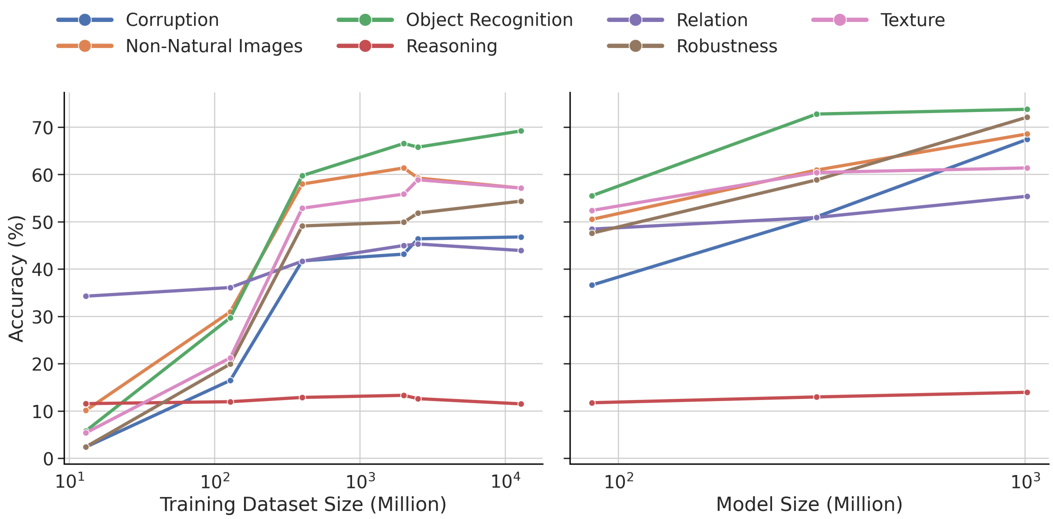


Figure 2: **The effect of scaling model and training dataset size using a fixed architecture and learning paradigm.** Zero-shot performance of models on various benchmark types. We investigate the impact of training dataset size (left), and model size on various benchmark types (right). To isolate the effect of scale, we fix the architecture, learning paradigm, model size (for right panel), and training dataset size (for left) by using the same CLIP ViT-B/32 model and LAION 400M dataset, respectively.We observe a similar trend when measured across all 59 models as shown in Appendix Figure 6

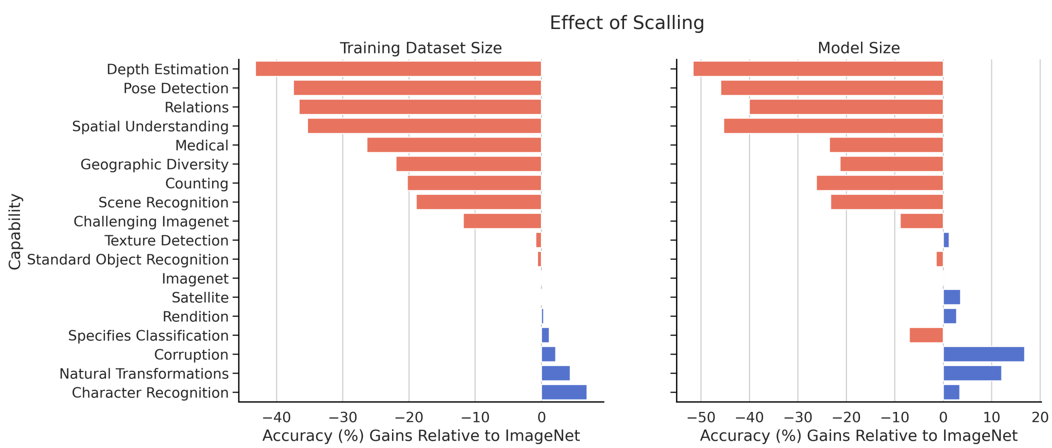


Figure 3: **The effect scaling of training dataset (left) and model size (right) across capabilities for all models.** Accuracy is the difference in performance between the most scaled and the least scaled model across capabilities relative to ImageNet performance.

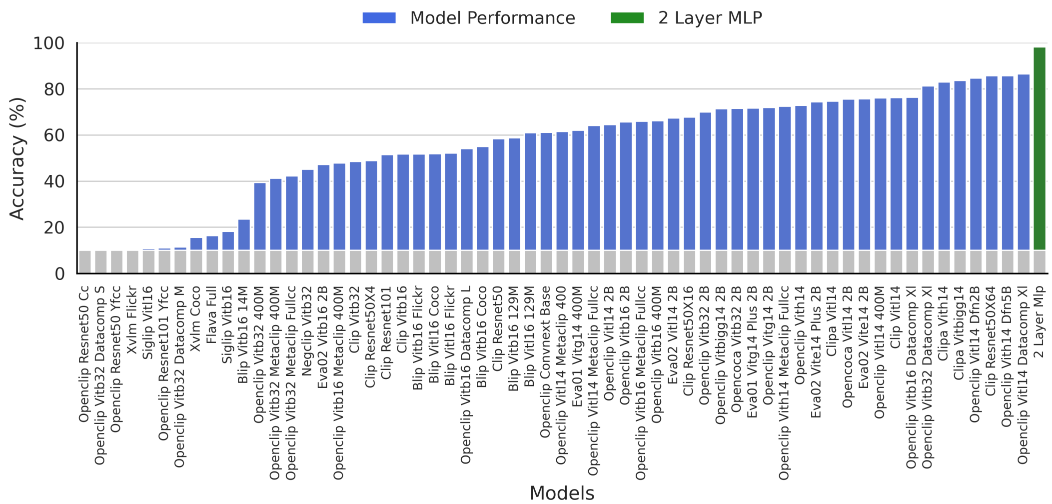


Figure 4: **Performance of 59 VLMs on MNIST, showing despite progress, VLMs still struggle on MNIST**. Blue bars represent zero-shot performance of models, grey bars represent the chance-level for MNIST, and green bar shows performance for a 2-Layer MLP.

