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'

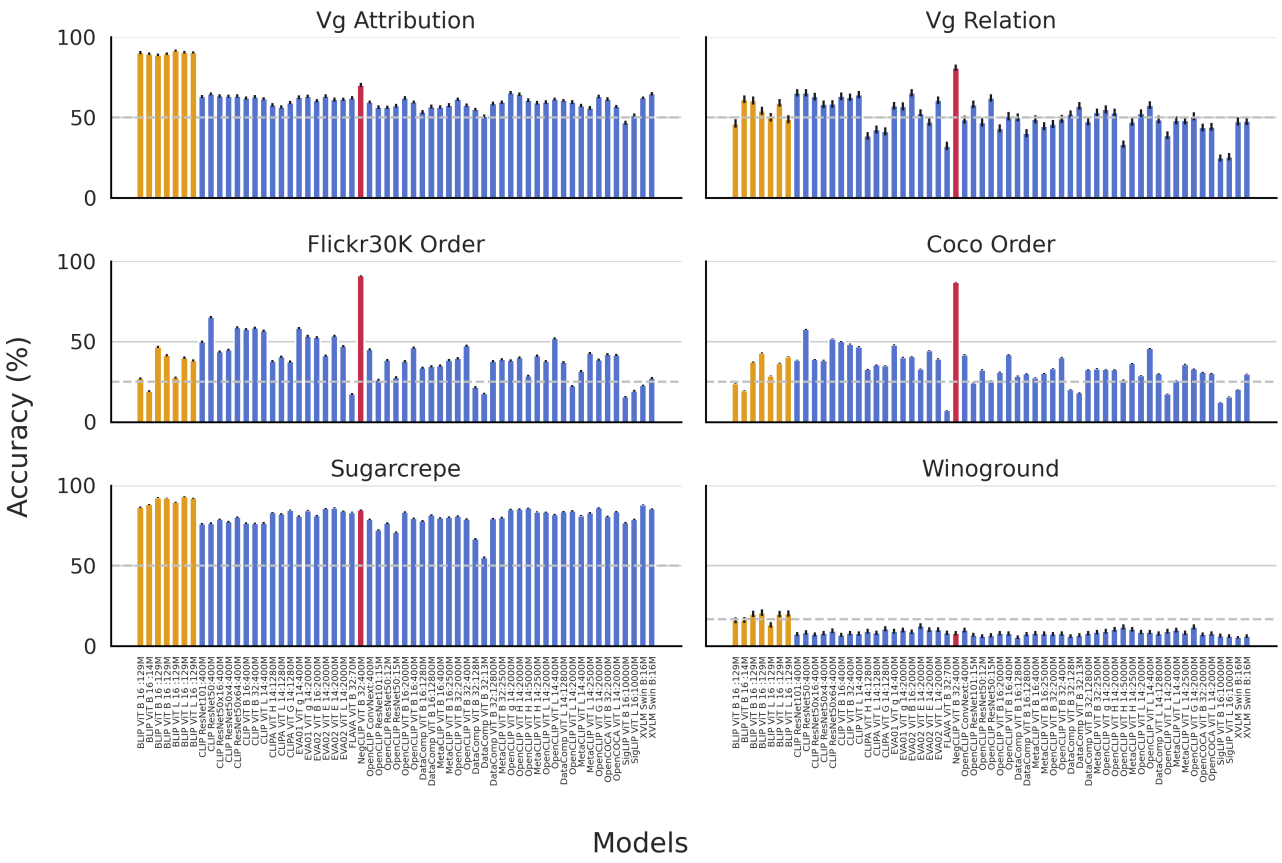

Figure 4. Average zero-shot performance of all models across Relation benchmarks (Section 2.1). Orange-colored bars reflect performance of BLIP, and red-colored bars reflect performance of NegCLIP. The x-axis outlines the names of the models, with the size of the dataset they were pre-trained on, $[ModelName]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

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.

