Self-supervised learning, foundation models

Vlad Shakhuro

27 November 2025

Outline

1. Intro

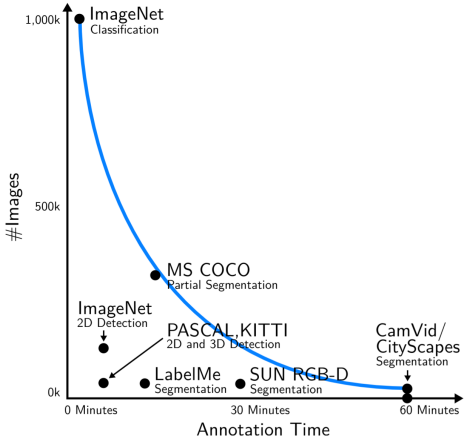
2. Pretext tasks

3. Contrastive and masked learning

4. Foundation models

5. Benchmarking foundation models

Labelling tradeoff



Labelling enough data for deep models is expensive

Labels always have errors

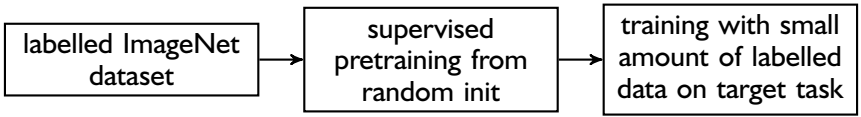
Training people to label effectively and without many errors is hard

Lots of unlabelled data exist, can we use it somehow?

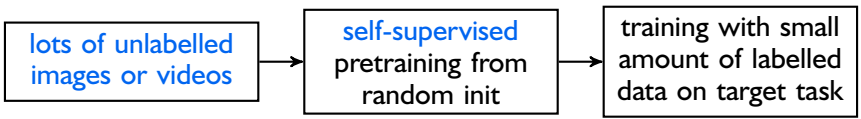
Xie et al. Semantic Instance Annotation of Street Scenes by 3D to 2D Label Transfer. CVPR 2016

Self-supervised pretraining

Usual ImageNet pretraining:



Self-supervised pretraining:



Outline

1. Intro

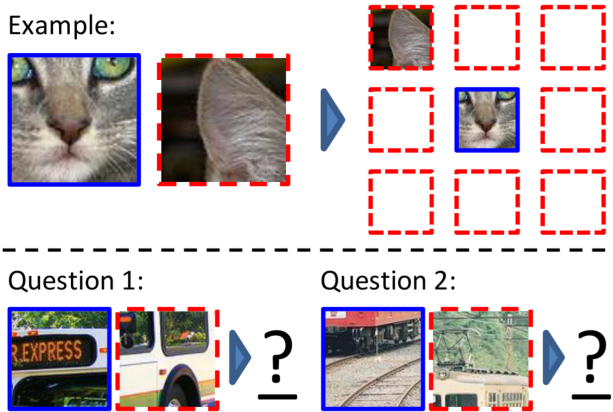
2. Pretext tasks

3. Contrastive and masked learning

4. Foundation models

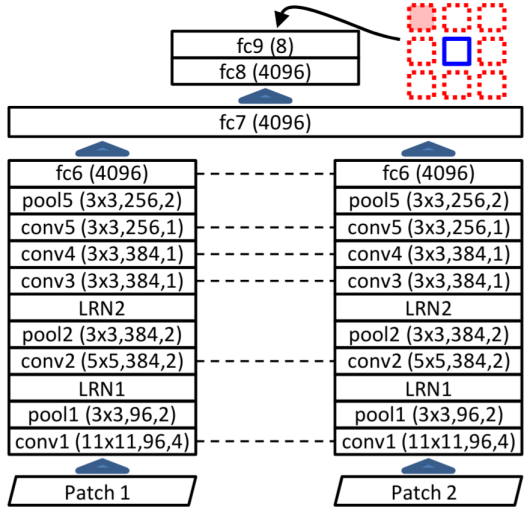
5. Benchmarking foundation models

Context prediction



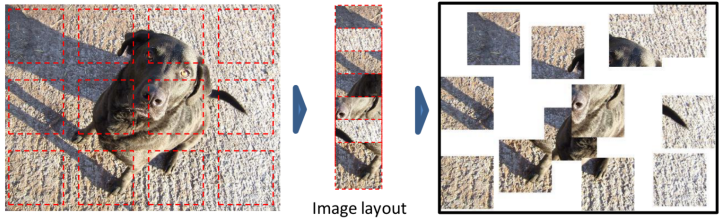
Doersch et al. Unsupervised Visual Representation Learning by Context Prediction. ICCV 2015

Context prediction



Doersch et al. Unsupervised Visual Representation Learning by Context Prediction. ICCV 2015

Context prediction



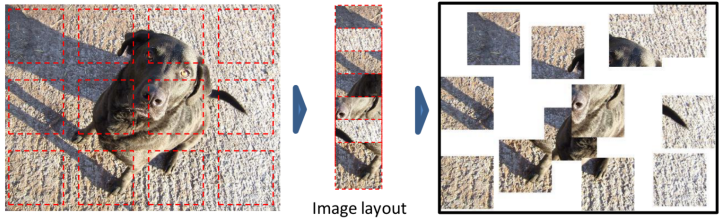
Initial layout, with sampled patches in red

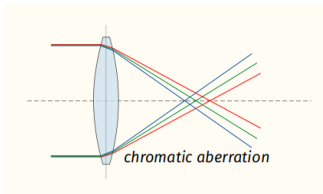
Image layout is discarded

We can recover image layout automatically

Doersch et al. Unsupervised Visual Representation Learning by Context Prediction. ICCV 2015