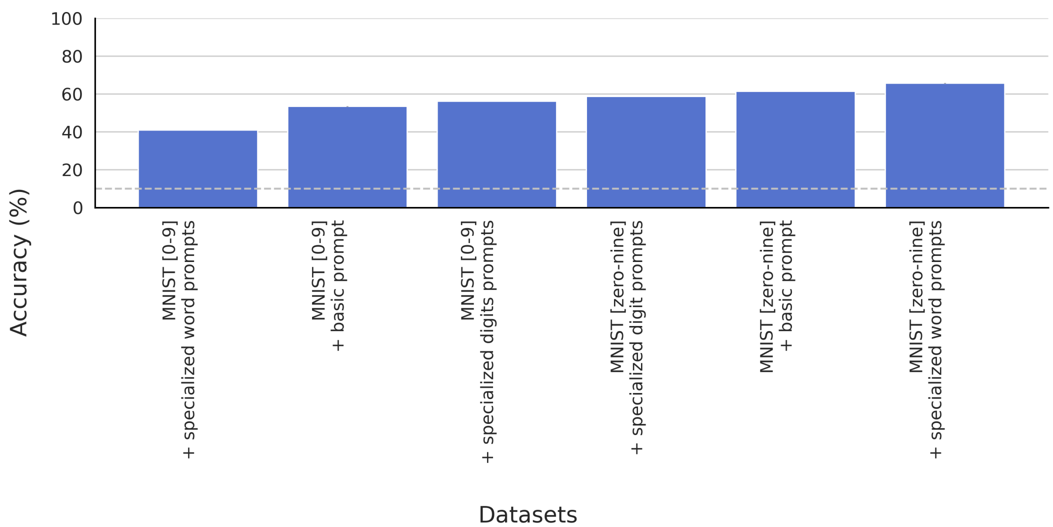


Figure 5: **Median performance of 59 VLMs on MNIST while varying prompts and labels**. Blue bars represent the median zero-shot performance of models and dashed-grey line represents the chance-level for MNIST.

4. **VLMs struggle on other digit benchmarks:** To further explore whether the poor performance on MNIST is indicative of broader issues in number comprehension, we extend our investigation to other benchmarks such as SVHN, CountBench, and ClevrCount (Appendix Figure 10). We find across all benchmarks VLMs struggled with number recognition and counting tasks.

**Takeaway** Despite training on vast datasets, even leading VLMs can struggle with simple tasks solved trivially by much smaller models, including tasks involving basic number comprehension. These findings highlight the need for a comprehensive evaluation pipeline that includes so called, simpler benchmarks, to uncover VLM limitations.

| Benchmark Type | Mean Performance | Top | | Top vs Worst Scale | | Worst | |
|---|---|---|---|---|---|---|---|
| | | Model | Performance | Training Dataset Size | Model Size | Performance | Model |
| Corruption | 46.2 | EVA02 ViT E 14 | 74.3 | 153× | 50× | 2.4 | DataComp ViT B 32 |
| Non-Natural Images | 54.1 | EVA02 ViT E 14 | 74.6 | 153× | 50× | 16.1 | DataComp ViT B 32 |
| Object Recognition | 55.0 | CLIPA ViT G 14 | 71.1 | 98× | 21× | 12.1 | DataComp ViT B 32 |
| Reasoning | 14.9 | OpenCLIP ViT g 14 | 19.0 | 133× | 18× | 10.6 | OpenCLIP ResNet101 |
| Relation | 46.7 | NegCLIP ViT B 32 | 66.8 | 30× | 1× | 33.2 | DataComp ViT B 32 |
| Robustness | 52.1 | EVA02 ViT E 14 | 72.8 | 153× | 50× | 3.8 | DataComp ViT B 32 |
| Texture | 53.5 | MetaCLIP ViT H 14 | 72.5 | 192× | 7× | 5.4 | DataComp ViT B 32 |
| Overall | 46.1 | EVA02 ViT E 14 | 61.2 | 153× | 50× | 12.1 | DataComp ViT B 32 |

Table 1: List of all evaluated dataset types with their corresponding mean performance across models, the best and worst performing models. The Top vs. Worst Scale shows the proportion difference between the worst and best model on the training dataset size and the model size.

### 3.3 What contributes to better model performance?

We have shown both the promise and limitations of scale for VLM performance. We now examine what other levers can overcome the limitations of scale. In particular, we examine other promising factors, such as data quality and learning objectives for improving relational understanding and reasoning.

**Data quality matters more than data quantity.** Among the 59 VLMs we evaluated, there are models trained from 12.8 million samples to 12.8 billion samples. While the quantity of data is often highlighted as a key driver for improving model performance, the quality of the data can be even more critical. For instance, among all models in Appendix Tables 1 and 4, the top performing models are generally the ones trained on 2 billion samples, which use more strict CLIP score filtering as described in Gadre et al. [2024]. This observation suggests that beyond a certain threshold, simply adding more data does not necessarily translate to better performance. Instead, the composition and quality of the data set become paramount. Models trained on such data are better equipped to generalize from their training environments to real-world applications, demonstrating that strategic curation of data can be more valuable than the sheer volume of data collected.

**Tailored learning objectives can help where scale does not.** The learning objectives defined during model training phase are pivotal in steering the model's learning process and ultimately its performance on various tasks. A notable example is NegCLIP [Yuksekgonul et al., 2022], with a tailored learning objective for capturing relations via hard-negatives seems to substantially aid NegCLIP's performance on relational understanding (Appendix Tables 1 and 4). As shown in the original paper NegCLIP's performance isn't simply the result of finetuning with additional data (see Table 6 of Yuksekgonul et al. [2022]), but is thanks to a tailored learning objective involving hard negatives. NegCLIP, with only 86M parameters, significantly outperforms models up to 50× larger with an overall performance of 70.4%, compared to only 50.5% for the largest EVA ViT-E/14 model with 4.3B parameters. Similarly, Paiss et al. [2023] tailored learning objective for VLMs can have significantly improve performance on counting tasks.

### 3.4 Which model should I use?

Finally, we provide recommendations for practitioners to select the most suitable openly available VLM. For an overall high-performing model across the axes we measured, models with large ViT encoders trained on large datasets exhibit the highest overall performance, with Eva-2 ViT-E/14 leading the way. For relations, counting, or related capabilities, we rank the top and worst performing models in Appendix Table 4.

## References


Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*, volume 32.

Curran Associates, Inc., 2019. URL `https://papers.nips.cc/paper_files/paper/2019/hash/97af07a14cacba681feacf3012730892-Abstract.html`.

Florian Bordes, Shashank Shekhar, Mark Ibrahim, Diane Bouchacourt, Pascal Vincent, and Ari S. Morcos. Pug: Photorealistic and semantically controllable synthetic data for representation learning, 2023.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning, December 2022. URL `http://arxiv.org/abs/2212.07143`. arXiv:2212.07143 [cs].

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

Yufeng Cui, Lichen Zhao, Feng Liang, Yangguang Li, and Jing Shao. Democratizing contrastive language-image pre-training: A clip benchmark of data, model, and supervision, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848. ISSN: 1063-6919.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks, 2023a.

Yuxin Fang, Wen Wang, Binhui Xie, Quan-Sen Sun, Ledell Yu Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. 2023 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19358–19369, 2022.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *arXiv preprint arXiv:2303.11331*, 2023b.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004.

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

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024.

William Gaviria Rojas, Sudnya Diamos, Keertan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. *Advances in Neural Information Processing Systems*, 35:12979–12990, 2022.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012.

Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A. Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Introducing eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 204–207. IEEE, 2018.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations, 2019.

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, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15262–15271, 2021b.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in Neural Information Processing Systems*, 36, 2024.

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.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yann LeCun, Fu Jie Huang, and Leon Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–104. IEEE, 2004.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

Xianhang Li, Zeyu Wang, and Cihang Xie. An inverse scaling law for clip training, 2023a.

