Scaling Vision-Language Models Does Not Improve Relational Understanding: The Right Learning Objective Helps

Haider Al-Tahan†, Quentin Garrido†, Randall Balestriero, Diane Bouchacourt†, Caner Hazirbas†, Mark Ibrahim†
†FAIR, Meta
{haideraltahan, garridoq, dianeb, hazirbas, marksibrahim}@meta.com, randallbalestriero@gmail.com

Abstract

Vision-language models (VLMs) have emerged as an effective strategy for multi-modal representation learning. The field has invested extensive efforts to develop a multitude of models pushing the boundaries of model and training data scales. Given the growing set of VLM benchmarks, however, model releases often selectively evaluate a subset of benchmarks. Consequently, drawing principled conclusions about optimal strategies for advancing VLMs in this fragmented landscape is a challenge. In this study, we systematically benchmark four axes of performance: zero-shot classification, relational understanding, robustness, and resilience to corruptions. We evaluate all axes across nearly 60 VLMs, including recent large-scale models such as EVA-CLIP, with scales up to 4.3B parameters and 12.8B training samples. Despite the field's investment, we find scale, while helpful for other axes of performance, does not aid relational understanding. We also find transformer-based architectures are more resilient against image corruptions compared to other architectures. Finally, we highlight improved learning objectives as a promising avenue for advancing relational understanding.

1. Introduction

Pre-training visual models with language supervision, exemplified by CLIP [23], has emerged as an effective and straightforward strategy for multimodal representation learning. VLMs have been shown to exhibit exceptional adaptability, displaying zero-shot capabilities in classification and transferability [23], text and image retrieval [4, 10, 22], robustness [34], and compositional relationships [14, 28, 33]. Despite these advancements, the field lacks a comprehensive corpus that evaluates the performance of VLMs across a diverse range of benchmarks and model types. This fragmentation hinders the ability of researchers to draw principled conclusions about the optimal strategies for further advancing the capabilities of VLMs. Thus, there is a clear need for a comprehensive analysis that bridges this gap, providing valuable guidance for future research directions and model development in the field of visual representation learning.

Factors	Benchmarks			
	ImageNet	Relation	Robustness	Corruption
Dataset size	✔	✗	✔	✔
Model size	✔	✗	✔	✔
Architecture	∼	∼	∼	✔
Learning Objective	∼	✔	∼	∼

Table 1. **Relational understanding requires more than scaling dataset and model size**. We shows the impact of various factors on the performance of VLMs. Learning objective is the only factor that helps improving the robustness (✔) on the Relational benchmarks (Figure 1) while it has insignificant impact (∼) on the rest of the benchmarks. In contrast, both scaling dataset and model size have adverse impact (✗).

Current work in this direction provides isolated insights [3], showing that scaling training data and model sizes improves performance on ImageNet and robustness benchmarks. However, we believe there is a notable absence of an integrated overview for understanding how VLMs fair across other axes of performance. Furthermore, many existing investigations are limited in the number of VLMs studied with many focusing only on the original CLIP model [3].

To address this gap, we study the performance of 59 vision-language models across diverse learning objectives and architectures with scales of up 12.8 billion training samples and 4.3 billion parameters. We evaluate several axes of performance including relational understanding and resilience to image corruptions to provide an overarching perspective of VLMs capabilities and shed light on promising future research directions.

We illustrate the takeaways in Table 1 and summarize specific our contributions as follows

1. **Scaling does not improve relational understanding**: While expanding the model size and training data vol-

ume serves as effective methods for boosting performance on many benchmarks, they do not suffice to enhance relational understanding (Figure 2). Despite training colossal models with up to 12.8 billion samples [9] and 4.3 billion parameters [7, 8], they show no improvement on relational understanindg.

2. **ViTs are more robust to corruptions**: Vision transformer (ViT) encoders demonstrate to be more capable at handling image corruptions compared to convolutional architectures.
3. **CLIP's learning objective requires rethinking**: We find models with richer learning objectives, such as NegCLIP and BLIP, that either include hard negatives or use an objective which adds image-to-text matching and image-conditioned language modeling, perform significantly better on relational understanding.

2. Evaluation Setup

We outline below the axes of performance we evaluate along with the choice of benchmarks for each. Next we discuss the diverse set of 59 VLMs (see Table 2) we investigate across architectures, learning paradigms, and model/data scales. Together the set of benchmarks and models provide a comprehensive apples-to-apples overview of the facets of model performance as well as the strategies most effective for each.

2.1. Benchmarks

The evaluation of the models in this study is conducted on four distinct axes, each providing a unique perspective on models' performance. These benchmarks have been carefully selected to cover a wide range of scenarios and challenges that the models may encounter in real-world applications.

1. **ImageNet:** The ImageNet dataset [17] is a large-scale, diverse dataset that is widely used for benchmarking in the field of computer vision. It provides a broad base for assessing the model's ability to understand and represent a wide variety of objects and scenes.
2. **Relation:** We include relational benchmarks, that are, *Visual Genome* [33], Winoground [28], and SugarCrepe [14] are designed to evaluate the models' ability to understand and represent relationships between objects within an image (Figure 1). This is a crucial aspect of vision-language models, as understanding the relationships between objects can provide valuable context for interpreting the image. For instance, *Visual Genome* benchmark includes a variety of relationships (denoted VG-Relation) and attributions (denoted VG-Attribution) tasks, such as spatial relationships (*e.g.*, "above", "next to"), action relationships (*e.g.*, "riding", "holding"), and appropriate attribution (*e.g.*, "the brown horse and the orange cat" vs. "the orange horse and the orange brown"). *Visual Genome* also includes COCO-order and Flickr30k-order, which assess the models' sensitivity to word order (*e.g.*, "a brown cat" vs. "cat a brown").
3. **Robustness:** This set is a collection of several datasets, including ImageNet-E [18], ObjectNet [2], ImageNet-R [13], ImageNet-9 [29], and ImageNet-V2 [25]. These datasets are designed to test the model's robustness to various transformations and perturbations. For example, the ObjectNet dataset introduces changes in object position, scale, and background, while the ImageNet-R dataset focuses on transformations related to many types of image renditions.
4. **Corruption:** Consisting of ImageNet-C dataset [12] introduces various types of image corruptions, such as noise, blur, and digital artifacts. These corruptions simulate the types of degradation that images may undergo in real-world scenarios, such as poor lighting conditions, low-quality cameras, or transmission errors.

By evaluating the models on these diverse datasets, we aim to provide a comprehensive assessment of their performance, robustness, and ability to handle a variety of real-world scenarios.

2.2. VLMs under Interrogation

We evaluate 59 VLMs (see Table 2) on 4 benchmarks (Section 2.1) across a range of model size, pre-training dataset size, learning objective, and architectures (full list in Table 2). For dataset size, we included models trained and/or fine-tuned with datasets ranging from 13 million to 12.8 billion samples; which include DataComp [9] (small, medium, large, and extra-large), LIAON [26] (400M, 2B, 5B), MetaCLIP [30] (400M and 2.5B), Flickr [31], PMD [27], and COCO [20]. For model size and architecture, we categorize models based on the number of parameters and either convolutional or transformer based models, ranging from ResNet50 [11] with 38 million parameters to EVA02 ViT E [8] with 4.3 Billion parameters. Lastly, for learning objective, we included SigLIP [36], NegCLIP [33], BLIP [16], among others.