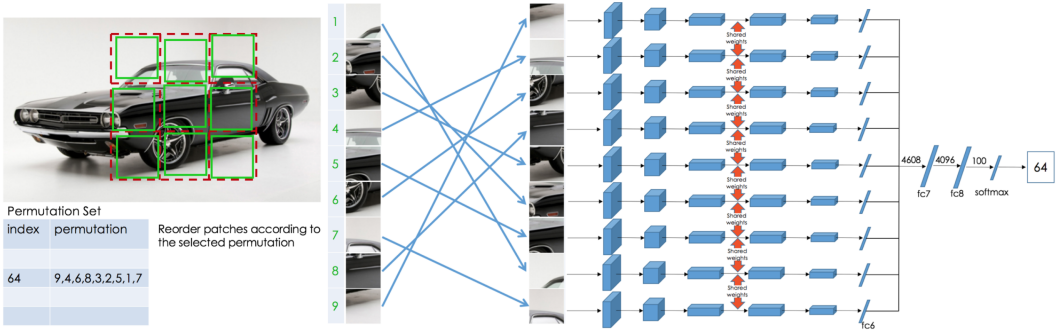
Context prediction





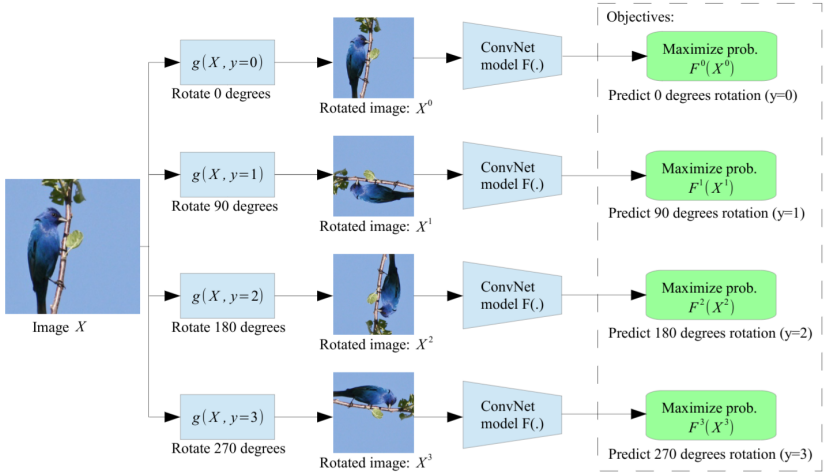
Doersch et al. Unsupervised Visual Representation Learning by Context Prediction. ICCV 2015

Jigsaw puzzle



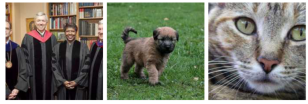
Noroozi, Favaro. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. ECCV 2016

Image rotation

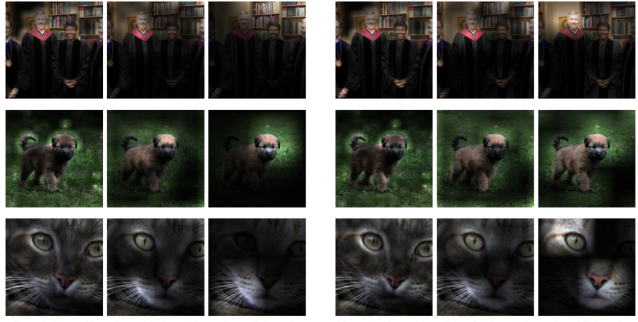


Gidaris et al. Unsupervised Representation Learning by Predicting Image Rotations. ICLR 2018

Image rotation



Input images on the models



Conv1 27×27 Conv3 13×13 Conv5 6×6 Conv1 27×27 Conv3 13×13 Conv5 6×6

(a) **Attention maps of supervised model** (b) **Attention maps of our self-supervised model**

Gidaris et al. Unsupervised Representation Learning by Predicting Image Rotations. ICLR 2018

Pretext tasks summary

Pretext tasks are constructed emprically based on some visual prior knowledge, e.g. predict rotations, spatial locations, colors of the image

The models are trained to extract semantic knowledge from images in order to solve pretext task

Usefullness of trained features is assessed on the target task

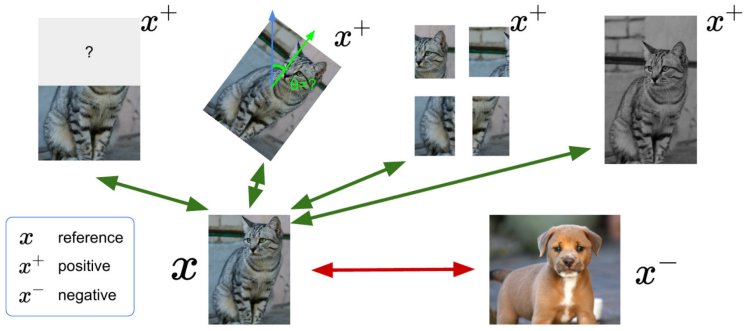
Pretext tasks are hard to come up with

It's hard to predict whether learned features will be general enough

Outline

1. Intro
2. Pretext tasks
3. Contrastive and masked learning
4. Foundation models
5. Benchmarking foundation models

Contrastive learning



Given a score function $s(\cdot, \cdot)$, we want to learn a mapping f_θ that yields high score for positive pairs (x, x^+) and low score for negative pairs (x, x^-):

$$s(f_\theta(x), f_\theta(x^+)) \gg s(f_\theta(x), f_\theta(x^-))$$

Contrastive learning

Assume that we have 1 reference (x), 1 positive (x^+) and $N-1$ negative (x^-) samples. We will use multiclass cross entropy loss function:

$$\mathscr{L} = -\mathsf{E}_{\mathscr{X}}\left[\log \frac{\exp\left(s(f_\theta(x), f_\theta(x^+))\right)}{\exp\left(s(f_\theta(x), f_\theta(x^+))\right) + \sum_{j=1}^{N-1}\exp\left(s(f_\theta(x), f_\theta(x_j^-))\right)}\right]$$