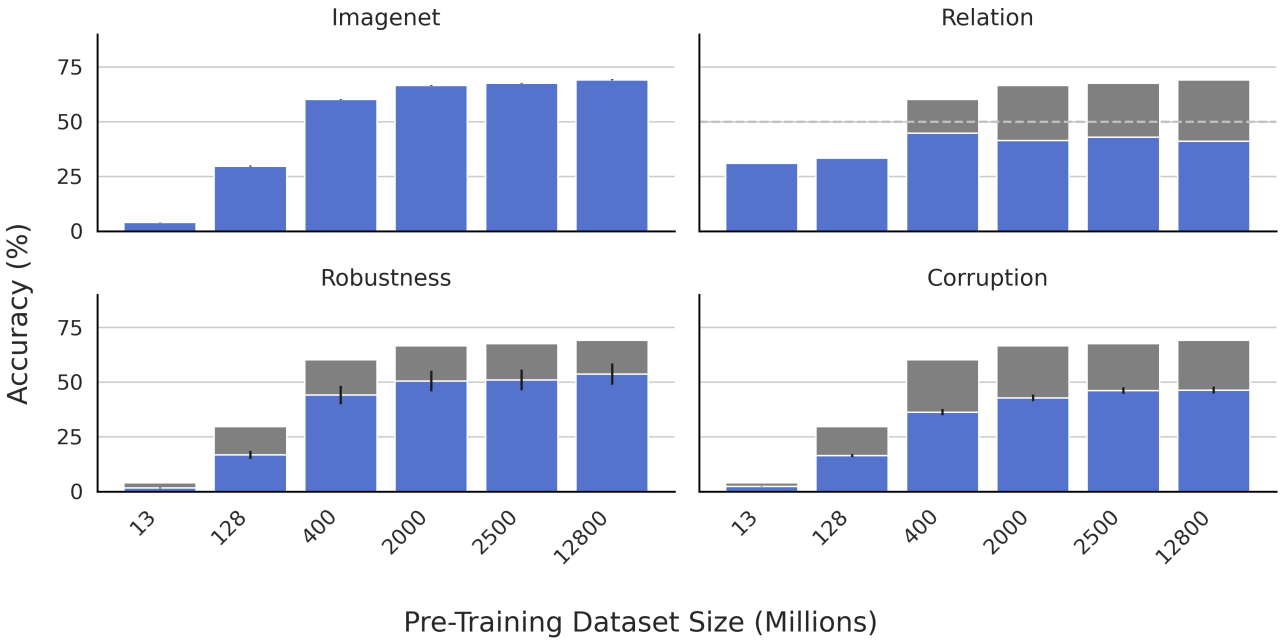

Figure 5. Average zero-shot performance of models scaled only in the number of samples across various benchmarks (Section 2.1). Grey-colored bars reflect ImageNet zero-shot performance, blue-colored bars reflect performance across other benchmarks. Grey-dashed line represent chance level.