2.3. Evaluation Procedure

We evaluate performance of zero-shot classification datasets similar to [24], by contrasting the representations of class labels (averaged across 80 prompts) with image representation and using the class with the highest probability as the predicted class. Alternatively, we evaluated performance of relation datasets by contrasting correct and incorrect captions with image representation and using the correct caption as the ground-truth label.

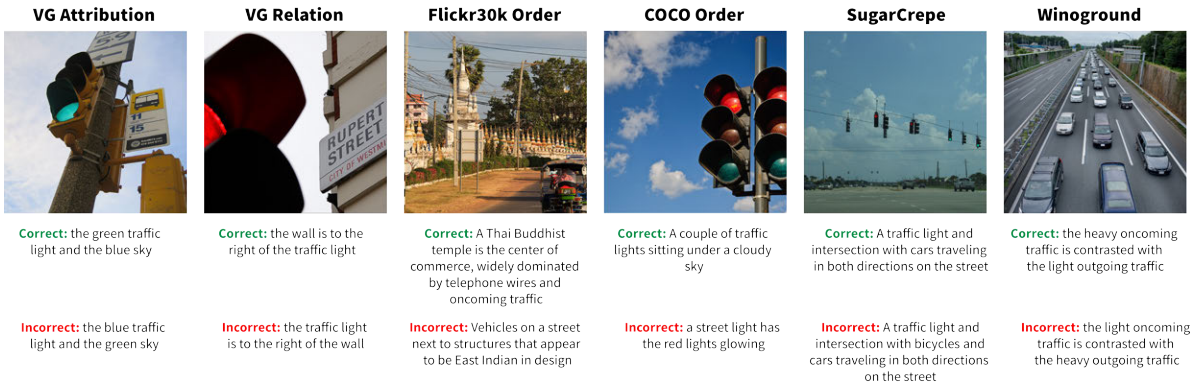

Figure 1. **Relational understanding benchmarks.** Examples with their respective expected correct vs incorrect captions.

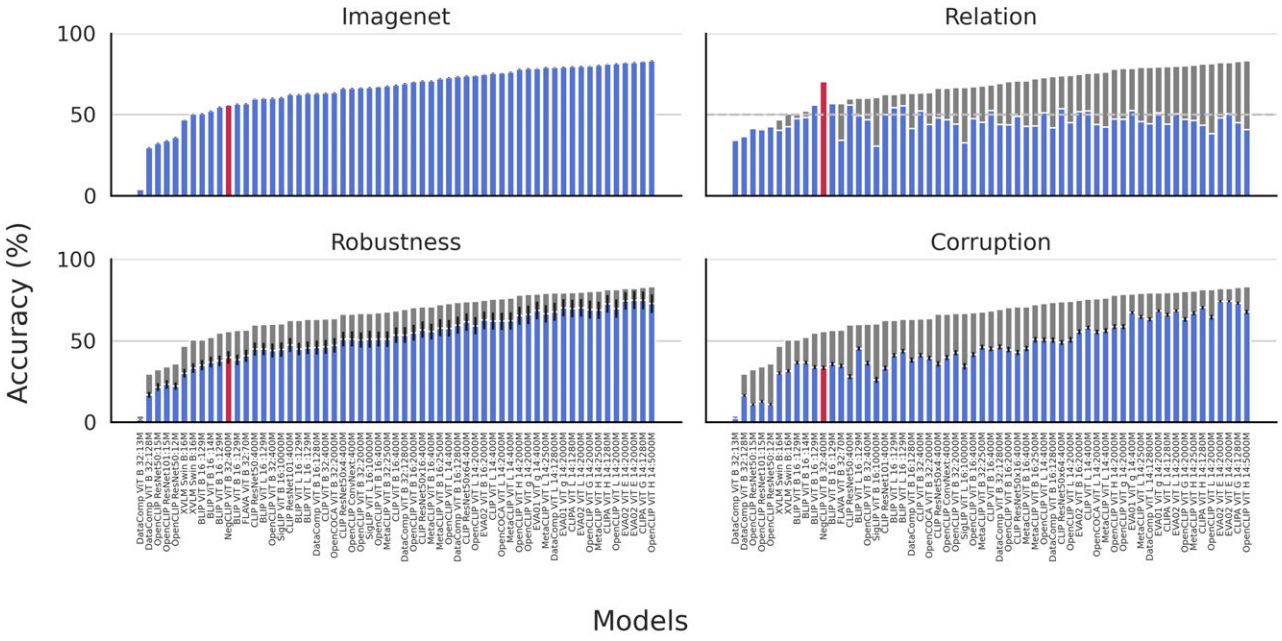

Figure 2. **Relational understanding does not improve with better ImageNet performance.** Average zero-shot performance of all models across all dataset benchmarks (Section 2.1). Grey-colored bars reflect ImageNet zero-shot performance, blue-colored bars reflect performance across other benchmarks, red-colored bars reflect performance of NegCLIP. The x-axis outlines the names of the models, with the size of the dataset they were pre-trained on, $[ModelName] : [DatasetSize]$. Grey-dashed line represent chance level.

3. Results

We show the overall performance of the nearly 60 VLMs we examined in Figure 2 ranked by their zero-shot classification performance on ImageNet. We find performance trends for robustness and corruption improve with increasing ImageNet accuracy, but relational understanding is mixed, suggesting relational performance may not be correlated with standard classification and robustness benchmarks. Next we investigate the effect scaling training data, model size, learning paradigm, and architecture to isolate the contribution of each to the facets of VLM performance.

3.1. Scaling training data does not improve relational understanding

In line with previous research [3, 24], we found that increasing the size of the training data improves the zero-shot classification and robustness of VLMs across various benchmarks. Our findings, as illustrated in Figure 3, uncover that for relational understanding the trend does not hold. The performance on Relation benchmarks does not increase proportionally with the size of the training dataset. In fact, most VLMs barely reach chance level on these benchmarks, even when dataset sizes are scaled up to 12.8B sam-

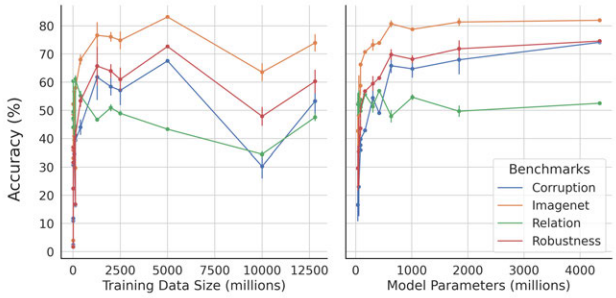

Figure 3. **Relational understanding requires more than scaling dataset and model size.** Average zero-shot performance of models across Relation benchmarks. We investigate the impact of dataset size (left), and model size on relational understanding (right).

ples (Figure 6). NegCLIP is a notable exception show in red in Figure 2 (see Appendix A.5). We run an additional control in Figure 5 to isolate the effect of training data size keeping other factors such as architecture and learning objective fixed. We observe similar trends with respect to scaling. Next we examine the trends when the model size rather than data is scaled.