It is also known as InfoNCE loss and its negative is a lower bound on the mutual information between $f_\theta(x)$ and $f_\theta(x^+)$:

$$MI[f_\theta(x), f_\theta(x^+)] \geq \log N - \mathscr{L}$$

Maximizing mutual information between different "views" of an image forces features to capture high-level information from images

Oord et al. Representation Learning with Contrastive Predictive Coding. arXiv: 1807.03748

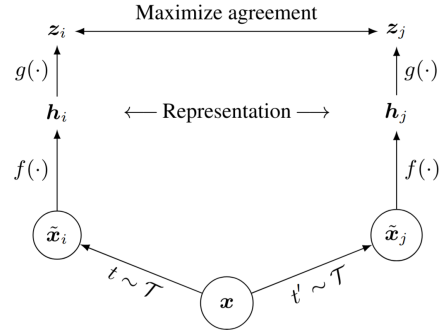
SimCLR

Cosine similarity as the score function:

$$s(\boldsymbol{z}_i, \boldsymbol{z}_j) = \frac{\boldsymbol{z}_i^T \boldsymbol{z}_j}{\|\boldsymbol{z}_i\| \|\boldsymbol{z}_j\|}$$

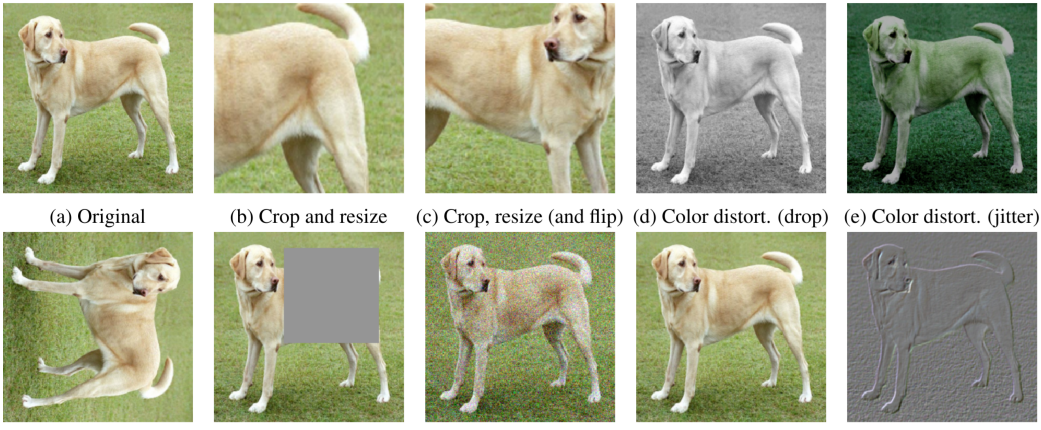
Use a projection network $g(\cdot)$ to project features to a space where contrastive learning is applied

The projection improves learning (more important information is preserved in $\boldsymbol{h}$)



Chen et al. A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020

SimCLR



(a) Original (b) Crop and resize (c) Crop, resize (and flip) (d) Color distort. (drop) (e) Color distort. (jitter)

(f) Rotate $\{90°, 180°, 270°\}$ (g) Cutout (h) Gaussian noise (i) Gaussian blur (j) Sobel filtering

Chen et al. A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020

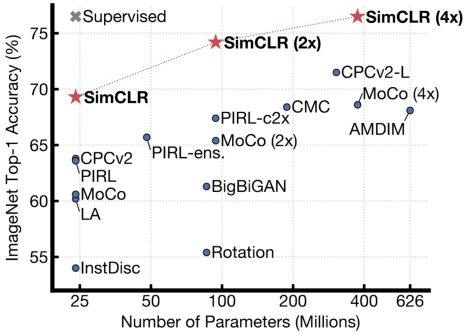
SimCLR

Algorithm 1 SimCLR's main learning algorithm.

input: batch size N, constant τ, structure of f, g, $\mathcal{T}$.
for sampled minibatch $\{\boldsymbol{x}_k\}_{k=1}^{N}$ **do**
  for all $k \in \{1, \dots, N\}$ **do**
    draw two augmentation functions $t \sim \mathcal{T}$, $t' \sim \mathcal{T}$
    # the first augmentation
    $\tilde{\boldsymbol{x}}_{2k-1} = t(\boldsymbol{x}_k)$
    $\boldsymbol{h}_{2k-1} = f(\tilde{\boldsymbol{x}}_{2k-1})$ # representation
    $\boldsymbol{z}_{2k-1} = g(\boldsymbol{h}_{2k-1})$ # projection
    # the second augmentation
    $\tilde{\boldsymbol{x}}_{2k} = t'(\boldsymbol{x}_k)$
    $\boldsymbol{h}_{2k} = f(\tilde{\boldsymbol{x}}_{2k})$ # representation
    $\boldsymbol{z}_{2k} = g(\boldsymbol{h}_{2k})$ # projection
  end for
  for all $i \in \{1, \dots, 2N\}$ and $j \in \{1, \dots, 2N\}$ **do**
    $s_{i,j} = \boldsymbol{z}_i^\top \boldsymbol{z}_j / (\|\boldsymbol{z}_i\| \|\boldsymbol{z}_j\|)$ # pairwise similarity
  end for
  define $\ell(i, j)$ **as** $\ell(i, j) = -\log \frac{\exp(s_{i,j}/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(s_{i,k}/\tau)}$
  $\mathcal{L} = \frac{1}{2N} \sum_{k=1}^{N} [\ell(2k-1, 2k) + \ell(2k, 2k-1)]$
  update networks f and g to minimize $\mathcal{L}$
end for
return encoder network $f(\cdot)$, and throw away $g(\cdot)$

