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 2. 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.

```python
from unibench.models_zoo import ClipModel
import unibench

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 = unibench.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 2: Custom Model Example

A.2 Natural Language Output Models on UniBench

As described in Section 2.2, LLM-style models defined as models that generate tokens/text as output. Thereby, making them hard to compare with CLIP-style VLMs. In UniBench, we also incorporated LLM-style models in a control experiments. While, LLM-style benchmarks are not suitable for evaluating CLIP-like VLMs, benchmarks in UniBench are capable of testing both LLM and CLIP style models. Following Matsuura et al. [2023] methodology, we evaluated Llava 1.5 [Liu et al., 2023] - a LLM-style VLM - on various benchmark types in UniBench (Table 2). In Table 2, we evaluated 7 and 13 billion scales of Llava.

Model Name	Corruption	Non-natural Images	Object Recognition	Reasoning	Relation	Robustness	Texture
Llava 1.5 13B [Liu et al., 2023]	31	50	36	11	41	24	34
Llava 1.5 7B [Liu et al., 2023]	29	51	32	12	42	23	28

Table 2: Performance (%) of Llava 1.5 on different Benchmark types.

A.3 Gauging progress in Vision Language Models

Scaling improves many benchmarks, but offers little benefit for reasoning and relation. Appendix Figure 7 shows that despite increasing the training dataset size by a factor of 1000× and model size by a factor of 11×, 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 8.

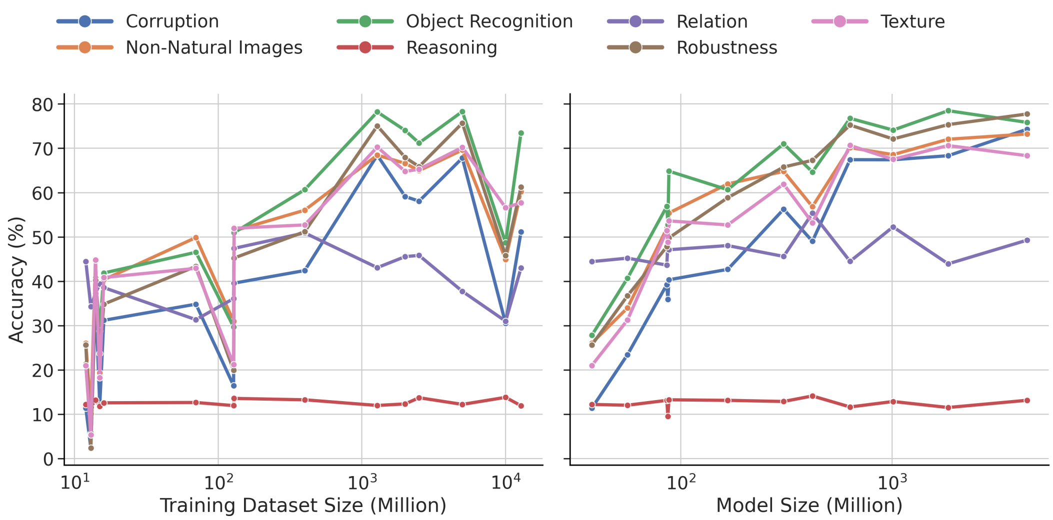

Figure 7: **The effect of scaling model and training dataset size on all models.** Median zero-shot performance of models on various benchmark types. We investigate the impact of training dataset size (left), and model size on various benchmark types (right).