3.2. Scaling VLM size does not improve relational understanding

To examine scaling with respect to model size, we plot in the right panel of Figure 3 performance across all four axes as a function of model sizes ranging from 32M to 4.3B. We find, while scaling model size improves ImageNet, robustness, and corruptions, it has little effect on relational understanding. We find even large models such aas EVA02 E14 with 4.3B parameters are at near chance level for many relational benchmarks. We perform an additional control to confirm this trend by fixing the learning paradigm and training data size in Figure 7. We examine architecture separately in Appendix A.4. Analogous to scaling data size, while scaling boosts ImageNet performance by 39.3% relational understanding remains flat suggesting it's an open challenge that scale alone does not address. These findings underscores the importance of considering other factors, such as the choice of learning objectives and training strategies, when aiming to improve VLMs' performance on relational understanding tasks

3.3. ViT encoders perform better on corruption tasks

Next, we examine the role of architecture across the axes of performance. As shown in Figure 2, all of the top 20 (out of the 59 models) for ImageNet, robustness and corruption are ViT-based. To isolate other confounding factors, we compare the choice of encoder architecture while controlling for the model size, learning paradigm, and training data size in Figure 9. We find the choice of encoder architecture, whether ViT or convolutional, has little effect on performance across standard ImageNet, relational, and robustness. However, we find ViT models perform much better on corrupted images consistent with prior findings for vision-only supervised models in Bai et al. [1]. For example, transformer-based VLM performs $5.31 - 9.13\%$ better on Corruption benchmarks relative to comparably sized convolutional-encoder models.

3.4. Rethinking learning objectives might be the solution for relation understanding

While scaling model and training data sizes improves standard classification and robustness benchmarks, relational understanding does not improve with scale. Can better learning objectives help?

We show in Figure 2 the highlighted NegCLIP model with a tailored learning objective for capturing relations via hard-negatives seems to perform remarkably better on relational understanding. NegCLIP, with only 86M parameters, significantly outperforms models up to $50\times$ 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. We note NegCLIP is finetuned with 330k labeled relations, which suggests carefully curated data may also be a helpful lever. We further breakdown the performance of relational understanding benchmarks in Figure 4. We find in addition to NegCLIP, BLIP models emerge with better performance on attribute-based relational understanding (see highlighted bars in Figure 4). Similar to NegCLIP, BLIP models incorporate a richer learning objective that includes image-to-text matching and image-conditioned language modeling in addition to the standard contrastive objective. This preliminary finding suggests rather than scale, the right learning objective can be a promising strategy for improving relational understanding.

4. Discussion

Limitations Our analysis relies on the faithfulness and consistency of existing benchmarks. In particular, relational benchmarks vary in difficulty [5] as well as their construction. For example, Winoground is more challenging as it contains a balanced number of counterfactual samples whereas others do not [21]. We provide a breakdown of relational understanding by benchmark in Appendix A.

By providing a comprehensive apples-to-apples overview of performance across nearly 60 VLMs, we surfaced an important trend for the research community: while scale is a promising lever for several axes of performance, targeted learning objectives for relational understanding along with a better understanding of data quality are needed to advance VLM's relational understanding capabilities.

References

[1] Yutong Bai, Jieru Mei, Alan Yuille, and Cihang Xie. Are transformers more robust than cnns?, 2021. 4

[2] 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*. Curran Associates, Inc., 2019. 2

[3] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 1, 3

[4] 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. 1

[5] Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. Why is winoground hard? investigating failures in visuolinguistic compositionality, 2022. 4

[6] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks, 2023. 7

[7] 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. 2, 7

[8] 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*, 2023. 2, 7

[9] 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. 2, 7

[10] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A. Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining, 2022. 1

[11] 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. 2

[12] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations, 2019. 2

[13] 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, 2021. 2

[14] 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. 1, 2

[15] 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, 2021. If you use this software, please cite it as below. 7

[16] 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. 2, 7

[17] Xiaodan Li, Yuefeng Chen, Yao Zhu, Shuhui Wang, Rong Zhang, and Hui Xue. Imagenet-e: Benchmarking neural network robustness via attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20371–20381, 2023. 2

[18] Xiaodan Li, Yuefeng Chen, Yao Zhu, Shuhui Wang, Rong Zhang, and Hui Xue. ImageNet-E: Benchmarking Neural Network Robustness via Attribute Editing, 2023. arXiv:2303.17096 [cs]. 2

[19] Xianhang Li, Zeyu Wang, and Cihang Xie. An inverse scaling law for clip training, 2023. 7

[20] 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. 2

[21] Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. Revisiting the role of language priors in vision-language models, 2024. 4

[22] 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. 1

[23] 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, 2021. arXiv:2103.00020 [cs]. 1

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 7

[25] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet?, 2019. 2

[26] 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. 2

[27] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model, 2022. 2, 7

[28] 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. 1, 2

[29] 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. 2

[30] 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. 2, 7

[31] 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. 2

[32] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models, 2022. 7

[33] 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?, 2023. arXiv:2210.01936 [cs]. 1, 2, 7

[34] Tian Yun, Usha Bhalla, Ellie Pavlick, and Chen Sun. Do vision-language pretrained models learn composable primitive concepts?, 2023. 1

[35] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts, 2022. 7

[36] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. 2, 7

A. Appendix

In this section, we further detail our experimental setup and provide more results. Appendix A.1 gives an overview to the Relation benchmarks and we detail our control factors in Appendices A.2 to A.5. Finally, we list the models we used in our experiments in Appendix A.6.

A.1. Overview Relation Benchmarks

Figure 4 shows zero-shot models performance all Relation benchmarks (Section 2.1). Figure 4 provides a detailed comparison of the performance of various VLMs, particularly highlighting the effectiveness of the NegCLIP and BLIP models across different relational benchmarks. This figure illustrates how the NegCLIP model, with its learning objective that incoporates hard negatives, excels in relational understanding compared to other models. Interestingly, BLIP outperforms NegCLIP and other models on VG Attribution, Winoground, and Sugarcrepe benchmarks, while falling short on Flickr30K order, COCO order, and VG Relation benchmarks. This demonstrate that BLIP's objective which adds image-to-text matching and image-conditioned language modeling allows models to perform better on attribution-based tasks. Through Figure 4, we gain a comprehensive view of how different models stack up against each other in the realm of relational understanding, highlighting the necessity of richer learning objectives and training strategies for relational understanding tasks.

A.2. Training Data Size

Figures 5 and 6 provides a focused examination of how the scaling of training dataset sizes influences the performance of VLMs on various benchmarks. Figure 5 shows that increasing dataset size beyond 2 billion samples reaches a demonishing return on ImageNet, Robustness, and Corruption benchmarks. For instance, increasing dataset size from 400 million to 2 billion samples, improves performance by 6.36%. Alternatively, increasing dataset size from 2 billion to 12.8 billion samples, improves performance by 1.57%.

Figure 6 also shows that contrary to the positive impact of increased dataset size on benchmarks like ImageNet, Robustness, and Corruption, the figure illustrates a starkly different scenario for relational tasks. It highlights that, despite the substantial escalation of training data up to 12.8 billion samples, most VLMs do not exhibit significant improvement in relational understanding, often performing near or at chance levels. This suggests a plateau in performance gains from dataset scaling in the context of relational benchmarks. This divergence underscores the limited effectiveness of mere data scaling in relational contexts and hints at the necessity for targeted learning strategies to overcome the inherent challenges in relational understanding for VLMs.

A.2.1 Figure Controls

In Figures 5 and 6, we isolate the effect of training data size by controlling for other factors. To do so, we use the same ViT-B/32 architecture trained with the same contrastive CLIP objective over different number of training samples. These include models trained with DataComp (small, medium, large, and extra-large), LIAON (400 millions and 2 billions), and MetaCLIP (400 millions and 2.5 billions).