Chen et al. A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020

SimCLR

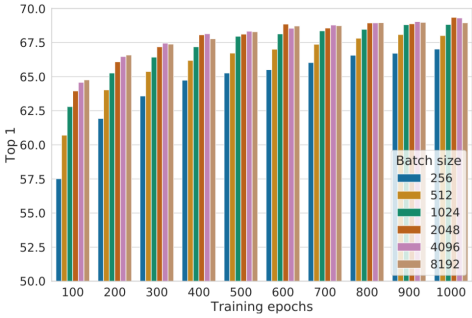


Method	Architecture	Label fraction 1%	Label fraction 10%
		Top 5	
Supervised baseline	ResNet-50	48.4	80.4
Methods using other label-propagation:			
Pseudo-label	ResNet-50	51.6	82.4
VAT+Entropy Min.	ResNet-50	47.0	83.4
UDA (w. RandAug)	ResNet-50	-	88.5
FixMatch (w. RandAug)	ResNet-50	-	89.1
S4L (Rot+VAT+En. M.)	ResNet-50 (4×)	-	91.2
Methods using representation learning only:			
InstDisc	ResNet-50	39.2	77.4
BigBiGAN	RevNet-50 (4×)	55.2	78.8
PIRL	ResNet-50	57.2	83.8
CPC v2	ResNet-161(∗)	77.9	91.2
SimCLR (ours)	ResNet-50	75.5	87.8
SimCLR (ours)	ResNet-50 (2×)	83.0	91.2
SimCLR (ours)	ResNet-50 (4×)	**85.8**	**92.6**

Table 7. ImageNet accuracy of models trained with few labels.

Chen et al. A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020

SimCLR

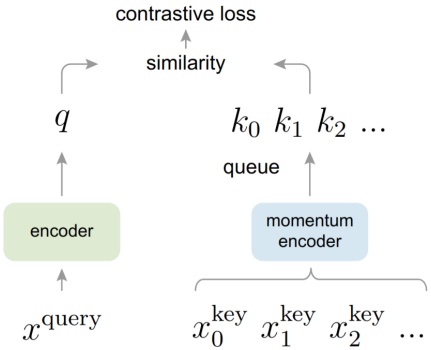


SimCLR requires training with large batch size. That is possible only with distributed training

SimCLR v2: larger models, more layers in projection head

Chen et al. A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020

MoCo



$$\mathcal{L}_q = -\log \frac{\exp(q \cdot k^+/\tau)}{\exp(q \cdot k^+/\tau) + \sum_{k^-} \exp(q \cdot k^-/\tau)}$$

Encoder is updated with large momentum ($m = 0.999$)

Batch is shuffled before passing to f_q and f_k to prevent cheating via BatchNorm

He et al. Momentum contrast for unsupervised visual representation learning. CVPR 2020

MoCo v2

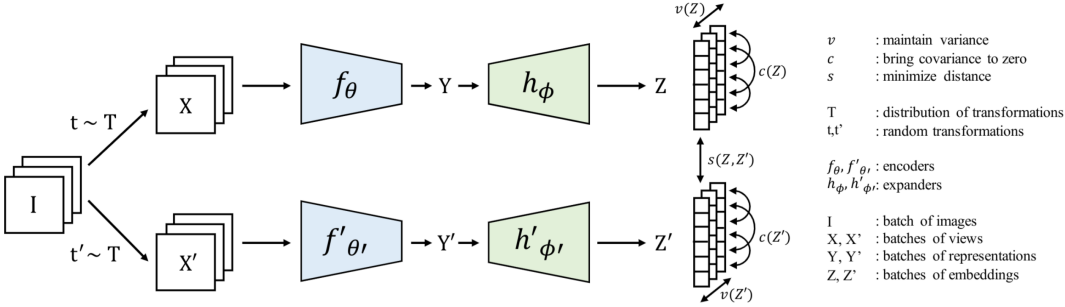
case	unsup. pre-train MLP	aug+	cos	epochs	batch	ImageNet acc.
MoCo v1 [6]				200	256	60.6
SimCLR [2]	✓	✓	✓	200	256	61.9
SimCLR [2]	✓	✓	✓	200	8192	66.6
MoCo v2	✓	✓	✓	200	256	**67.5**
*results of **longer** unsupervised training follow:*						
SimCLR [2]	✓	✓	✓	1000	4096	69.3
MoCo v2	✓	✓	✓	800	256	**71.1**

Non-linear projection head and strong augmentation are critical for good quality

MoCo outperforms SimCLR with much smaller batches
(it's possible to train MoCo v2 on a node with 8×V100 GPUs)

Chen et al. Improved Baselines with Momentum Contrastive Learning. arXiv:2003.04297

VICReg



$$v(Z) = \frac{1}{d}\sum_{i=1}^{d} \max\left(0, 1 - \sqrt{\mathrm{Var}(z^i) + \varepsilon}\right)$$

$$c(Z) = \frac{1}{d}\sum_{i \neq j} [C(Z)]_{ij}^2, \quad C(Z) = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})(z_i - \bar{z})^T$$

$$s(Z, Z') = MSE(Z, Z')$$

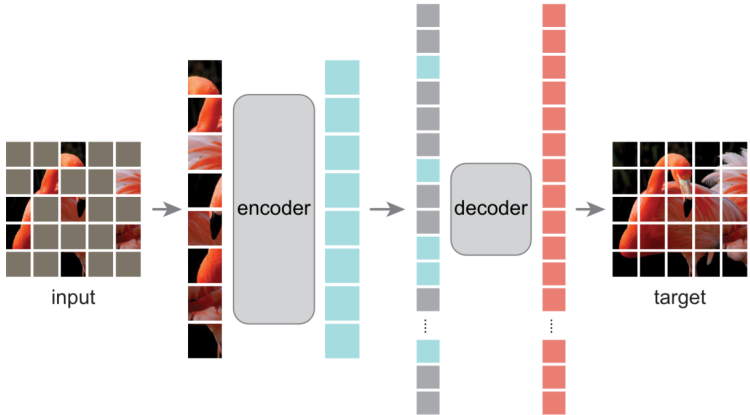
Bardes et al. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. ICLR 2022