Xiaodan Li, Yuefeng Chen, Yao Zhu, Shuhui Wang, Rong Zhang, and Hui Xue. ImageNet-E: Benchmarking Neural Network Robustness via Attribute Editing, March 2023b. URL `http://arxiv.org/abs/2303.17096`. arXiv:2303.17096 [cs].

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 7. Granada, Spain, 2011.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. *arXiv preprint arXiv:2302.12066*, 2023.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.

Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. Filtering, distillation, and hard negatives for vision-language pre-training, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, February 2021a. URL `http://arxiv.org/abs/2103.00020`. arXiv:2103.00020 [cs].

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

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet?, 2019.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model, 2022.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 32:323–332, 2012.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, pages 210–218. Springer, 2018.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. *The Caltech-UCSD Birds-200-2011 Dataset*. Jul 2011.

Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. Regularization of neural networks using dropconnect. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1058–1066, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL `https://proceedings.mlr.press/v28/wan13.html`.

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

Zhiguang Wang and Jianbo Yang. Diabetic retinopathy detection via deep convolutional networks for discriminative localization and visual explanation. In *Workshops at the thirty-second AAAI conference on artificial intelligence*, 2018.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *ArXiv preprint arXiv:2006.09994*, 2020.

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

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models, 2022.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it?, October 2022. URL `https://arxiv.org/abs/2210.01936v2`.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it?, March 2023. URL `http://arxiv.org/abs/2210.01936`. arXiv:2210.01936 [cs].

Tian Yun, Usha Bhalla, Ellie Pavlick, and Chen Sun. Do vision-language pretrained models learn composable primitive concepts?, 2023.

Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts, 2022.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019.

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

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

# A Appendix

## A.1 UniBench Implementation Details

We have developed UniBench to be easy-to-run library to allow researchers to systematically compare and contrast exsisting (n=59 ) and new VLMs on 53 benchmarks. To evaluate new VLMs that expand beyond the already implemented 59 VLMs, users need to follow Code Snippet 1. Users would need to create a class that inherent from ClipModel from uni_bench.models_zoo with get_image_embeddings and get_text_embeddings methods implemented. get_image_embeddings and get_text_embeddings methods takes images and captions as input, respectively, and returns a tensor of encoded representations.

```
from uni_bench.models_zoo import ClipModel
import uni_bench

class CustomModel(ClipModel):

    @torch.no_grad()
    # Output tensor of final layer of image encoder
    def get_image_embeddings(self, images):
        ...

    @torch.no_grad()
    # Output tensor of final layer of text encoder given captions
    def get_text_embeddings(self, captions):
        ...


evaluator = uni_bench.Evaluator() # Initialize Evaluator class
new_model = CustomModel() # Initialize new model
evaluator.add_model(new_model) # add new model to the evaluation pipeline
evaluator.evaluate() # run the evaluation
```

Code Snippet 1: Custom Model Example

## A.2 Gauging progress in Vision Language Models

**Scaling improves many benchmarks, but offers little benefit for reasoning and relation.** Appendix Figure 6 shows that despite increasing the training dataset size by a factor of $1000\times$ and model size by a factor of $11\times$, relational and reasoning benchmarks performance is fairly flat compared to the significant boost in performance on other tasks. We further pinpoint capabilities such as Depth Estimation, Spatial Understanding, Counting, Scene and Text Recognition, as the underlying capabilities where scale does not lead to improvements as shown in Figure 7.

## A.3 Impact of Prompts on MNIST Performance

The MNIST dataset, featuring handwritten digits, was subjected to various prompting strategies to evaluate their impact on model performance. Our findings reveal a distinct hierarchy in performance based on the type of prompts used. The benchmark was tested with both numeral formats ("zero-nine" and "0-9") and different prompt styles (specialized word prompts, specialized digit prompts, and a basic prompt) (Figure 5).

### A.3.1 Hierarchy of Prompt Performance

The performance of the MNIST model varied significantly across different prompt types and formats, arranged here from best to worst performing setups: 1. Word digits ("zero-nine") with specialized word prompts 2. Word digits ("zero-nine") with basic prompt 3. Word digits ("zero-nine") with specialized digit prompts 4. Digits ("0-9") with specialized digit prompts 5. Digits ("0-9") with basic prompt 6. Digits ("0-9") with specialized word prompts

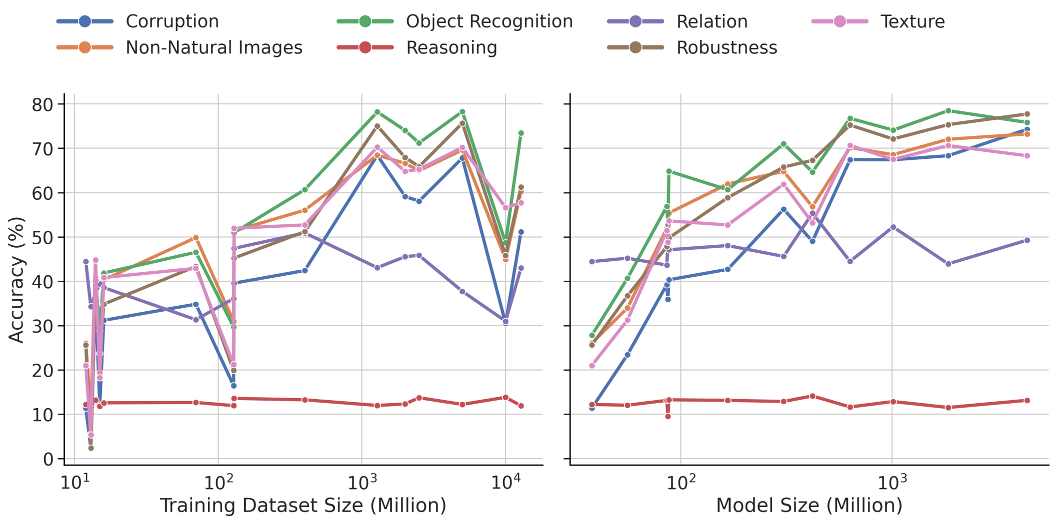


Figure 6: **The effect of scaling model and training dataset size on all models.** Median zero-shot performance of models on various benchmark types. We investigate the impact of training dataset size (left), and model size on various benchmark types (right).