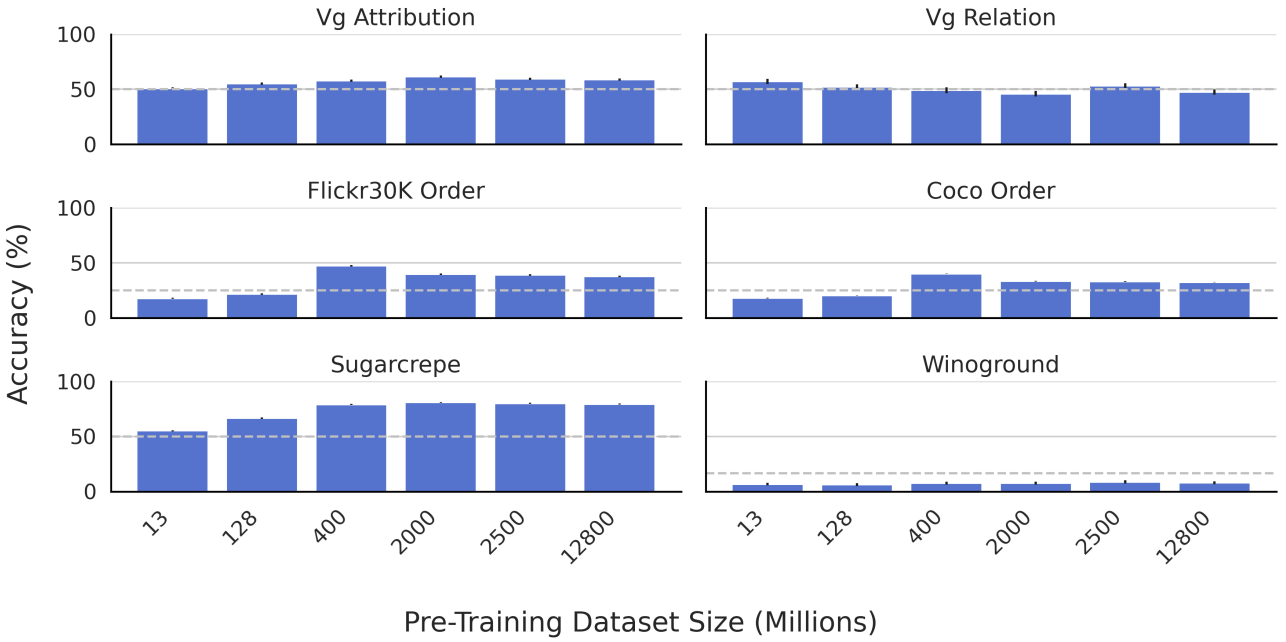

Figure 6. Average zero-shot performance on Relation benchmarks (Section 2.1) of VLMs trained on varying dataset sizes. Grey-dashed line represent chance level.