VICReg results

Method	Linear		Semi-supervised			
	Top-1	Top-5	Top-1		Top-5	
			1%	10%	1%	10%
Supervised	76.5	-	25.4	56.4	48.4	80.4
MoCo He et al. (2020)	60.6	-	-	-	-	-
PIRL Misra & Maaten (2020)	63.6	-	-	-	57.2	83.8
CPC v2 Hénaff et al. (2019)	63.8	-	-	-	-	-
CMC Tian et al. (2019)	66.2	-	-	-	-	-
SimCLR Chen et al. (2020a)	69.3	89.0	48.3	65.6	75.5	87.8
MoCo v2 Chen et al. (2020c)	71.1	-	-	-	-	-
SimSiam Chen & He (2020)	71.3	-	-	-	-	-
SwAV Caron et al. (2020)	71.8	-	-	-	-	-
InfoMin Aug Tian et al. (2020)	73.0	91.1	-	-	-	-
OBoW Gidaris et al. (2021)	73.8	-	-	-	82.9	90.7
BYOL Grill et al. (2020)	74.3	91.6	53.2	68.8	78.4	89.0
SwAV (w/ multi-crop) Caron et al. (2020)	75.3	-	53.9	70.2	78.5	89.9
Barlow Twins Zbontar et al. (2021)	73.2	91.0	55.0	69.7	79.2	89.3
VICReg (ours)	73.2	91.1	54.8	69.5	79.4	89.5

Training is done on 32×V100 GPUs

Masked Autoencoders

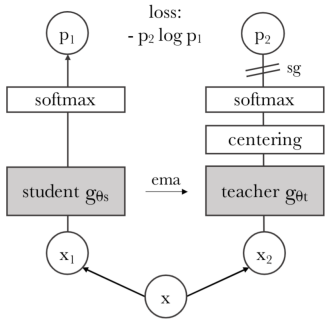


He et al. Masked autoencoders are scalable vision learners. CVPR 2022

Masked Autoencoders

method	pre-train data	ViT-B	ViT-L	ViT-H	ViT-H_{448}
scratch, our impl.	-	82.3	82.6	83.1	-
DINO [5]	IN1K	82.8	-	-	-
MoCo v3 [9]	IN1K	83.2	84.1	-	-
BEiT [2]	IN1K+DALLE	83.2	85.2	-	-
MAE	IN1K	83.6	85.9	86.9	**87.8**

DINO



	Random	Supervised	DINO
ViT-S/16	22.0	27.3	45.9
ViT-S/8	21.8	23.7	44.7

Caron et al. Emerging Properties in Self-Supervised Vision Transformers. ICCV 2021

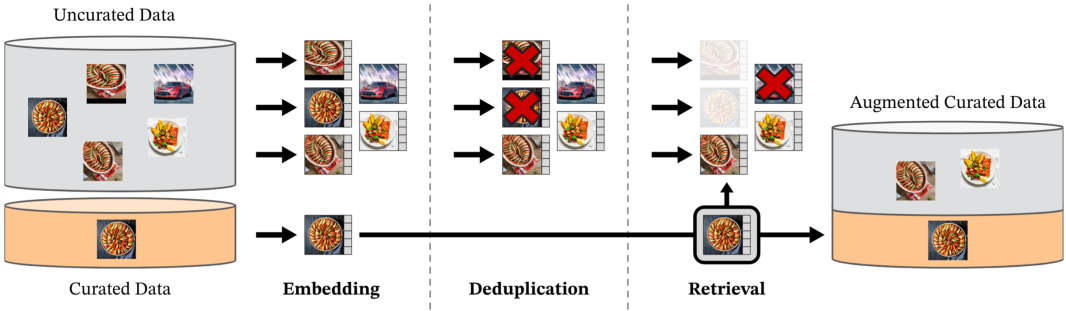
DINO results

Method	Arch.	Param.	im/s	Linear	k-NN
Supervised	RN50	23	1237	79.3	79.3
SCLR [12]	RN50	23	1237	69.1	60.7
MoCov2 [15]	RN50	23	1237	71.1	61.9
InfoMin [67]	RN50	23	1237	73.0	65.3
BarlowT [81]	RN50	23	1237	73.2	66.0
OBoW [27]	RN50	23	1237	73.8	61.9
BYOL [30]	RN50	23	1237	74.4	64.8
DCv2 [10]	RN50	23	1237	75.2	67.1
SwAV [10]	RN50	23	1237	**75.3**	65.7
DINO	RN50	23	1237	**75.3**	**67.5**
Supervised	ViT-S	21	1007	79.8	79.8
BYOL* [30]	ViT-S	21	1007	71.4	66.6
MoCov2* [15]	ViT-S	21	1007	72.7	64.4
SwAV* [10]	ViT-S	21	1007	73.5	66.3
DINO	ViT-S	21	1007	**77.0**	**74.5**
Comparison across architectures					
SCLR [12]	RN50w4	375	117	76.8	69.3
SwAV [10]	RN50w2	93	384	77.3	67.3
BYOL [30]	RN50w2	93	384	77.4	–
DINO	ViT-B/16	85	312	78.2	76.1
SwAV [10]	RN50w5	586	76	78.5	67.1
BYOL [30]	RN50w4	375	117	78.6	–
BYOL [30]	RN200w2	250	123	79.6	73.9
DINO	ViT-S/8	21	180	79.7	**78.3**
SCLRv2 [13]	RN152w3+SK	794	46	79.8	73.1
DINO	ViT-B/8	85	63	**80.1**	77.4