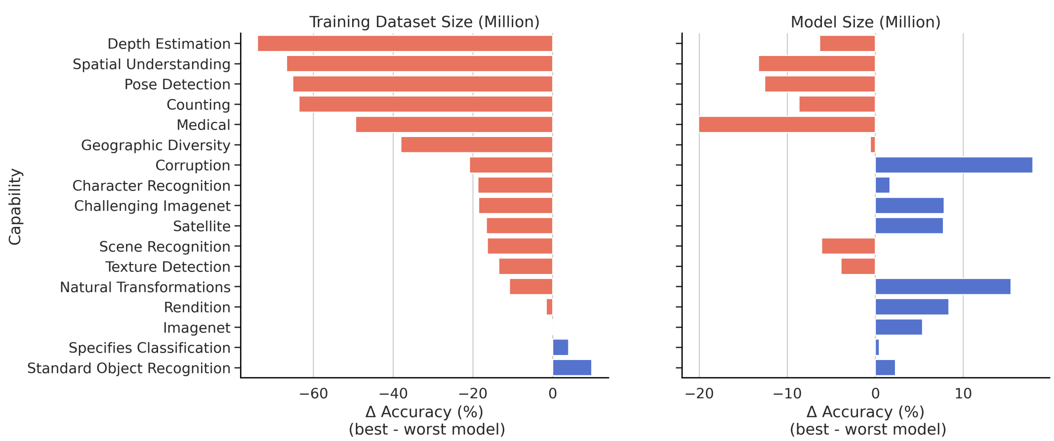


Figure 7: **Benchmark capabilities performance does not scale with dataset and model size** Median zero-shot performance of models on various benchmark capabilities. We investigate the impact of dataset size (left), and model size on various benchmark capabilities (right). We isolate the effect of training data size keeping other factors such as architecture, learning objective, and model size fixed only using ViT B32 (left). For right panel subfigure, we isolate the effect of model size keeping other factors such as architecture, learning objective, and training data size fixed only using LIAON 400M (right).

#### A.3.2 Specialized Word Prompts

These prompts provided detailed descriptions and contexts, significantly enhancing the model's ability to recognize and interpret the digits accurately. Examples include:

- "showcasing the digit {}, is this image."
- "this number {} is represented in a handwritten form."
- "the numeral {} is captured in this snapshot."
- "the digit {} is depicted visually in this image."
- "this image is a graphical representation of the number {}."
- "this is an illustration of the digit {}."
- "this image represents the digit {} in a handwritten form."
- "the number {} is sketched as a digit in this image."
- "this is a photograph of the digit {}."
- "the number {} is drawn as a digit in this image."

#### A.3.3 Specialized Digit Prompts

These prompts explicitly mention the format or style of the digit, aiding in recognition but to a lesser extent compared to specialized word prompts. Examples include:

- "A photo of the number: '{}'."
- "A digit drawing of the number: '{}'."
- "A digit sketch of the number: '{}'."
- "A handwritten digit image of: '{}'."
- "A digit illustration of: '{}'."
- "A graphical representation of the number: '{}'."
- "A visual depiction of the digit: '{}'."
- "A snapshot of the numeral: '{}'."
- "A handwritten representation of the number: '{}'."
- "An image showcasing the digit: '{}'."

#### A.3.4 Basic Prompt

The basic prompt used:

- "a photo of the number: '{}'."

This structured analysis clearly demonstrates how the specificity and relevance of the prompt significantly influence the performance of VLMs. We investigated whether the subpar performance could be attributed to a lack of training images containing digit concepts by analyzing the popular LAION 400M dataset. Our findings reveal a substantial number of captions with both word digits (100k-2M) and integer digits (15M-48M) in the training captions, suggesting that the poor performance is not merely due to insufficient training data (see Figure 9 for exact counts by digit). To further understand the performance results on MNIST, we compute more generous top-2,-3,-4, and -5 accuracy measures to understand whether models confuse similar digits. We show in Appendix Figure 8 that even when we compute top-5 accuracy (with 50% being chance), VLMs barely reach 90% accuracy suggesting poor performance is not due to minor confusions among digits.

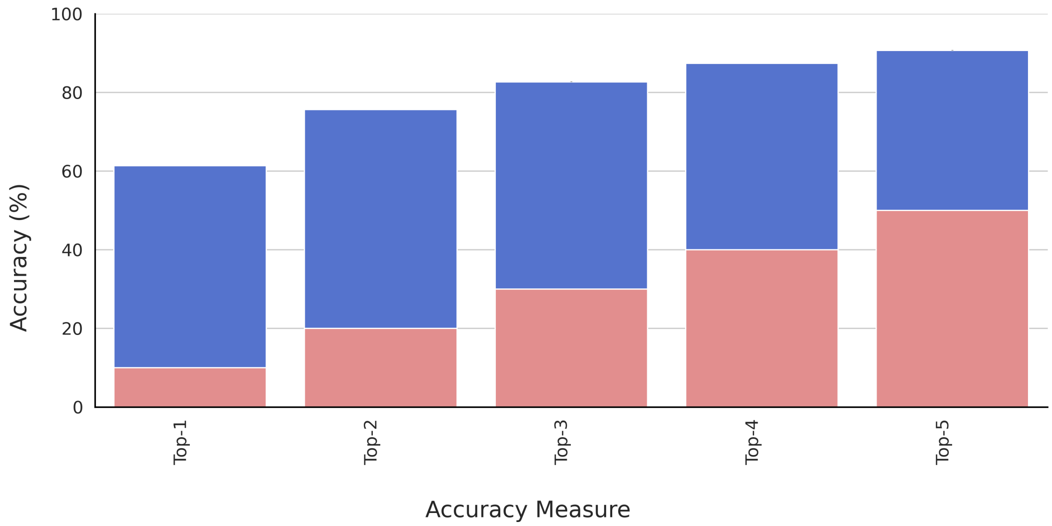


Figure 8: **Median performance of 59 VLMs on MNIST while varying accuracy measure from top-1 to top-5**. The following further shows that VLMs' performance on MNIST is not due mismatch between top-1 and top-5 guesses. Blue bars represent the median zero-shot performance of models and red bars represents the chance-level for benchmarks.

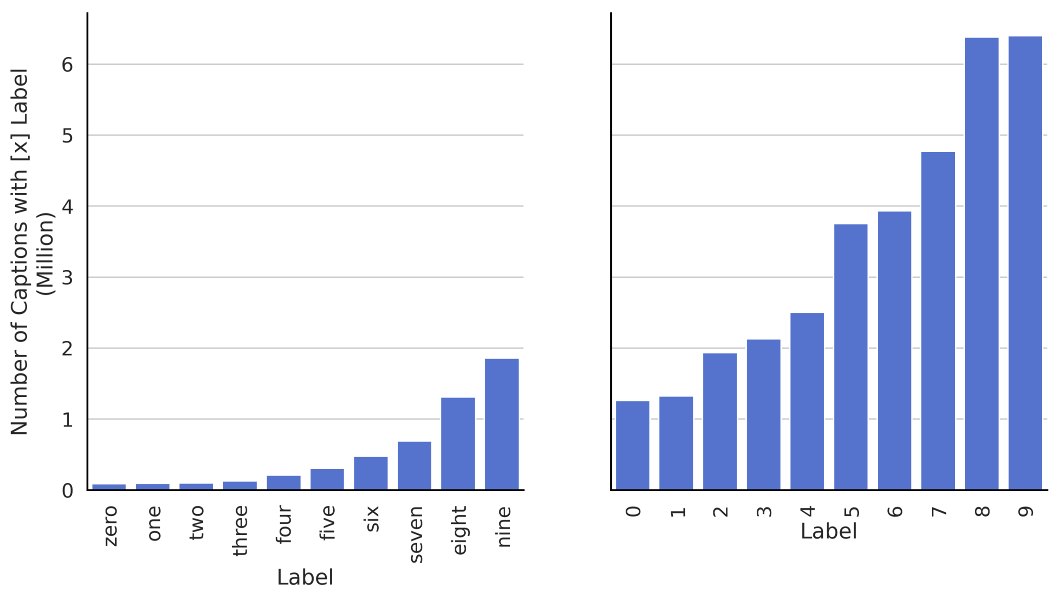


Figure 9: **Frequency of different digits in LAION-400M, showing substantial frequency of digits in visual diet of VLMs.** Left panel counts the number of words of the digits i.e. [zero-nine] and right panel counts the number of digits in LAION-400M.