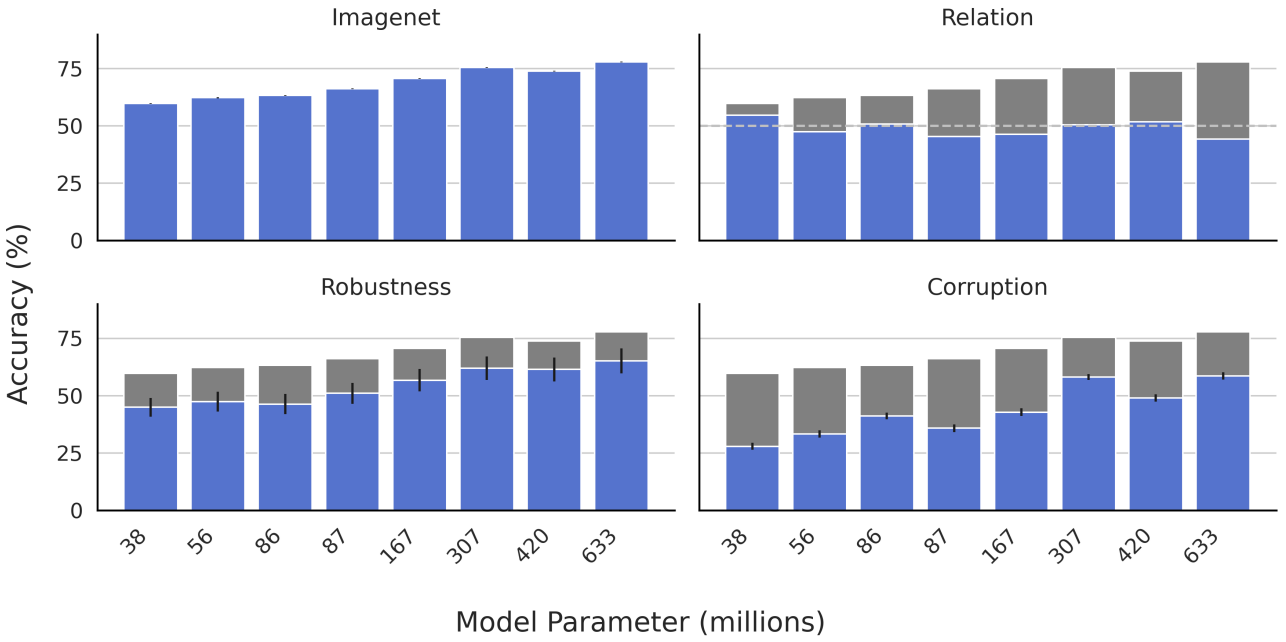

Figure 7. Average zero-shot performance of models scaled only in the number of parameters across various benchmarks (Section 2.1). Grey-colored bars reflect ImageNet zero-shot performance, blue-colored bars reflect performance across other benchmarks. Grey-dashed line represent chance level.

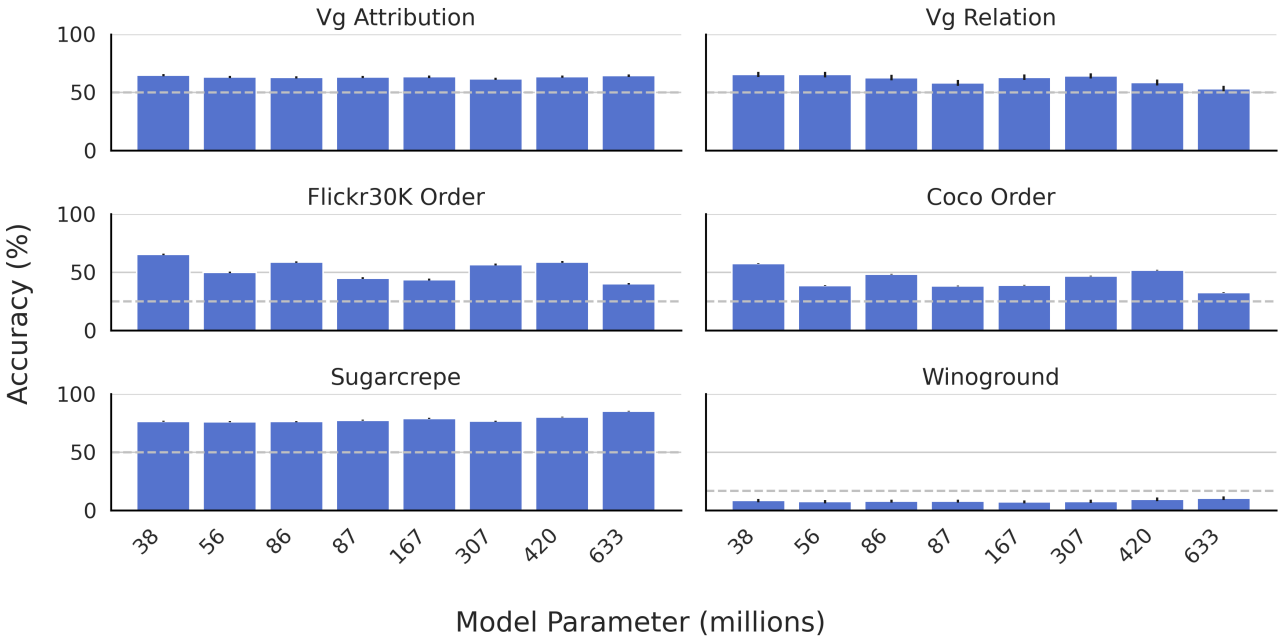

Figure 8. Average zero-shot performance on Relation benchmarks of VLMs trained on varying dataset sizes. Grey-dashed line represent chance level.