Training is done on 8×V100 GPUs

DINOv2

Data + large model (ViT-g, 1.1B params) distillation + several loss functions and regularizers + effective implementation. Training code and weights are open sourced under commercial license

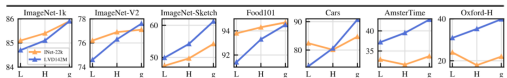


Oquab et al. DINOv2: Learning Robust Visual Features without Supervision. arXiv:2304.07193

DINOv2

Task	Dataset / Split	Images	Retrieval	Retrieved	Final
classification	ImageNet-22k / –	14,197,086	as is	–	14,197,086
classification	ImageNet-22k / –	14,197,086	sample	56,788,344	56,788,344
classification	ImageNet-1k / train	1,281,167	sample	40,997,344	40,997,344
fine-grained classif.	Caltech 101 / train	3,030	cluster	2,630,000	1,000,000
fine-grained classif.	CUB-200-2011 / train	5,994	cluster	1,300,000	1,000,000
fine-grained classif.	DTD / train1	1,880	cluster	1,580,000	1,000,000
fine-grained classif.	FGVC-Aircraft / train	3,334	cluster	1,170,000	1,000,000
fine-grained classif.	Flowers-102 / train	1,020	cluster	1,060,000	1,000,000
fine-grained classif.	Food-101 / train	75,750	cluster	21,670,000	1,000,000
fine-grained classif.	Oxford-IIIT Pet / trainval	3,680	cluster	2,750,000	1,000,000
fine-grained classif.	Stanford Cars / train	8,144	cluster	7,220,000	1,000,000
fine-grained classif.	SUN397 / train1	19,850	cluster	18,950,000	1,000,000
fine-grained classif.	Pascal VOC 2007 / train	2,501	cluster	1,010,000	1,000,000
segmentation	ADE20K / train	20,210	cluster	20,720,000	1,000,000
segmentation	Cityscapes / train	2,975	cluster	1,390,000	1,000,000
segmentation	Pascal VOC 2012 (seg.) / trainaug	1,464	cluster	10,140,000	1,000,000
depth estimation	Mapillary SLS / train	1,434,262	as is	–	1,434,262
depth estimation	KITTI / train (Eigen)	23,158	cluster	3,700,000	1,000,000
depth estimation	NYU Depth V2 / train	24,231	cluster	10,850,000	1,000,000
depth estimation	SUN RGB-D / train	4,829	cluster	4,870,000	1,000,000
retrieval	Google Landmarks v2 / train (clean)	1,580,470	as is	–	1,580,470
retrieval	Google Landmarks v2 / train (clean)	1,580,470	sample	6,321,880	6,321,880
retrieval	AmsterTime / new	1,231	cluster	960,000	960,000
retrieval	AmsterTime / old	1,231	cluster	830,000	830,000
retrieval	Met / train	397,121	cluster	62,860,000	1,000,000
retrieval	Revisiting Oxford / base	4,993	cluster	3,680,000	1,000,000
retrieval	Revisiting Paris / base	6,322	cluster	3,660,000	1,000,000
					142,109,386

Training Data	INet-1k	Im-A	ADE-20k	Oxford-M
INet-22k	85.9	73.5	46.6	62.5
INet-22k \ INet-1k	85.3	70.3	46.2	58.7
Uncurated data	83.3	59.4	48.5	54.3
LVD-142M	85.8	73.9	47.7	64.6



DINOv2 results

Method	Arch.	Data	Text sup.	kNN	linear		
				val	val	ReaL	V2
Weakly supervised							
CLIP	ViT-L/14	WIT-400M	✓	79.8	84.3	88.1	75.3
CLIP	ViT-L/14_{336}	WIT-400M	✓	80.5	85.3	88.8	75.8
SWAG	ViT-H/14	IG3.6B	✓	82.6	85.7	88.7	77.6
OpenCLIP	ViT-H/14	LAION	✓	81.7	84.4	88.4	75.5
OpenCLIP	ViT-G/14	LAION	✓	83.2	86.2	89.4	77.2
EVA-CLIP	ViT-g/14	custom*	✓	**83.5**	86.4	89.3	77.4
Self-supervised							
MAE	ViT-H/14	INet-1k	×	49.4	76.6	83.3	64.8
DINO	ViT-S/8	INet-1k	×	78.6	79.2	85.5	68.2
SEERv2	RG10B	IG2B	×	–	79.8	–	–
MSN	ViT-L/7	INet-1k	×	79.2	80.7	86.0	69.7
EsViT	Swin-B/W=14	INet-1k	×	79.4	81.3	87.0	70.4
Mugs	ViT-L/16	INet-1k	×	80.2	82.1	86.9	70.8
iBOT	ViT-L/16	INet-22k	×	72.9	82.3	87.5	72.4
DINOv2	ViT-S/14	LVD-142M	×	79.0	81.1	86.6	70.9
	ViT-B/14	LVD-142M	×	82.1	84.5	88.3	75.1
	ViT-L/14	LVD-142M	×	**83.5**	86.3	89.5	78.0
	ViT-g/14	LVD-142M	×	**83.5**	**86.5**	**89.6**	**78.4**