A.3. Model Size

Figures 7 and 8 provide a detailed examination of the impact of model size on the performance of VLMs across various benchmarks. Figures 9 and 10 highlights that increasing the model size does not correspond with better performance on relational benchmarks, suggesting that relational understanding requires more than just larger models.

A.3.1 Figure Controls

We show a controlled analysis of performance as a function of model size keeping training data size and learning paradigm fixed in Figure 9 and Figure 10. To do so, we use either ViT or ResNet architectures trained with the same contrastive CLIP objective and dataset (LIAON400M) with different number of parameters. These include ResNet50, ResNet101, ResNet50x64, ViTB32, and ViTL14.

A.4. Architecture

Figures 9 and 10 extends analysis of Appendix A.3 to compare different encoder architectures, showing that while the choice between ViT and convolutional architectures does not significantly affect performance on standard ImageNet, relational, and robustness benchmarks, transformer-based models exhibit a notable advantage in handling corrupted images.

A.4.1 Figure Control

We show a controlled analysis of performance as a function of model size and architecture keeping training data size and learning paradigm fixed in Figure 9 and Figure 10. To do so, we use either ViT or ResNet architectures trained with the same contrastive CLIP objective and dataset (LIAON400M) with different number of parameters. These include ResNet50, ResNet50x64, ViTB32, and ViTL14.

A.5. Learning Objective

Figures 11 and 12 provide a comprehensive overview of how different learning objectives influence the performance of VLMs across a range of benchmarks. Figure 11 zeroes in on the impact of various learning objectives on models'

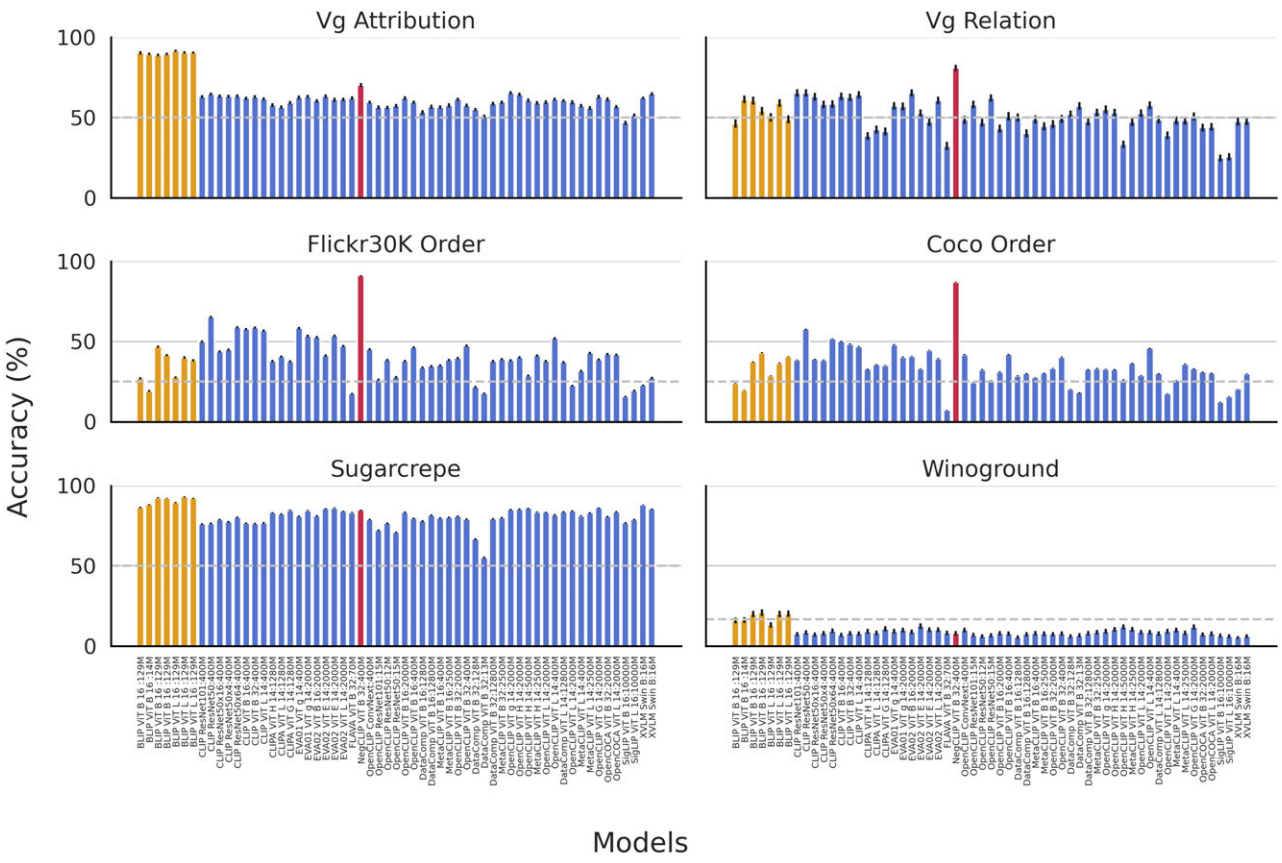

Figure 4. Average zero-shot performance of all models across Relation benchmarks (Section 2.1). Orange-colored bars reflect performance of BLIP, and red-colored bars reflect performance of NegCLIP. The x-axis outlines the names of the models, with the size of the dataset they were pre-trained on, $[ModelName] : [DatasetSize]$.

abilities to tackle relational benchmarks, illustrating that specific objectives such as NegCLIP and BLIP can significantly improve performance on relational understanding. On the other hand, Figure 12 broadens this analysis to other benchmarks, showing how the adoption of different learning objectives can also lead to varied performance across a spectrum of tasks, not just relational ones. For example, despite SigLIP being trained on a substantial dataset of 10 billion samples and comparable number of parameters to other methods such as pure contrastive and NegCLIP, it substantially underperforms in specific areas, notably Corruption and Relation benchmarks. This instance shows that even with extensive training data and substantial model complexity, the right learning objective is crucial. These figures highlights the versatility and adaptability required in selecting and designing learning objectives, emphasizing that the right choice can enhance a model's proficiency in specific tasks while potentially impacting its general performance across others.

A.6. Evaluation Setup

We show in Table 2 the list of models with their corresponding architecture, learning paradigm, model size, and training data size.