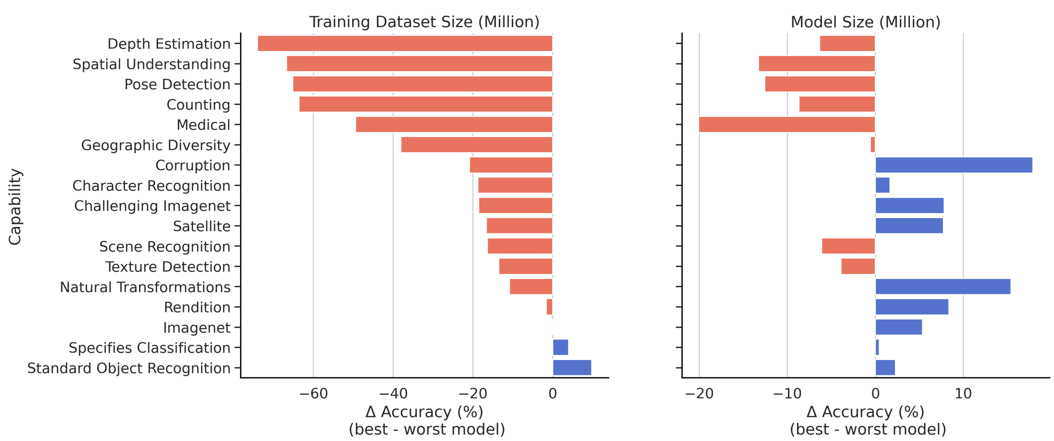

Figure 8: **Benchmark capabilities performance does not scale with dataset and model size** Median zero-shot performance of models on various benchmark capabilities. We investigate the impact of dataset size (left), and model size on various benchmark capabilities (right). We isolate the effect of training 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.4 Impact of Prompts on MNIST Performance

The MNIST benchmark, 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 9).

A.4.1 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

A.4.2 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.4.3 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.4.4 Basic Prompt

The basic prompt used:

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

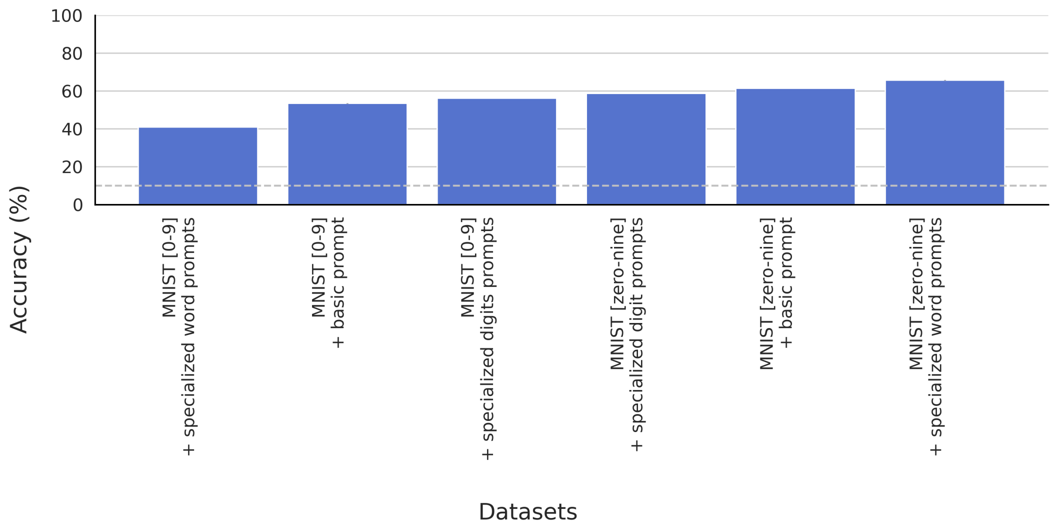

Figure 9: **Median performance of 59 VLMs on MNIST while varying prompts and labels**. Blue bars represent the median zero-shot performance of models and dashed-grey line represents the chance-level for MNIST.