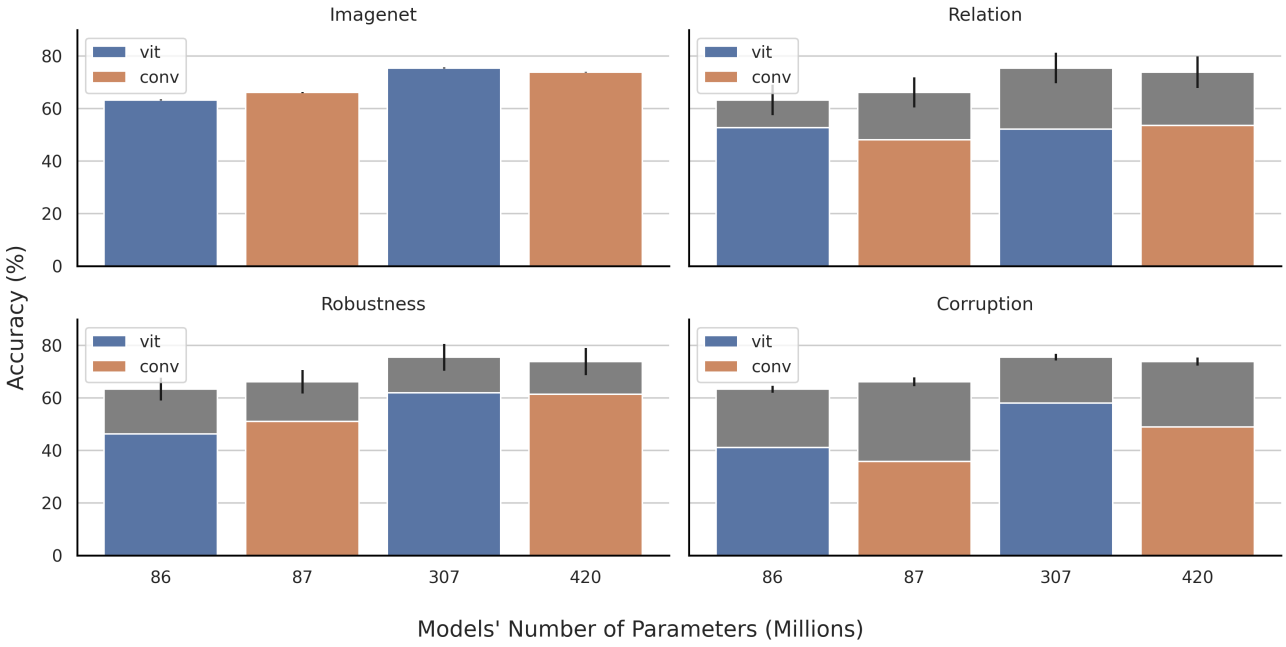

Figure 9. Average zero-shot performance of models scaled only in the number of parameters across various benchmarks (Section 2.1). Blue-colored bars reflect ViT models, and orange-colored bars reflect convolutional models. While varying model sizes and achitecture, we control for other factors that could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