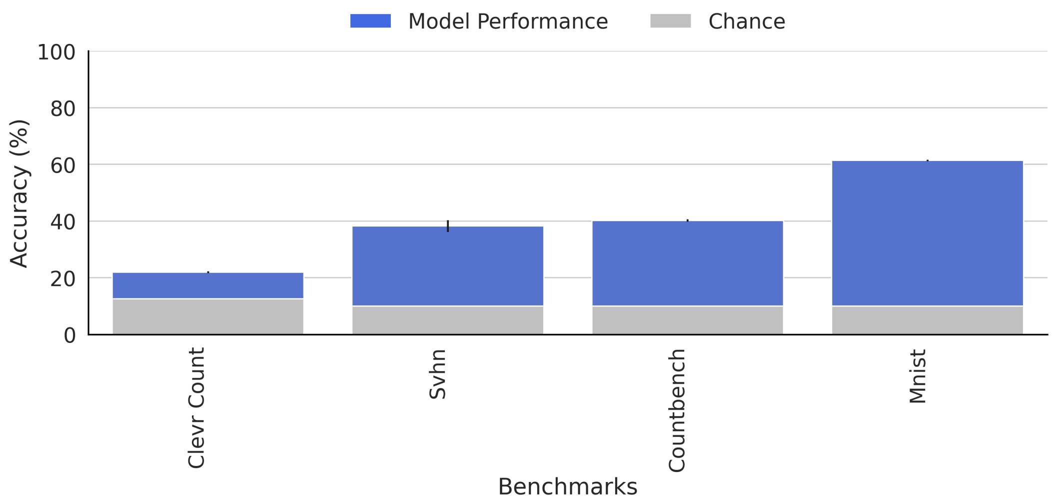


Figure 10: **Median performance of 59 VLMs on counting and character recognition benchmarks, showing MNIST performance is not an isolated instance and VLMs generally sruggle with these tasks**. Blue bars represent the median zero-shot performance of models and gray bars represent random chance-level.

### A.4 Correlation of ImageNet with Other Benchmarks

ImageNet, often considered a cornerstone in the field of computer vision, has been widely used as a benchmark to evaluate the performance of image recognition models. Its extensive dataset and challenging classification tasks have set a standard for algorithm development and comparison. However, while ImageNet correlates well with many benchmarks, it does not exhibit a universal correlation across all tasks. Our analysis reveals that for a significant number of benchmarks, specifically 18 out of the 53 benchmarks analyzed, the performance on ImageNet is poorly or negatively correlated. This is illustrated in Appendix Figure 11, which provides a detailed comparison of benchmark performances. This finding suggests that success on ImageNet does not necessarily translate to proficiency in all visual tasks.

### A.5 A Practical Subset of Benchmarks

While ideally, evaluating VLMs across all 53 benchmarks would provide the most comprehensive insights, the computational demands and complexity of parsing such extensive data can be overwhelming (6 million images to evaluate; 2+ hours for one model on an A100 GPU). To streamline evaluation, we distill the full set of benchmarks in `UniBench` into seven benchmark types and 17 capabilities. These categorizations were deraved based on benchmarks that correlate strongly with other benchmarks for each benchmark type and capability (Tables 2 and 3).

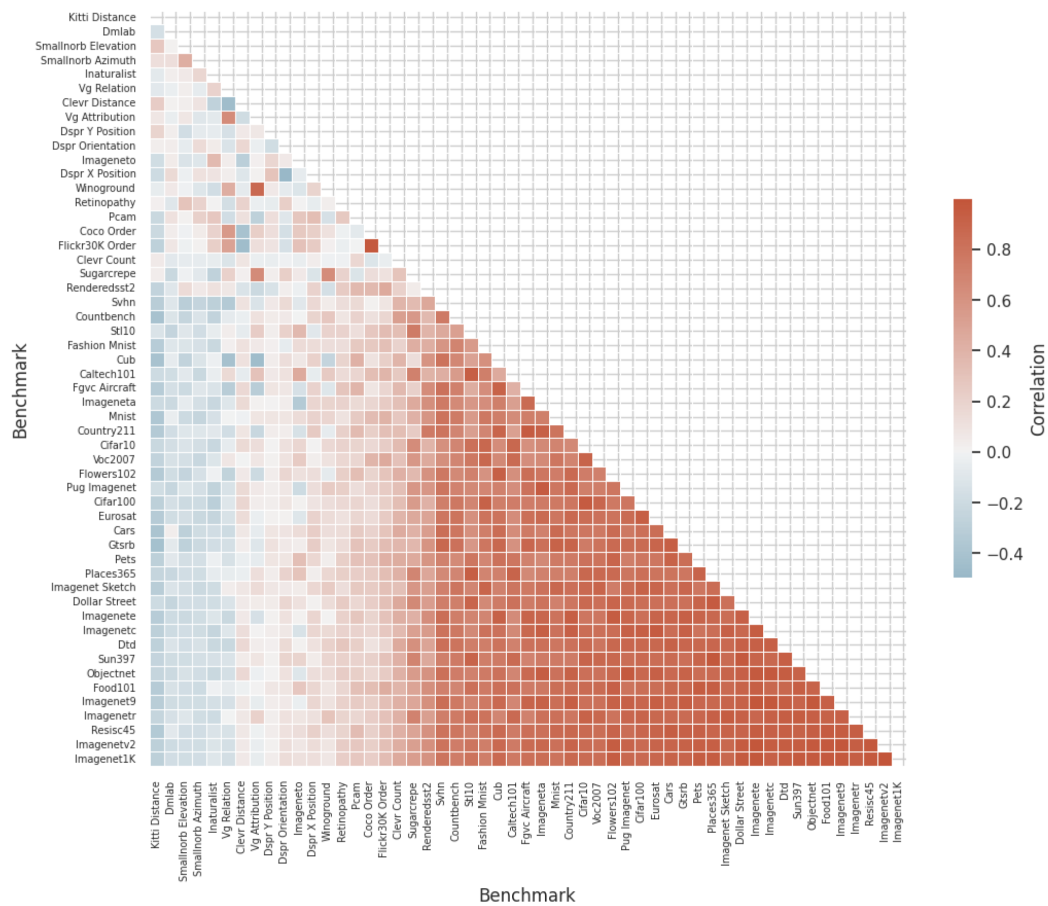


Figure 11: **Correlation matrix of models' performance across all benchmarks.**

| Benchmark Type | Most Correlated Benchmark | Correlation Value |
|---|---|---|
| Object recognition | ImageNet-1k | 0.82 |
| Reasoning (Counting) | CountBench | 0.76 |
| Reasoning (Spatial) | DSPR Position | 0.29 |
| Relation | VG Attribution | 0.57 |
| Texture | DTD | 1 |
| Non-Natural Images | Resisc45 | 0.72 |
| Robustness | ImageNet-v2 | 0.81 |
| Corruption | ImageNet-c | 1 |