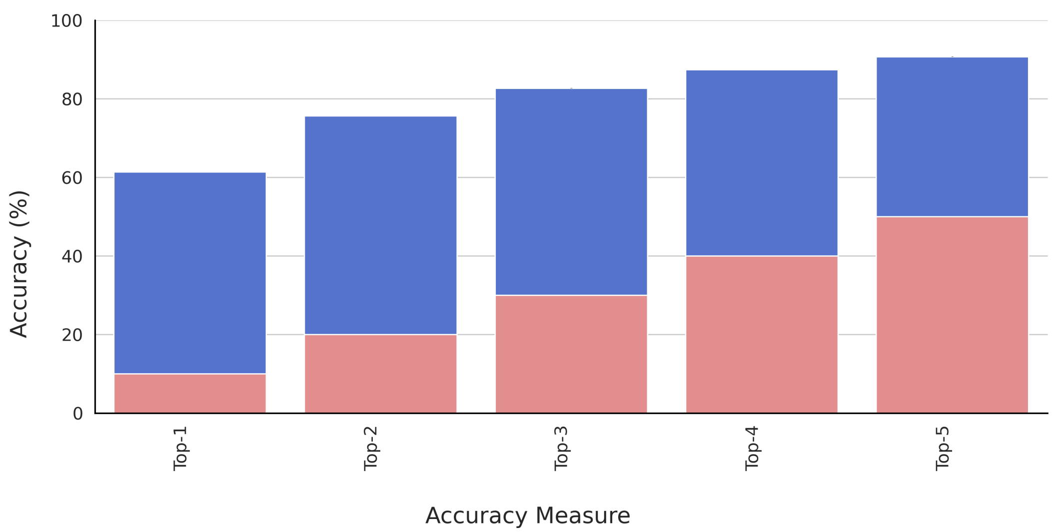

Figure 10: **Median performance of 59 VLMs on MNIST while varying accuracy measure from top-1 to top-5**. The following further shows that VLMs' performance on MNIST is not due mismatch between top-1 and top-5 guesses. Blue bars represent the median zero-shot performance of models and red bars represents the chance-level for benchmarks.

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 11 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 10 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.

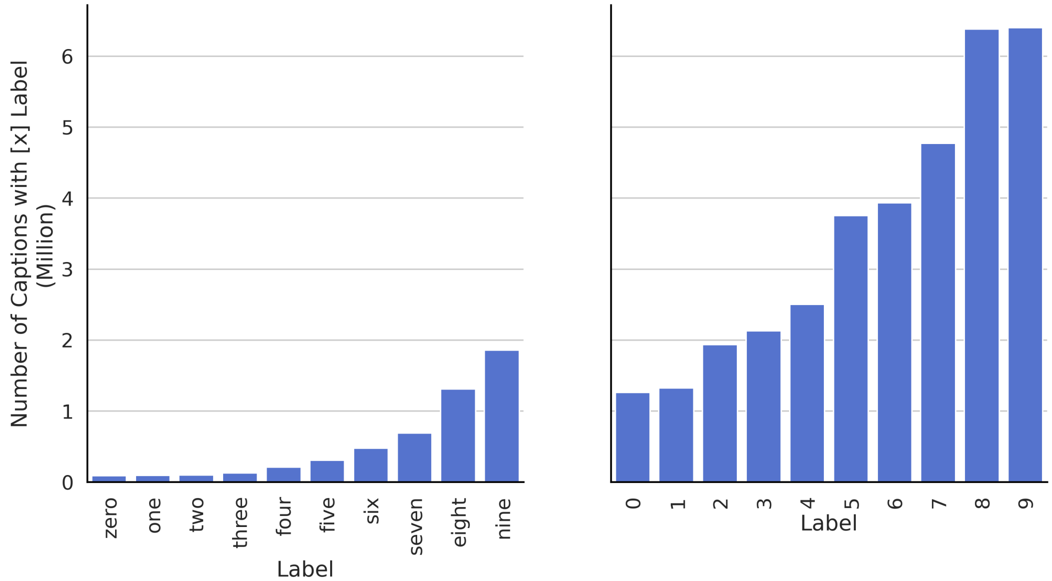

Figure 11: **Frequency of different digits in LAION-400M, showing substantial frequency of digits in visual diet of VLMs.** Left panel counts the number of words of the digits i.e. [zero-nine] and right panel counts the number of digits in LAION-400M.