Outline

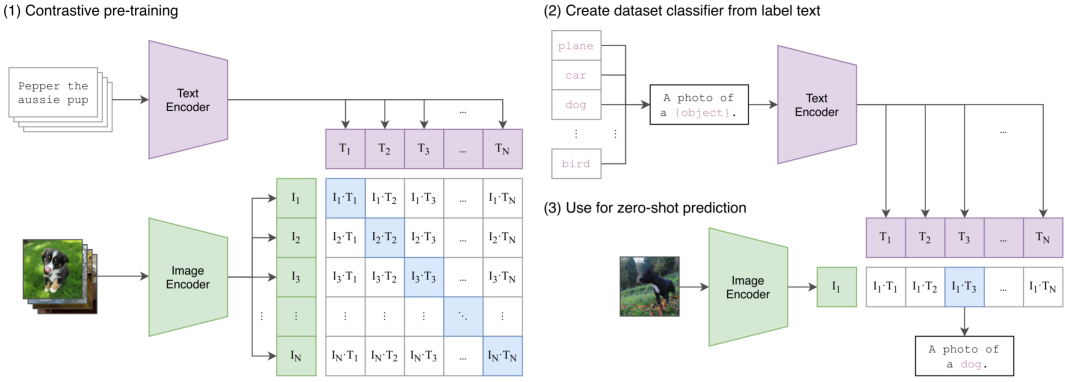
1. Intro
2. Pretext tasks
3. Contrastive and masked learning
4. Foundation models
5. Benchmarking foundation models

What is a foundation model?

A model may be called a foundation model if it:

- works with multiple modalities, i.e. text and images
- solves several tasks in different domains, i.e. classification, segmentation, captioning, question answering
- promptable, i.e. supports several types of queries regarding analyzed information
- works well without finetuning

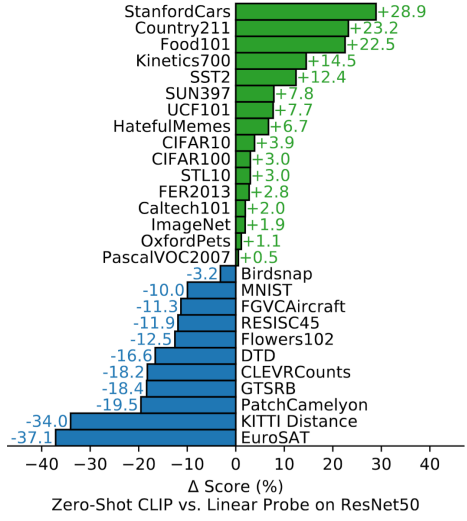
CLIP



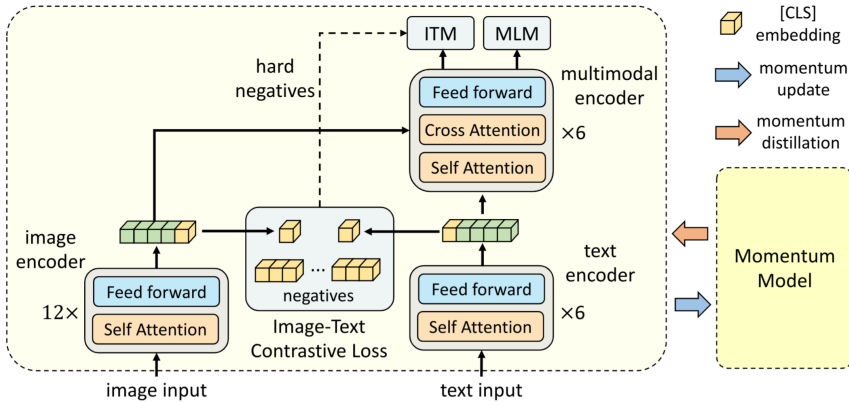
400M (image, text) pairs, 500×V100 GPUs for pretraining

Radford et al. Learning transferable visual models from natural language supervision. ICML 2021

CLIP zero-shot results



ALBEF



Loss: contrastive loss + image-text matching + masked language modelling

Li et al. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. NeurIPS 2021

ALBEF pseudo-targets

masked language modelling ↓

"polar bear in the [MASK]"

GT: wild
Top-5 pseudo-targets:
1. zoo
2. pool
3. water
4. pond
5. wild

"a man [MASK] along a road in front of nature in summer"

GT: standing
Top-5 pseudo-targets:
1. walks
2. walking
3. runs
4. running
5. goes

"a [MASK] waterfall in the deep woods"

GT: remote
Top-5 pseudo-targets:
1. small
2. beautiful
3. little
4. secret
5. secluded

GT: breakdown of the car on the road
Top-5 pseudo-targets:
1. young woman get out of the car near the road
2. a woman inspects her damaged car under a tree
3. a woman looking into a car after locking her keys inside
4. young woman with a broken car calling for help
5. breakdown of the car on the road

GT: the harbor a small village
Top-5 pseudo-targets:
1. the harbour with boats and houses
2. replica of the sailing ship in the harbour
3. ships in the harbor of the town
4. the harbor a small village
5. boats lined up alongside the geographical feature category in the village

image-text matching ↑

ALBEF comparison with CLIP

Method	# Pre-train Images	Flickr30K (1K test set)					
		TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10
UNITER [2]	4M	83.6	95.7	97.7	68.7	89.2	93.9
CLIP [6]	400M	88.0	98.7	99.4	68.7	90.6	95.2
ALIGN [7]	1.2B	88.6	98.7	99.7	75.7	93.8	96.8
ALBEF	4M	90.5	98.8	99.7	76.8	93.7	96.7
ALBEF	14M	**94.1**	**99.5**	**99.7**	**82.8**	**96.3**	**98.1**

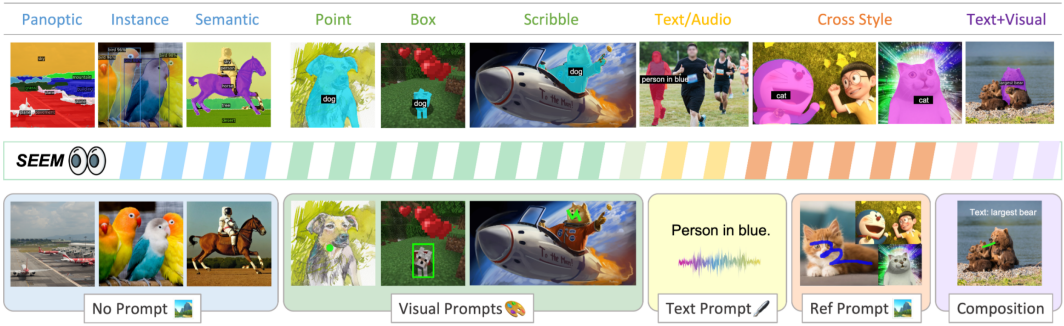
Table 3: Zero-shot image-text retrieval results on Flickr30K.