Table 2: **Evaluate on a curated list of benchmark types, rather than the full set, to save time.** The list includes benchmarks that correlate strongly with other benchmarks for each benchmark type.

| Capabilities | Most Correlated Benchmark | Correlation Value |
|---|---|---|
| standard object recognition | food101 | 0.85 |
| counting | countbench | 0.76 |
| spatial understanding | dspr y position | 0.29 |
| relations | vg attribution | 0.57 |
| geographic diversity | dollar street | 0.89 |
| specifies classification | flowers102 | 0.7 |
| depth estimation | dmlab | 0.42 |
| pose detection | smallnorb azimuth | 0.57 |
| texture detection | dtd | 1 |
| satellite | eurosat | 0.95 |
| character recognition | mnist | 0.88 |
| imagenet | imagenet1k | 1 |
| natural transformations | imagenet9 | 0.99 |
| rendition | imagenetr | 0.97 |
| challenging imagenet | imagenetv2 | 0.65 |
| corruption | imagenetc | 1 |
| medical | retinopathy | 0.64 |
| scene recognition | sun397 | 0.99 |

Table 3: **Evaluate on a curated list of capabilities, rather than the full set, to save time.** The list includes benchmarks that correlate strongly with other benchmarks for each capability.

| Benchmark Type | Mean Performance | **Top** | | **Top vs Worst Scale** | | **Worst** | |
|---|---|---|---|---|---|---|---|
| | | Model | Performance | Training Dataset Size | Model Size | Performance | Model |
| Challenging Imagenet | 47.8 | EVA02 ViT E 14 | 64.4 | 153 | 50 | 5.0 | DataComp ViT B 32 |
| Character Recognition | 54.8 | CLIPA ViT G 14 | 74.3 | 85 | 48 | 20.5 | OpenCLIP ResNet50 |
| Corruption | 46.1 | EVA02 ViT E 14 | 74.3 | 153 | 50 | 2.3 | DataComp ViT B 32 |
| Counting | 31.4 | OpenCOCA ViT L 14 | 53.1 | 153 | 3 | 11.5 | DataComp ViT B 32 |
| Depth Estimation | 20.4 | DataComp ViT B 16 | 27.6 | 0.6 | 0.1 | 12.4 | OpenCLIP ViT H 14 |
| Geographic Diversity | 33.8 | CLIPA ViT G 14 | 46.8 | 98 | 21 | 5.3 | DataComp ViT B 32 |
| Imagenet | 65.7 | OpenCLIP ViT H 14 | 83.1 | 384 | 7 | 3.9 | DataComp ViT B 32 |
| Medical | 43.3 | MetaCLIP ViT L 14 | 68.6 | 0.3 | 3 | 26.8 | DataComp ViT B 16 |
| Natural Transformations | 56.2 | CLIPA ViT G 14 | 81.7 | 98 | 21 | 2.5 | DataComp ViT B 32 |
| Pose Detection | 3.9 | OpenCLIP ViT B 32 | 4.7 | 5 | 0.9 | 3.3 | OpenCLIP ConvNext |
| Relations | 46.7 | NegCLIP ViT B 32 | 66.7 | 30 | 1 | 33.2 | DataComp ViT B 32 |
| Rendition | 63.7 | CLIPA ViT G 14 | 84.2 | 98 | 21 | 3.8 | DataComp ViT B 32 |
| Satellite | 55.2 | EVA02 ViT E 14 | 75.7 | 153 | 50 | 12.3 | DataComp ViT B 32 |
| Scene Recognition | 53.0 | OpenCLIP ViT H 14 | 61.7 | 384 | 7 | 6.3 | DataComp ViT B 32 |
| Spatial Understanding | 9.1 | MetaCLIP ViT L 14 | 11.3 | 1 | 3 | 6.3 | CLIP ResNet50x4 |
| Specifies Classification | 51.7 | OpenCLIP ViT H 14 | 68.9 | 384 | 7 | 2.8 | DataComp ViT B 32 |
| Standard Object Recognition | 60.0 | CLIPA ViT G 14 | 77.1 | 98 | 21 | 13.8 | DataComp ViT B 32 |
| Texture Detection | 53.4 | MetaCLIP ViT H 14 | 72.4 | 192 | 7 | 5.3 | DataComp ViT B 32 |
| Overall | 44.2 | EVA02 ViT E 14 | 58.0 | 153 | 50 | 11.3 | DataComp ViT B 32 |

Table 4: List of all evaluated capabilities with their corresponding mean performance across models, the best and the worst performing models. The Top vs. Worst Scale shows the proportion difference between the worst and best model on the training dataset size and the model size.