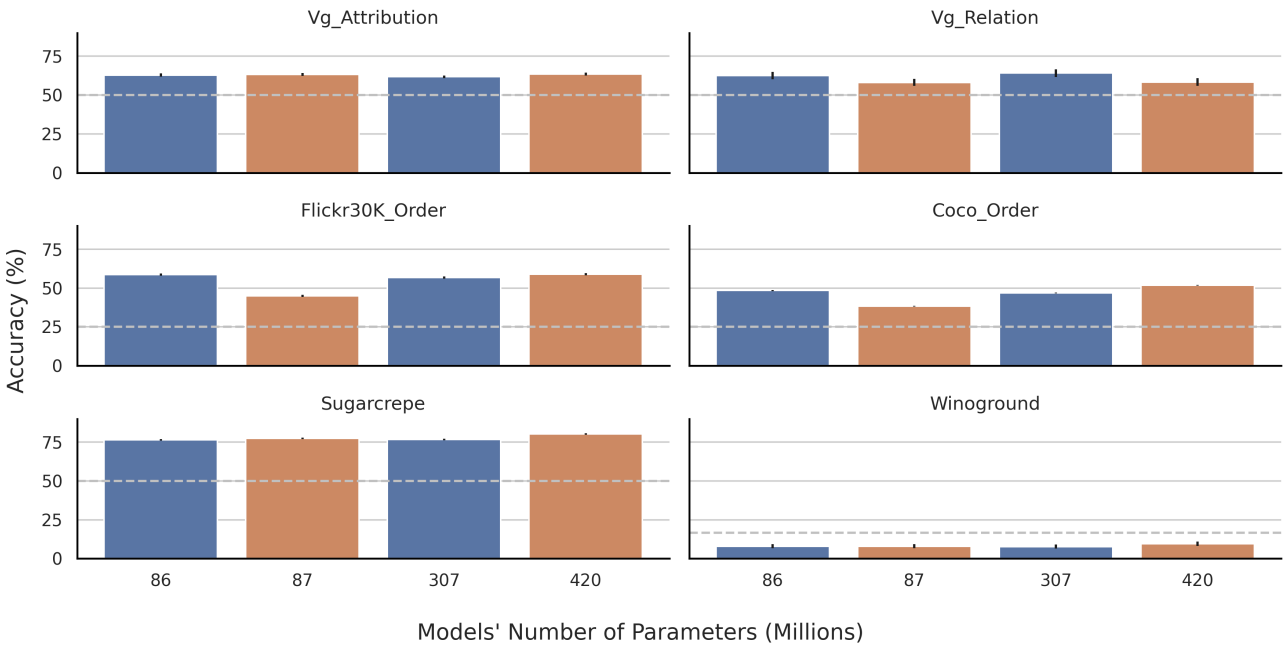

Figure 10. Average zero-shot performance on Relation datasets of VLMs trained on varying model sizes and architectures. Blue-colored bars reflect ViT models, and orange-colored bars reflect convolutional models. While varying model sizes and achitecture, we control for other factors that could influence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