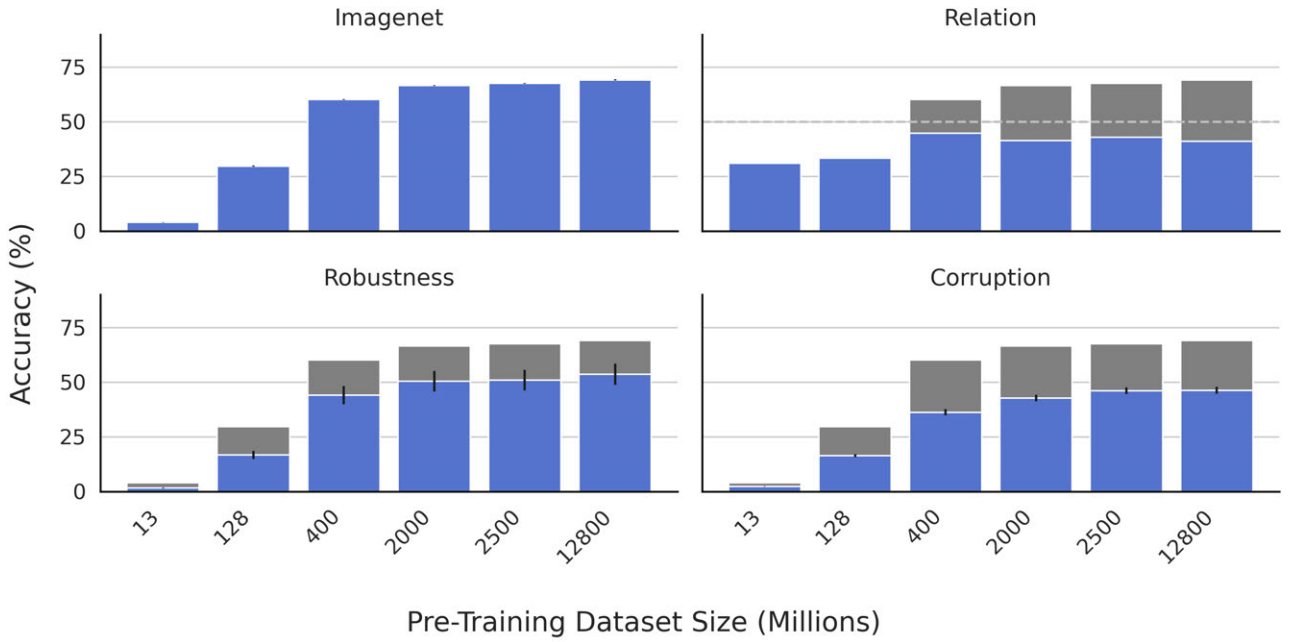

Figure 5. Average zero-shot performance of models scaled only in the number of samples across various benchmarks (Section 2.1). Grey-colored bars reflect ImageNet zero-shot performance, blue-colored bars reflect performance across other benchmarks. Grey-dashed line represent chance level.

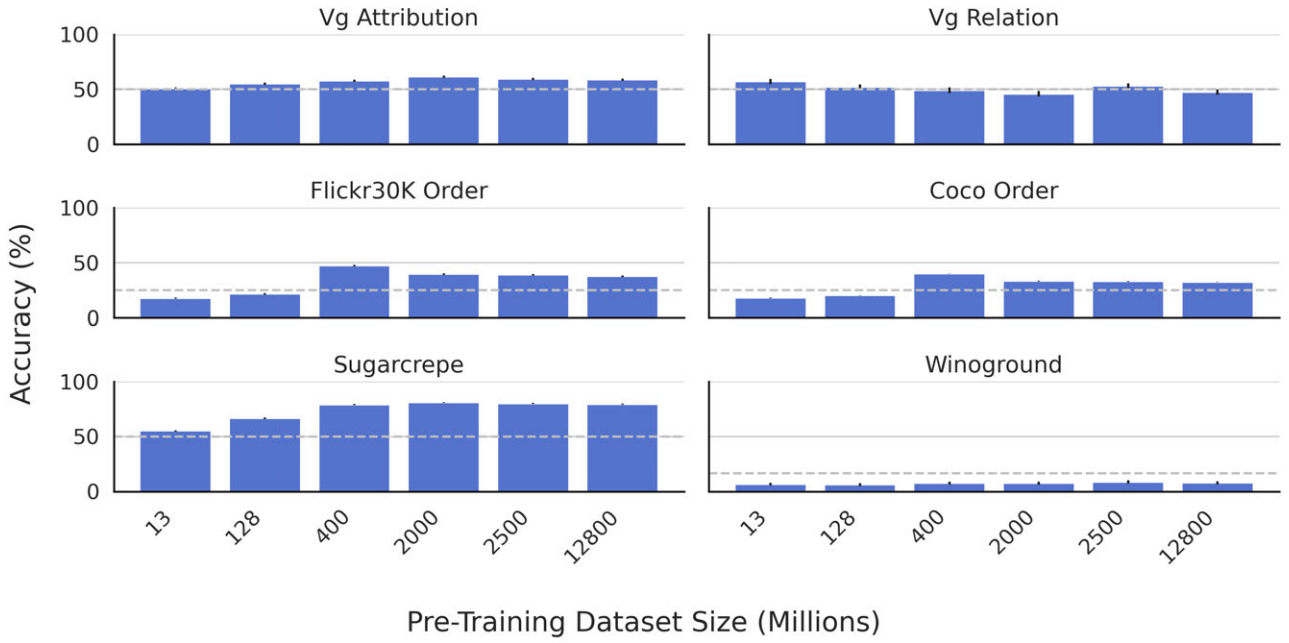

Figure 6. Average zero-shot performance on Relation benchmarks (Section 2.1) of VLMs trained on varying dataset sizes. Grey-dashed line represent chance level.

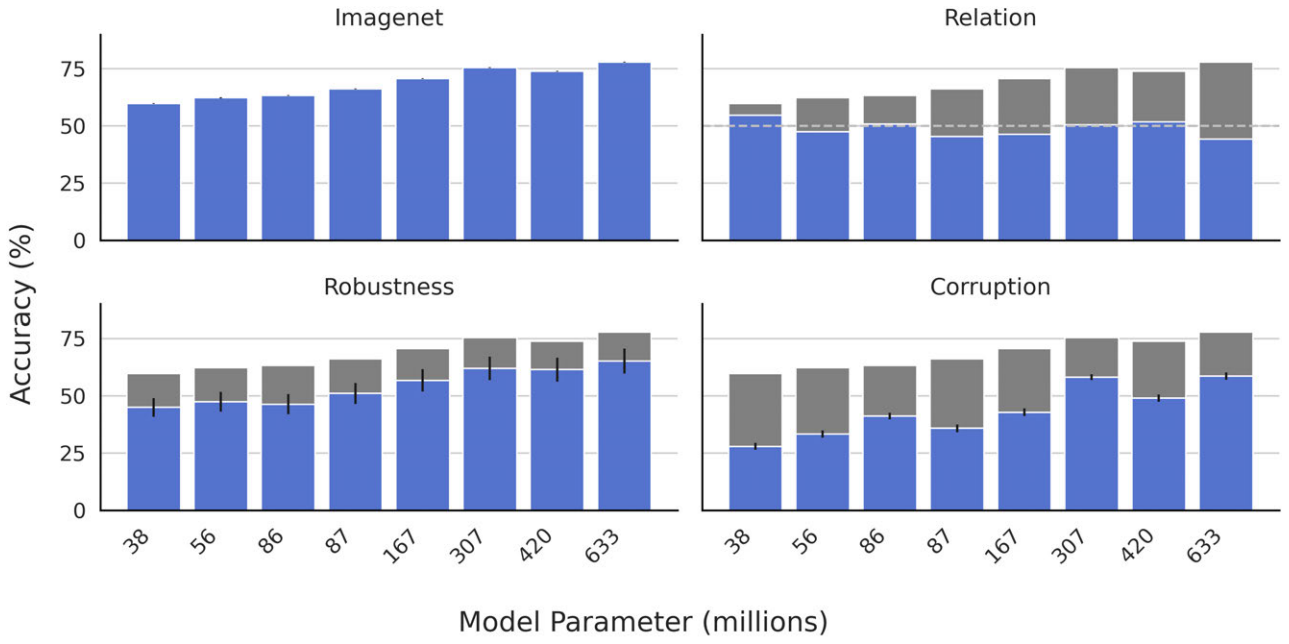

Figure 7. Average zero-shot performance of models scaled only in the number of parameters across various benchmarks (Section 2.1). Grey-colored bars reflect ImageNet zero-shot performance, blue-colored bars reflect performance across other benchmarks. Grey-dashed line represent chance level.