| Model | Dataset size | Model size | Learning objective | Architecture | Model name |
|---|---|---|---|---|---|
| blip_vitB16_14m Li et al. [2022] | 14 | 86 | BLIP | vit | BLIP ViT B 16 |
| blip_vitL16_129m Li et al. [2022] | 129 | 307 | BLIP | vit | BLIP ViT L 16 |
| blip_vitB16_129m Li et al. [2022] | 129 | 86 | BLIP | vit | BLIP ViT B 16 |
| blip_vitB16_coco Li et al. [2022] | 129 | 86 | BLIP | vit | BLIP ViT B 16 |
| blip_vitB16_flickr Li et al. [2022] | 129 | 86 | BLIP | vit | BLIP ViT B 16 |
| blip_vitL16_coco Li et al. [2022] | 129 | 307 | BLIP | vit | BLIP ViT L 16 |
| blip_vitL16_flickr Li et al. [2022] | 129 | 307 | BLIP | vit | BLIP ViT L 16 |
| eva02_vitE14_plus_2b Fang et al. [2023b] | 2000 | 4350 | Pure Contrastive | vit | EVA02 ViT E 14 |
| eva02_vitE14_2b Fang et al. [2023b] | 2000 | 4350 | Pure Contrastive | vit | EVA02 ViT E 14 |
| eva02_vitL14_2b Fang et al. [2023b] | 2000 | 307 | Pure Contrastive | vit | EVA02 ViT L 14 |
| eva02_vitB16_2b Fang et al. [2023b] | 2000 | 86 | Pure Contrastive | vit | EVA02 ViT B 16 |
| eva01_vitG14_plus_2b Fang et al. [2022] | 2000 | 1011 | Pure Contrastive | vit | EVA01 ViT g 14 |
| eva01_vitG14_400m Fang et al. [2022] | 400 | 1011 | Pure Contrastive | vit | EVA01 ViT g 14 |
| clipa_vitbigG14 Li et al. [2023a] | 1280 | 1843 | Pure Contrastive | vit | CLIPA ViT G 14 |
| clipa_vitH14 Li et al. [2023a] | 1280 | 633 | Pure Contrastive | vit | CLIPA ViT H 14 |
| clipa_vitL14 Li et al. [2023a] | 1280 | 307 | Pure Contrastive | vit | CLIPA ViT L 14 |
| siglip_vitL16 Zhai et al. [2023] | 10000 | 307 | Contrastive (sigmoid) | vit | SigLIP ViT L 16 |
| siglip_vitB16 Zhai et al. [2023] | 10000 | 86 | Contrastive (sigmoid) | vit | SigLIP ViT B 16 |
| openclip_vitB32_metaclip_fullcc Xu et al. [2023] | 2500 | 86 | Pure Contrastive | vit | MetaCLIP ViT B 32 |
| openclip_vitB16_metaclip_400m Xu et al. [2023] | 400 | 86 | Pure Contrastive | vit | MetaCLIP ViT B 16 |
| openclip_vitB32_metaclip_400m Xu et al. [2023] | 400 | 86 | Pure Contrastive | vit | MetaCLIP ViT B 32 |
| openclip_vitB16_metaclip_fullcc Xu et al. [2023] | 2500 | 86 | Pure Contrastive | vit | MetaCLIP ViT B 16 |
| openclip_vitL14_dfn2b Fang et al. [2023a] | 2000 | 307 | Pure Contrastive | vit | OpenCLIP ViT L 14 |
| openclip_vitL14_metaclip_400 Xu et al. [2023] | 400 | 307 | Pure Contrastive | vit | MetaCLIP ViT L 14 |
| openclip_vitL14_metaclip_fullcc Xu et al. [2023] | 2500 | 307 | Pure Contrastive | vit | MetaCLIP ViT L 14 |
| openclip_vitH14_metaclip_fullcc Xu et al. [2023] | 2500 | 633 | Pure Contrastive | vit | MetaCLIP ViT H 14 |
| openclip_vitH14_dfn5b Fang et al. [2023a] | 5000 | 633 | Pure Contrastive | vit | OpenCLIP ViT H 14 |
| openclip_convnext_base Ilharco et al. [2021] | 400 | 88 | Pure Contrastive | conv | OpenCLIP ConvNext |
| openclip_vitB32_datacomp_s Gadre et al. [2023] | 13 | 86 | Pure Contrastive | vit | DataComp ViT B 32 |
| openclip_vitB32_datacomp_m Gadre et al. [2023] | 128 | 86 | Pure Contrastive | vit | DataComp ViT B 32 |
| openclip_vitB32_datacomp_xl Gadre et al. [2023] | 12800 | 86 | Pure Contrastive | vit | DataComp ViT B 32 |
| openclip_vitB16_datacomp_xl Gadre et al. [2023] | 12800 | 86 | Pure Contrastive | vit | DataComp ViT B 16 |
| openclip_vitB16_datacomp_l Gadre et al. [2023] | 1280 | 86 | Pure Contrastive | vit | DataComp ViT B 16 |
| openclip_vitH14 Ilharco et al. [2021] | 2000 | 633 | Pure Contrastive | vit | OpenCLIP ViT H 14 |
| xvlm_flickr Zeng et al. [2022] | 16 | 86 | XVLM | Swin | XVLM Swin B |
| flava_full Singh et al. [2022] | 70 | 86 | Other | vit | FLAVA ViT B 32 |
| openclip_vitL14_400m Ilharco et al. [2021] | 400 | 307 | Pure Contrastive | vit | OpenCLIP ViT L 14 |
| openclip_vitL14_datacomp_xl Gadre et al. [2023] | 12800 | 307 | Pure Contrastive | vit | DataComp ViT L 14 |
| openclip_vitL14_2b Ilharco et al. [2021] | 2000 | 307 | Pure Contrastive | vit | OpenCLIP ViT L 14 |
| clip_vitL14 Radford et al. [2021b] | 400 | 307 | Pure Contrastive | vit | CLIP ViT L 14 |
| xvlm_coco Zeng et al. [2022] | 16 | 86 | XVLM | Swin | XVLM Swin B |
| openclip_vitB32_400m Ilharco et al. [2021] | 400 | 86 | Pure Contrastive | vit | OpenCLIP ViT B 32 |
| openclip_vitB32_2b Ilharco et al. [2021] | 2000 | 86 | Pure Contrastive | vit | OpenCLIP ViT B 32 |
| openclip_vitG14_2b Ilharco et al. [2021] | 2000 | 1011 | Pure Contrastive | vit | OpenCLIP ViT g 14 |
| openclip_vitbigG14_2b Ilharco et al. [2021] | 2000 | 1843 | Pure Contrastive | vit | OpenCLIP ViT G 14 |
| openclip_vitB16_2b Ilharco et al. [2021] | 2000 | 86 | Pure Contrastive | vit | OpenCLIP ViT B 16 |
| openclip_vitB16_400m Ilharco et al. [2021] | 400 | 86 | Pure Contrastive | vit | OpenCLIP ViT B 16 |
| opencoca_vitL14_2b Yu et al. [2022], Ilharco et al. [2021] | 2000 | 307 | Other | vit | OpenCOCA ViT L 14 |
| opencoca_vitB32_2b Yu et al. [2022], Ilharco et al. [2021] | 2000 | 86 | Other | vit | OpenCOCA ViT B 32 |
| negclip_vitB32 Yuksekgonul et al. [2023] | 400 | 86 | Negative CLIP | vit | NegCLIP ViT B 32 |
| clip_vitB16 Radford et al. [2021b] | 400 | 86 | Pure Contrastive | vit | CLIP ViT B 16 |
| clip_resnet50 Radford et al. [2021b] | 400 | 38 | Pure Contrastive | conv | CLIP ResNet50 |
| openclip_resnet101_yfcc Ilharco et al. [2021] | 15 | 56 | Pure Contrastive | conv | OpenCLIP ResNet101 |
| openclip_resnet50_yfcc Ilharco et al. [2021] | 15 | 38 | Pure Contrastive | conv | OpenCLIP ResNet50 |
| openclip_resnet50_cc Ilharco et al. [2021] | 12 | 38 | Pure Contrastive | conv | OpenCLIP ResNet50 |
| clip_resnet101 Radford et al. [2021b] | 400 | 56 | Pure Contrastive | conv | CLIP ResNet101 |
| clip_resnet50x4 Radford et al. [2021b] | 400 | 87 | Pure Contrastive | conv | CLIP ResNet50x4 |
| clip_resnet50x16 Radford et al. [2021b] | 400 | 167 | Pure Contrastive | conv | CLIP ResNet50x16 |
| clip_resnet50x64 Radford et al. [2021b] | 400 | 420 | Pure Contrastive | conv | CLIP ResNet50x64 |
| clip_vitB32 Radford et al. [2021b] | 400 | 86 | Pure Contrastive | vit | CLIP ViT B 32 |