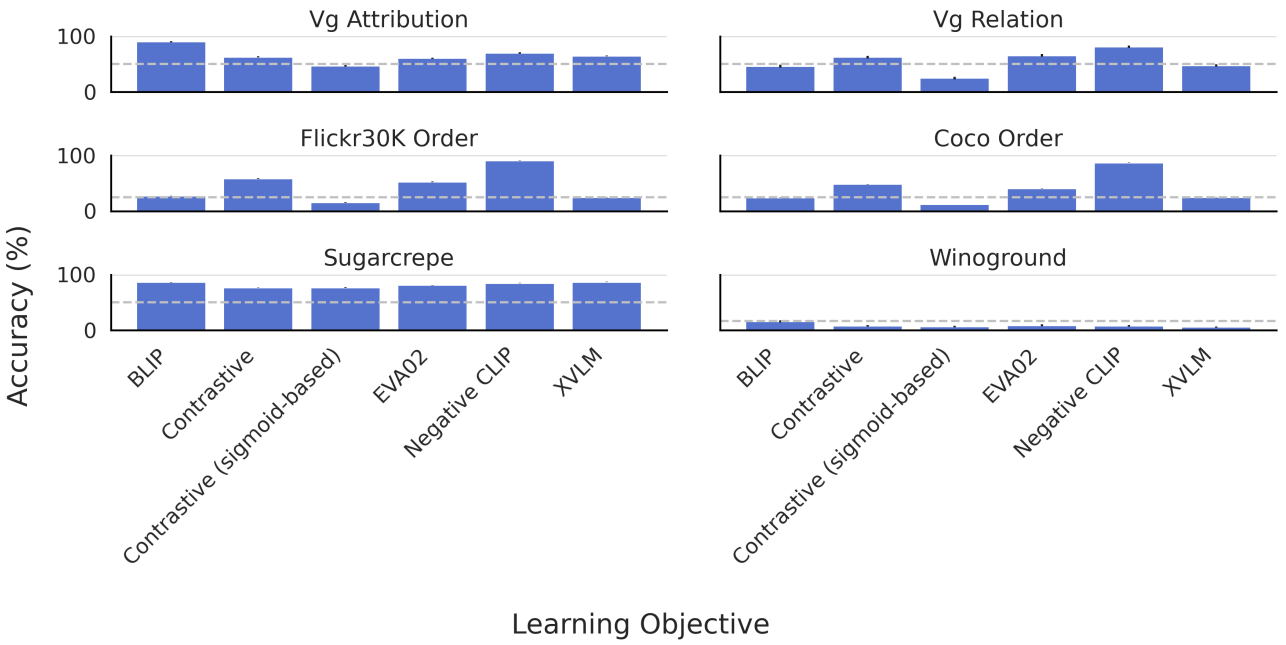

Figure 11. Average zero-shot performance of models across all datasets in the dataset zoo. There are four categories of datasets: ImageNet, Relation, Robustness, and Corruption. The following figure demonstrate that unlike ImageNet, Robustness, and Corruption datasets, Relation datasets are not correlated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

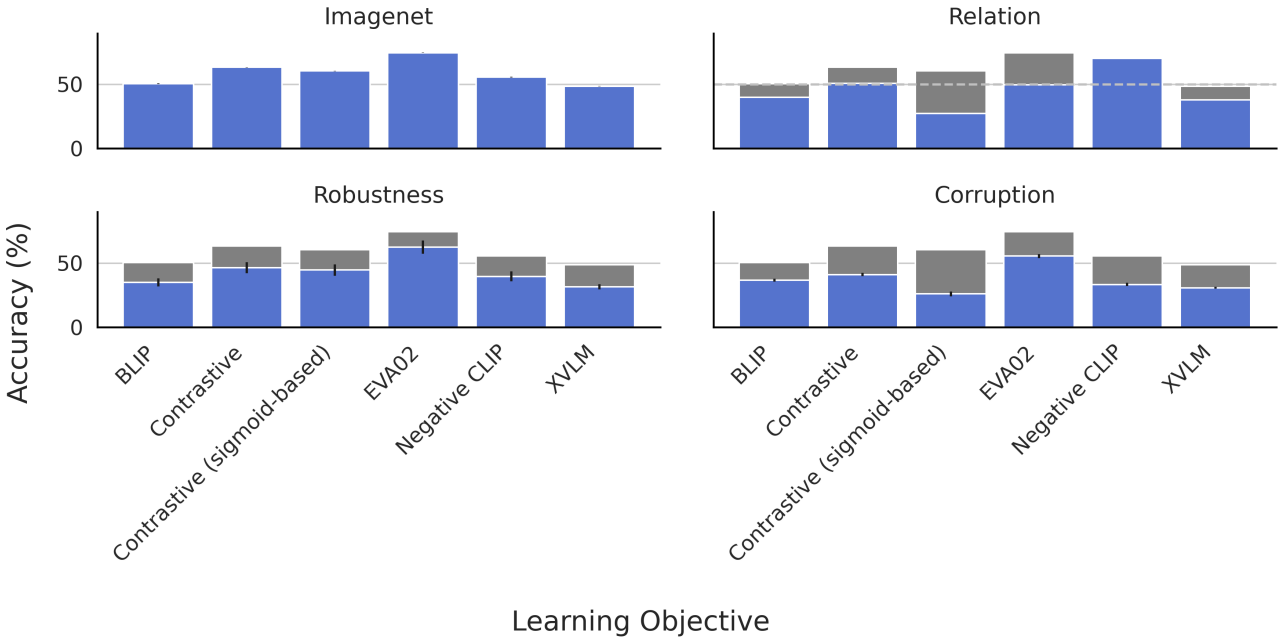

Figure 12. Average zero-shot performance of models across all datasets in the dataset zoo. There are four categories of datasets: ImageNet, Relation, Robustness, and Corruption. The following figure demonstrate that unlike ImageNet, Robustness, and Corruption datasets, Relation datasets are not correlated 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.