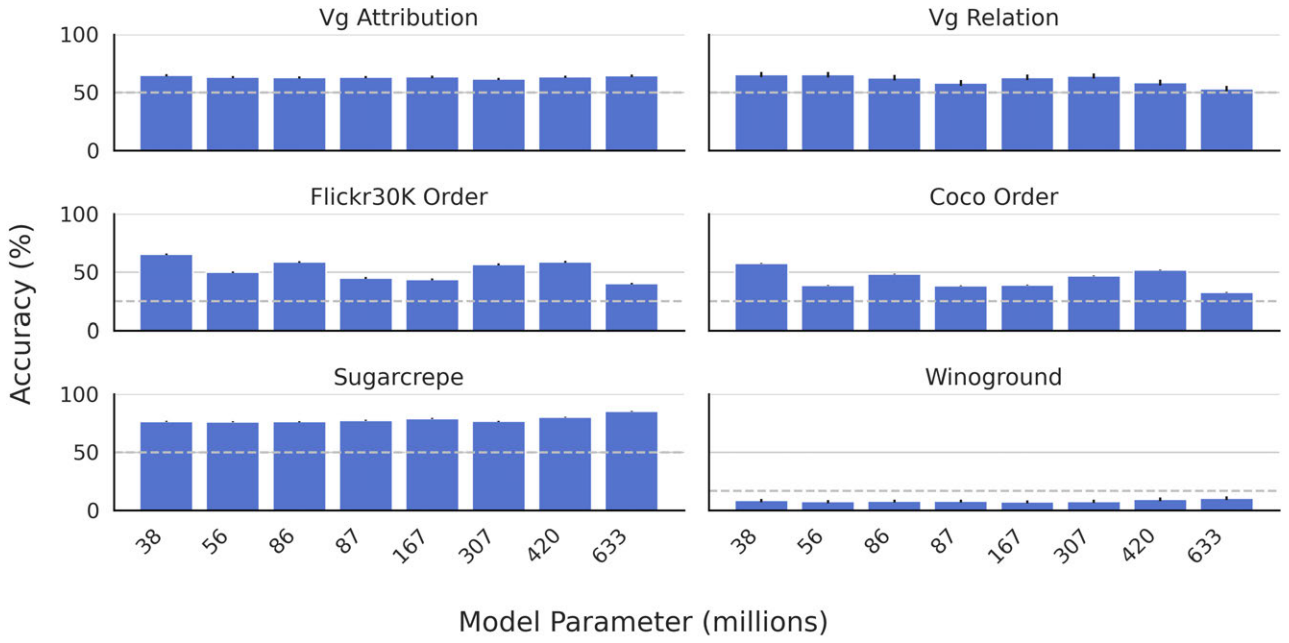

Figure 8. Average zero-shot performance on Relation benchmarks of VLMs trained on varying dataset sizes. Grey-dashed line represent chance level.

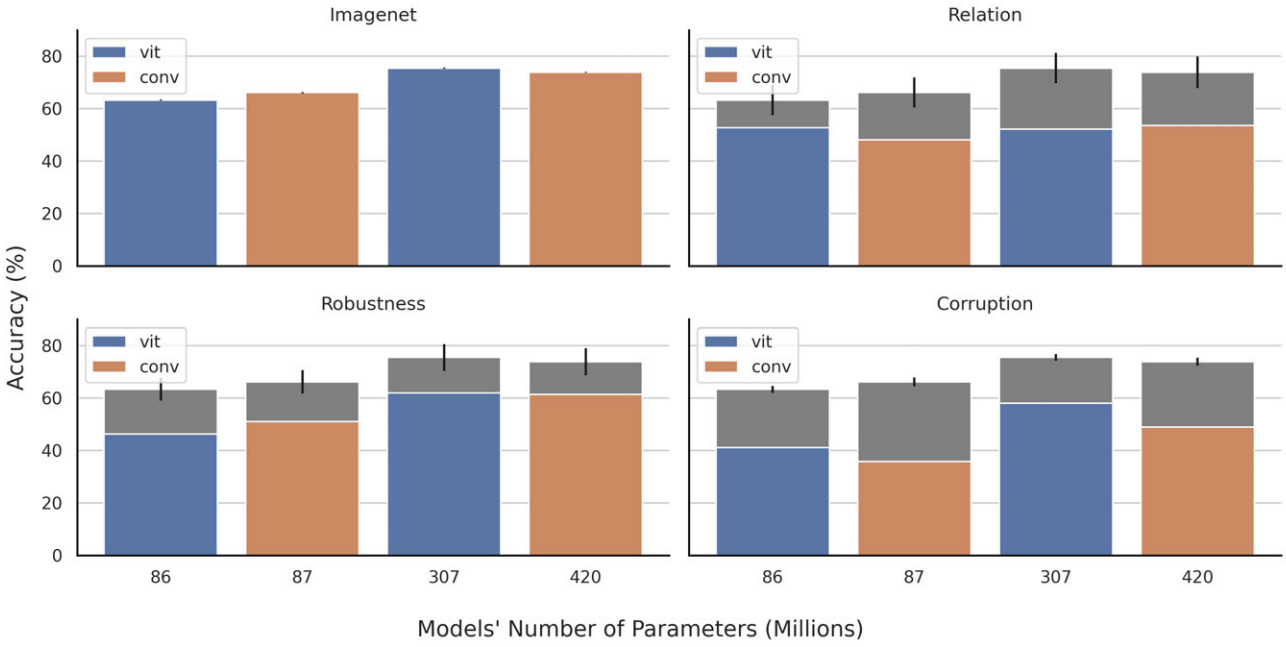

Figure 9. Average zero-shot performance of models scaled only in the number of parameters across various benchmarks (Section 2.1). Blue-colored bars reflect ViT models, and orange-colored bars reflect convolutional models. While varying model sizes and achitecture, we control for other factors that could influence performance. For instance, we only used models that are trained similar datasets.