A.5 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 12, 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.6 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 are based on benchmarks that correlate strongly with other benchmarks within each benchmark type and capability (Tables 3 and 4).

A.7 Weighted Average Performance

To account for the varying difficulties across tasks, we compute the weighted average performance of each model by normalizing their scores relative to the performance of CLIP B/32. We use CLIP B/32 as a baseline because its performance effectively captures the inherent complexity of each task, serving as a proxy for task difficulty.

Figure 13 illustrates the normalization results in lower overall performance scores for all models. However, it does not affect the relative rankings among them. This consistency suggests that while task difficulty impacts absolute performance metrics, the comparative effectiveness of the models remains stable across different levels of task complexity.

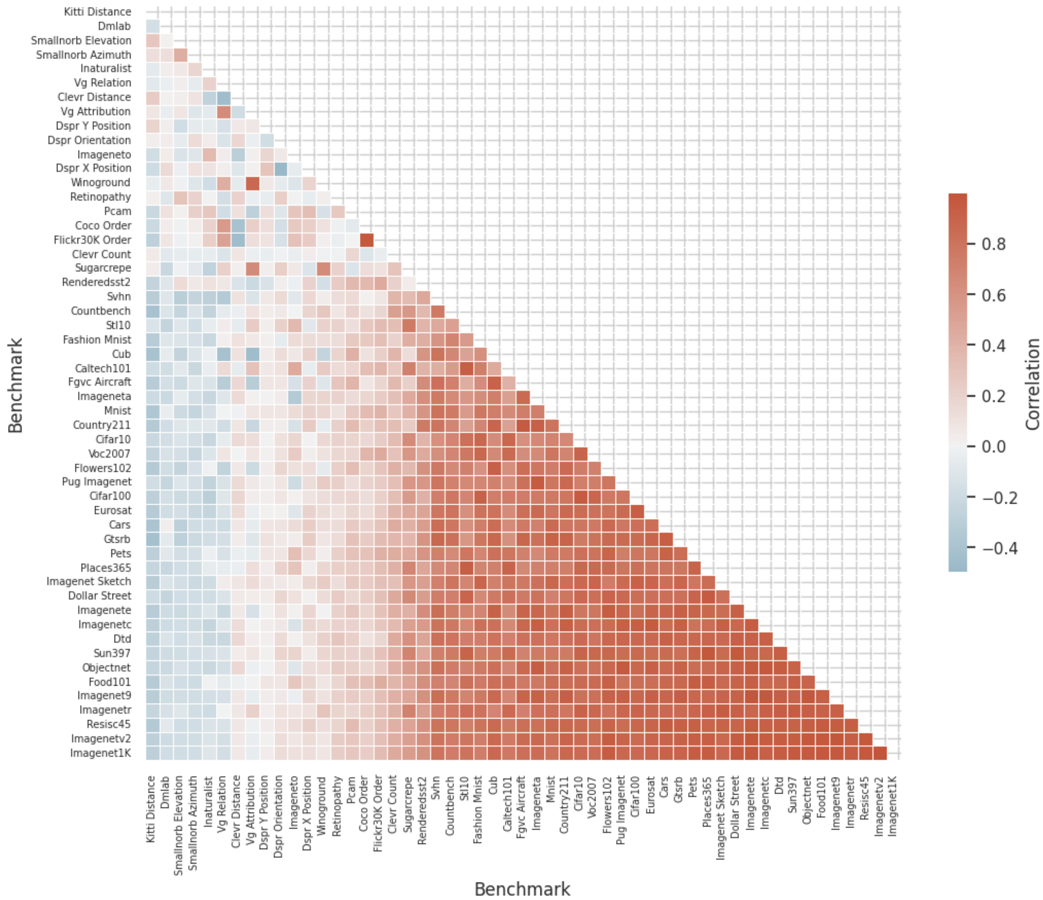

Figure 12: **Correlation matrix of models' performance across all benchmarks.**

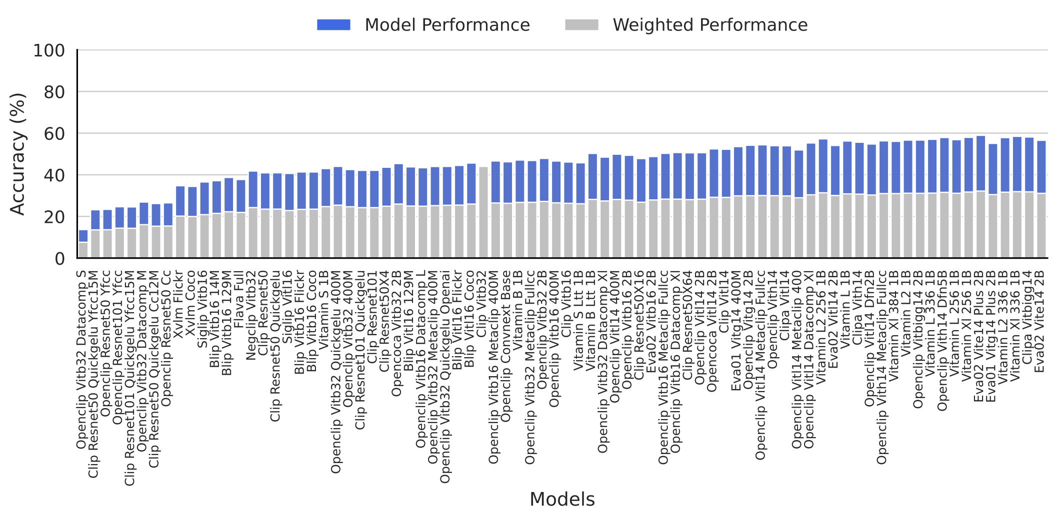

Figure 13: **Weighted Average Performance** for each model using CLIP B/32 as the baseline model performance (as a proxy for task difficulty

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 3: **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 4: **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 5: 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. [2022a]	14	86	BLIP	vit	BLIP ViT B 16