Table 5: List of all the models used in evaluations with their corresponding dataset size, model size (number of parameters), learning objective, and architecture.

| Benchmark | Measure | Benchmark Type | Capability | Curated | Object Centric | Number of Classes |
|---|---|---|---|---|---|---|
| caltech101 [Fei-Fei et al., 2004] | zero-shot | object recognition | standard object recognition | False | True | 102 |
| cars [Krause et al., 2013] | zero-shot | object recognition | standard object recognition | False | True | 196 |
| cifar10 [Krizhevsky et al., 2009] | zero-shot | object recognition | standard object recognition | False | True | 10 |
| cifar100 [Krizhevsky et al., 2009] | zero-shot | object recognition | standard object recognition | False | True | 100 |
| clevr count [Johnson et al., 2017] | zero-shot | reasoning | counting | True | False | 8 |
| clevr distance [Johnson et al., 2017] | zero-shot | reasoning | spatial understanding | True | False | 6 |
| coco order [Yuksekgonul et al., 2023] | relation | relation | relations | False | False | 5 |
| countbench [Paiss et al., 2023] | zero-shot | reasoning | counting | False | False | 10 |
| country211 [Radford et al., 2021a] | zero-shot | object recognition | geographic diversity | False | False | 211 |
| cub [Wah et al., 2011] | zero-shot | object recognition | specifies classification | False | False | 200 |
| dmlab [Zhai et al., 2019] | zero-shot | reasoning | depth estimation | True | False | 6 |
| dollar street [Gaviria Rojas et al., 2022] | zero-shot | object recognition | geographic diversity | False | True | 60 |
| dspr orientation [Matthey et al., 2017] | zero-shot | reasoning | pose detection | True | False | 40 |
| dspr x position [Matthey et al., 2017] | zero-shot | reasoning | spatial understanding | True | False | 32 |
| dspr y position [Matthey et al., 2017] | zero-shot | reasoning | spatial understanding | True | False | 32 |
| dtd [Cimpoi et al., 2014] | zero-shot | texture | texture detection | True | False | 47 |
| eurosat [Helber et al., 2019, 2018] | zero-shot | non-natural images | satellite | False | False | 10 |
| fashion mnist [Xiao et al., 2017] | zero-shot | object recognition | character recognition | True | True | 10 |
| fgvc aircraft [Maji et al., 2013] | zero-shot | object recognition | standard object recognition | False | True | 100 |
| flickr30k order [Yuksekgonul et al., 2023] | relation | relation | relations | False | False | 5 |
| flowers102 [Nilsback and Zisserman, 2008] | zero-shot | object recognition | specifies classification | False | True | 102 |
| food101 [Bossard et al., 2014] | zero-shot | object recognition | standard object recognition | False | True | 101 |
| gtsrb [Stallkamp et al., 2012] | zero-shot | object recognition | standard object recognition | False | True | 43 |
| imagenet1k [Deng et al., 2009] | zero-shot | object recognition | imagenet | False | True | 1000 |
| imagenet9[Xiao et al., 2020] | zero-shot | robustness | natural transformations | True | True | 1000 |
| imagenet sketch [Wang et al., 2019] | zero-shot | non-natural images | rendition | True | True | 1000 |
| imageneta [Hendrycks et al., 2021b] | zero-shot | robustness | challenging imagenet | True | True | 200 |
| imagenetc [Hendrycks and Dietterich, 2019] | zero-shot | corruption | corruption | True | True | 1000 |
| imagenete [Li et al., 2023b] | zero-shot | robustness | natural transformations | True | True | 1000 |
| imageneto [Hendrycks et al., 2021b] | zero-shot | robustness | challenging imagenet | True | True | 200 |
| imagenetr [Hendrycks et al., 2021a] | zero-shot | non-natural images | rendition | True | True | 200 |
| imagenetv2 [Recht et al., 2019] | zero-shot | robustness | challenging imagenet | True | True | 1000 |
| inaturalist [Van Horn et al., 2018] | zero-shot | object recognition | specifies classification | False | True | 5089 |
| kitti distance [Geiger et al., 2012] | zero-shot | reasoning | depth estimation | False | False | 4 |
| mnist[LeCun et al., 1998] | zero-shot | object recognition | character recognition | True | True | 10 |
| objectnet [Barbu et al., 2019] | zero-shot | robustness | natural transformations | False | True | 113 |
| pcam [Veeling et al., 2018] | zero-shot | non-natural images | medical | True | False | 2 |
| pets [Parkhi et al., 2012] | zero-shot | object recognition | specifies classification | False | True | 37 |
| places365 [Zhou et al., 2017] | zero-shot | object recognition | scene recognition | False | False | 365 |
| pug imagenet [Bordes et al., 2023] | zero-shot | object recognition | standard object recognition | False | True | 151 |
| renderedsst2 [Radford et al., 2021a] | zero-shot | object recognition | character recognition | True | True | 2 |
| resisc45[Cheng et al., 2017] | zero-shot | non-natural images | satellite | False | False | 45 |
| retinopathy [Wang and Yang, 2018] | zero-shot | non-natural images | medical | False | False | 5 |
| smallnorb azimuth [LeCun et al., 2004] | zero-shot | reasoning | pose detection | True | False | 18 |
| smallnorb elevation [LeCun et al., 2004] | zero-shot | reasoning | spatial understanding | True | False | 9 |
| stl10 [Coates et al., 2011] | zero-shot | object recognition | standard object recognition | False | True | 10 |
| sugarcrepe [Hsieh et al., 2024] | relation | relation | relations | False | False | 2 |
| sun397 [Xiao et al., 2010] | zero-shot | object recognition | scene recognition | False | False | 397 |
| svhn [Netzer et al., 2011] | zero-shot | object recognition | character recognition | False | True | 10 |
| vg attribution [Yuksekgonul et al., 2023] | relation | relation | relations | False | False | 2 |
| vg relation [Yuksekgonul et al., 2023] | relation | relation | relations | False | False | 2 |
| voc2007 [Everingham et al.] | zero-shot | object recognition | standard object recognition | False | True | 20 |
| winoground [Thrush et al., 2022] | relation | relation | relations | False | False | 2 |

Table 6: List of all the benchmarks used in evaluations with their corresponding dataset type, capability, number of classes, whether they are curated and whether they are curated object centric.