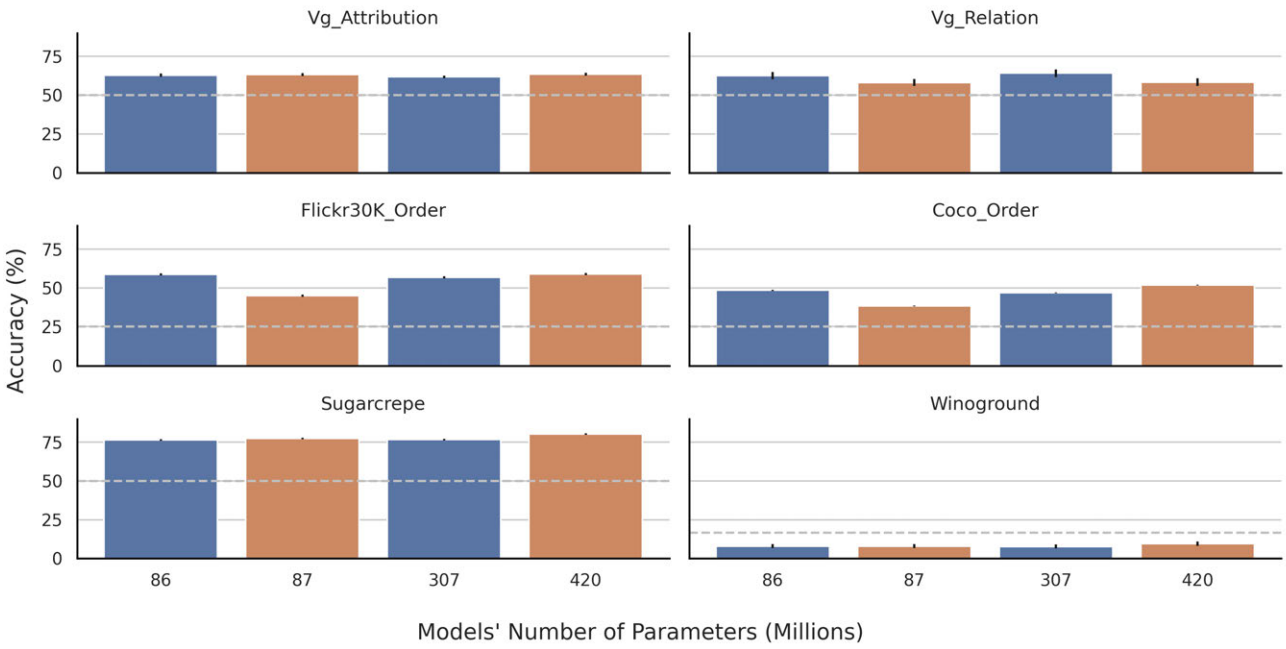

Figure 10. Average zero-shot performance on Relation datasets of VLMs trained on varying model sizes and architectures. Blue-colored bars reflect ViT models, and orange-colored bars reflect convolutional models. While varying model sizes and achitecture, we control for other factors that could influence performance. For instance, we only used models that are trained similar datasets.

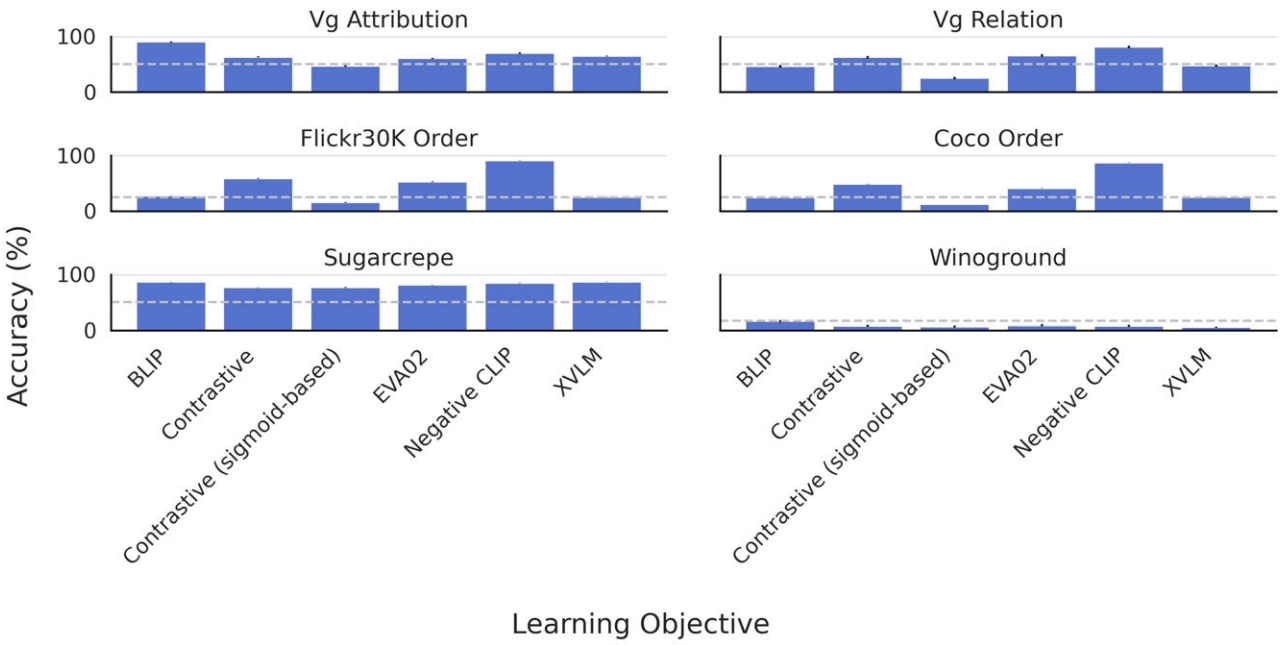

Figure 11. Average zero-shot performance of models across all datasets in the dataset zoo. There are four categories of datasets: ImageNet, Relation, Robustness, and Corruption. The following figure demonstrate that unlike ImageNet, Robustness, and Corruption datasets, Relation datasets are not correlated in models' performance. Models were ranked based on their ImageNet zero-shot performance in order to compare trends across the other categories of benchmarks.

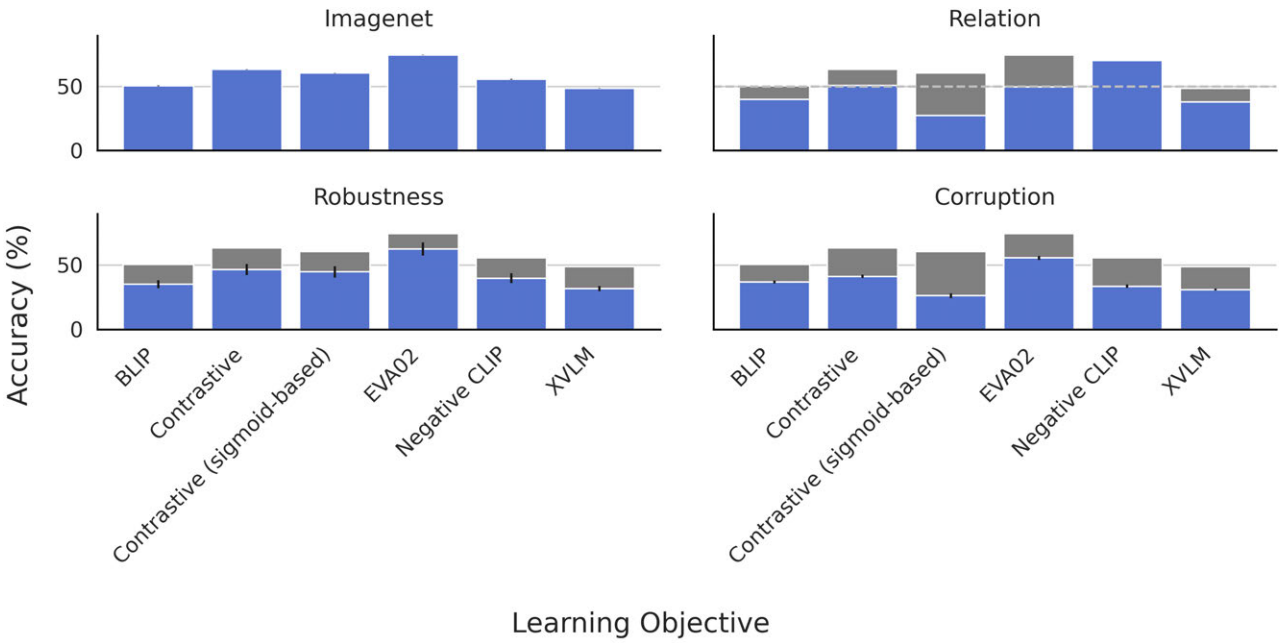

Figure 12. Average zero-shot performance of models across all datasets in the dataset zoo. There are four categories of datasets: ImageNet, Relation, Robustness, and Corruption. The following figure demonstrate that unlike ImageNet, Robustness, and Corruption datasets, Relation datasets are not correlated in models' performance. Models were ranked based on their ImageNet zero-shot performance in order to compare trends across the other categories of benchmarks.