blip_vitL16_129m Li et al. [2022a]	129	307	BLIP	vit	BLIP ViT L 16
blip_vitB16_129m Li et al. [2022a]	129	86	BLIP	vit	BLIP ViT B 16
blip_vitB16_coco Li et al. [2022a]	129	86	BLIP	vit	BLIP ViT B 16
blip_vitB16_flickr Li et al. [2022a]	129	86	BLIP	vit	BLIP ViT B 16
blip_vitL16_coco Li et al. [2022a]	129	307	BLIP	vit	BLIP ViT L 16
blip_vitL16_flickr Li et al. [2022a]	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. [2023b]	1280	1843	Pure Contrastive	vit	CLIPA ViT G 14
clipa_vitH14 Li et al. [2023b]	1280	633	Pure Contrastive	vit	CLIPA ViT H 14
clipa_vitL14 Li et al. [2023b]	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. [2023b]	13	86	Pure Contrastive	vit	DataComp ViT B 32
openclip_vitB32_datacomp_m Gadre et al. [2023b]	128	86	Pure Contrastive	vit	DataComp ViT B 32
openclip_vitB32_datacomp_xl Gadre et al. [2023b]	12800	86	Pure Contrastive	vit	DataComp ViT B 32
openclip_vitB16_datacomp_xl Gadre et al. [2023b]	12800	86	Pure Contrastive	vit	DataComp ViT B 16
openclip_vitB16_datacomp_l Gadre et al. [2023b]	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. [2022a]	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. [2023b]	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. [2022a], Ilharco et al. [2021]	2000	307	Other	vit	OpenCOCA ViT L 14
opencoca_vitB32_2b Yu et al. [2022a], 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 6: 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., 2023c]	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., 2022a]	relation	relation	relations	False	False	2

Table 7: 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.