	Dataset size	Model size	Learning objective	Architecture	Model name
blip_vitB16_14m [16]	14	86	BLIP	vit	BLIP ViT B 16
blip_vitL16_129m [16]	129	307	BLIP	vit	BLIP ViT L 16
blip_vitB16_129m [16]	129	86	BLIP	vit	BLIP ViT B 16
blip_vitB16_coco [16]	129	86	BLIP	vit	BLIP ViT B 16
blip_vitB16_flickr [16]	129	86	BLIP	vit	BLIP ViT B 16
blip_vitL16_coco [16]	129	307	BLIP	vit	BLIP ViT L 16
blip_vitL16_flickr [16]	129	307	BLIP	vit	BLIP ViT L 16
eva02_vitE14_plus_2b [8]	2000	4350	Pure Contrastive	vit	EVA02 ViT E 14
eva02_vitE14_2b [8]	2000	4350	Pure Contrastive	vit	EVA02 ViT E 14
eva02_vitL14_2b [8]	2000	307	Pure Contrastive	vit	EVA02 ViT L 14
eva02_vitB16_2b [8]	2000	86	Pure Contrastive	vit	EVA02 ViT B 16
eva01_vitG14_plus_2b [7]	2000	1011	Pure Contrastive	vit	EVA01 ViT g 14
eva01_vitG14_400m [7]	400	1011	Pure Contrastive	vit	EVA01 ViT g 14
clipa_vitbigG14 [19]	1280	1843	Pure Contrastive	vit	CLIPA ViT G 14
clipa_vitH14 [19]	1280	633	Pure Contrastive	vit	CLIPA ViT H 14
clipa_vitL14 [19]	1280	307	Pure Contrastive	vit	CLIPA ViT L 14
siglip_vitL16 [36]	10000	307	Contrastive (sigmoid)	vit	SigLIP ViT L 16
siglip_vitB16 [36]	10000	86	Contrastive (sigmoid)	vit	SigLIP ViT B 16
openclip_vitB32_metaclip_fullcc [30]	2500	86	Pure Contrastive	vit	MetaCLIP ViT B 32
openclip_vitB16_metaclip_400m [30]	400	86	Pure Contrastive	vit	MetaCLIP ViT B 16
openclip_vitB32_metaclip_400m [30]	400	86	Pure Contrastive	vit	MetaCLIP ViT B 32
openclip_vitB16_metaclip_fullcc [30]	2500	86	Pure Contrastive	vit	MetaCLIP ViT B 16
openclip_vitL14_dfn2b [6]	2000	307	Pure Contrastive	vit	OpenCLIP ViT L 14
openclip_vitL14_metaclip_400 [30]	400	307	Pure Contrastive	vit	MetaCLIP ViT L 14
openclip_vitL14_metaclip_fullcc [30]	2500	307	Pure Contrastive	vit	MetaCLIP ViT L 14
openclip_vitH14_metaclip_fullcc [30]	2500	633	Pure Contrastive	vit	MetaCLIP ViT H 14
openclip_vitH14_dfn5b [6]	5000	633	Pure Contrastive	vit	OpenCLIP ViT H 14
openclip_convnext_base [15]	400	88	Pure Contrastive	conv	OpenCLIP ConvNext
openclip_vitB32_datacomp_s [9]	13	86	Pure Contrastive	vit	DataComp ViT B 32
openclip_vitB32_datacomp_m [9]	128	86	Pure Contrastive	vit	DataComp ViT B 32
openclip_vitB32_datacomp_xl [9]	12800	86	Pure Contrastive	vit	DataComp ViT B 32
openclip_vitB16_datacomp_xl [9]	12800	86	Pure Contrastive	vit	DataComp ViT B 16
openclip_vitB16_datacomp_l [9]	1280	86	Pure Contrastive	vit	DataComp ViT B 16
openclip_vitH14 [15]	2000	633	Pure Contrastive	vit	OpenCLIP ViT H 14
xvlm_flickr [35]	16	86	XVLM	Swin	XVLM Swin B
flava_full [27]	70	86	Other	vit	FLAVA ViT B 32
openclip_vitL14_400m [15]	400	307	Pure Contrastive	vit	OpenCLIP ViT L 14
openclip_vitL14_datacomp_xl [9]	12800	307	Pure Contrastive	vit	DataComp ViT L 14
openclip_vitL14_2b [15]	2000	307	Pure Contrastive	vit	OpenCLIP ViT L 14
clip_vitL14 [24]	400	307	Pure Contrastive	vit	CLIP ViT L 14
xvlm_coco [35]	16	86	XVLM	Swin	XVLM Swin B
openclip_vitB32_400m [15]	400	86	Pure Contrastive	vit	OpenCLIP ViT B 32
openclip_vitB32_2b [15]	2000	86	Pure Contrastive	vit	OpenCLIP ViT B 32
openclip_vitG14_2b [15]	2000	1011	Pure Contrastive	vit	OpenCLIP ViT g 14
openclip_vitbigG14_2b [15]	2000	1843	Pure Contrastive	vit	OpenCLIP ViT G 14
openclip_vitB16_2b [15]	2000	86	Pure Contrastive	vit	OpenCLIP ViT B 16
openclip_vitB16_400m [15]	400	86	Pure Contrastive	vit	OpenCLIP ViT B 16
opencoca_vitL14_2b [15, 32]	2000	307	Other	vit	OpenCOCA ViT L 14
opencoca_vitB32_2b [15, 32]	2000	86	Other	vit	OpenCOCA ViT B 32
negclip_vitB32 [33]	400	86	Negative CLIP	vit	NegCLIP ViT B 32
clip_vitB16 [24]	400	86	Pure Contrastive	vit	CLIP ViT B 16
clip_resnet50 [24]	400	38	Pure Contrastive	conv	CLIP ResNet50
openclip_resnet101_yfcc [15]	15	56	Pure Contrastive	conv	OpenCLIP ResNet101
openclip_resnet50_yfcc [15]	15	38	Pure Contrastive	conv	OpenCLIP ResNet50
openclip_resnet50_cc [15]	12	38	Pure Contrastive	conv	OpenCLIP ResNet50
clip_resnet101 [24]	400	56	Pure Contrastive	conv	CLIP ResNet101
clip_resnet50x4 [24]	400	87	Pure Contrastive	conv	CLIP ResNet50x4
clip_resnet50x16 [24]	400	167	Pure Contrastive	conv	CLIP ResNet50x16
clip_resnet50x64 [24]	400	420	Pure Contrastive	conv	CLIP ResNet50x64
clip_vitB32 [24]	400	86	Pure Contrastive	vit	CLIP ViT B 32

Table 2. List of all the models used in evaluations with their corresponding dataset size, model size (number of parameters), learning objective, and architecture.