Method	VQA		$NLVR^2$		SNLI-VE	
	test-dev	test-std	dev	test-P	val	test
VisualBERT [13]	70.80	71.00	67.40	67.00	-	-
VL-BERT [10]	71.16	-	-	-	-	-
LXMERT [1]	72.42	72.54	74.90	74.50	-	-
12-in-1 [12]	73.15	-	-	78.87	-	76.95
UNITER [2]	72.70	72.91	77.18	77.85	78.59	78.28
VL-BART/T5 [54]	-	71.3	-	73.6	-	-
ViLT [21]	70.94	-	75.24	76.21	-	-
OSCAR [3]	73.16	73.44	78.07	78.36	-	-
VILLA [8]	73.59	73.67	78.39	79.30	79.47	79.03
ALBEF (4M)	74.54	74.70	80.24	80.50	80.14	80.30
ALBEF (14M)	**75.84**	**76.04**	**82.55**	**83.14**	**80.80**	**80.91**

Table 4: Comparison with state-of-the-art methods on downstream vision-language tasks.

8×A100 GPUs for pretraining

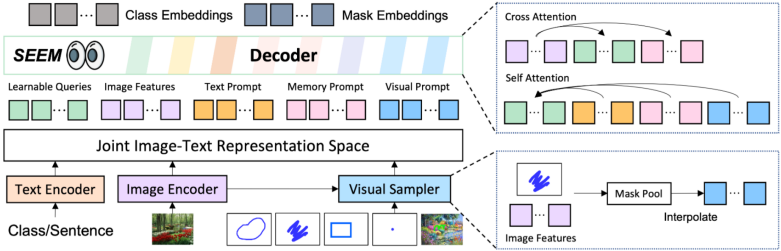
SEEM



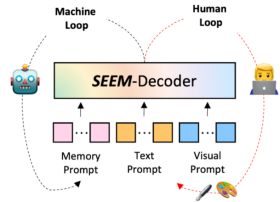
Panoptic + referring + interactive segmentation

Zou et al. Segment Everything Everywhere All at Once. NeurIPS 2023

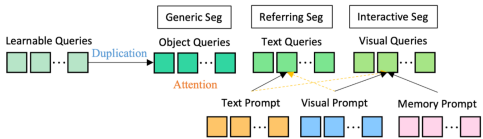
SEEM architecture



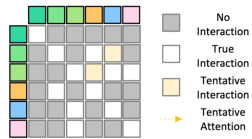
(a) Model Architecture



(b) Human-Model Interaction



(a) Queries and Prompt Interaction



(b) Self-Attention Mask

SEEM results

Method	Segmentation Data	Type	Generic Segmentation COCO			Referring Segmentation RefCOCOg			Interactive Segmentation PascalVOC					
			PQ	mAP	mIoU	cIoU	mIoU	AP50	5-NoC85	10-NoC85	20-NoC85	5-NoC90	10-NoC90	20-NoC90
Mask2Former (T) [6]	COCO (0.12M)		53.2	43.3	63.2	-	-	-	-	-	-	-	-	-
Mask2Former (B) [6]	COCO (0.12M)		56.4	46.3	67.1	-	-	-	-	-	-	-	-	-
Mask2Former (L) [6]	COCO (0.12M)		57.8	48.6	67.4	-	-	-	-	-	-	-	-	-
Pano/SegFormer (B) [45]	COCO (0.12M)	Segmentation	55.4	*	*	-	-	-	-	-	-	-	-	-
LAVT (B) [53]	Ref-COCO (0.03M)		-	-	-	61.2	*	*	-	-	-	-	-	-
PolyFormer (B) [17]	Ref-COCO+VG+... (0.16M)		-	-	-	69.3	*	*	-	-	-	-	-	-
PolyFormer (L) [17]	Ref-COCO+VG+... (0.16M)		-	-	-	71.1	*	*	-	-	-	-	-	-
RITM (<T) [18]	COCO+LVIS (0.12M)		-	-	-	-	-	-	*	*	2.19	*	*	2.57
PseudoClick (<T) [54]	COCO (0.12M)		-	-	-	-	-	-	*	*	1.94	*	*	2.25
FocalClick (T) [21]	COCO (0.12M)		-	-	-	-	-	-	*	*	2.97	*	*	3.52
FocalClick (B) [21]	COCO (0.12M)	Interactive	-	-	-	-	-	-	*	*	2.46	*	*	2.88
SimpleClick (B) [20]	COCO+LVIS (0.12M)		-	-	-	-	-	-	1.75	1.93	2.06	1.94	2.19	2.38
SimpleClick (L) [20]	COCO+LVIS (0.12M)		-	-	-	-	-	-	1.52	1.64	1.72	1.67	1.84	1.96
SimpleClick (H) [20]	COCO+LVIS (0.12M)		-	-	-	-	-	-	1.51	1.64	1.76	1.64	1.83	1.98
UViM (L) [55]	COCO (0.12M)		45.8	*	*	-	-	-	-	-	-	-	-	-
Pix2Seq v2 (B) [56]	COCO (0.12M)		-	38.2	-	-	-	-	-	-	-	-	-	-
X-Decoder (T) [11]	COCO (0.12M)		52.6	41.3	62.4	59.8	*	*	-	-	-	-	-	-
X-Decoder (B) [11]	COCO (0.12M)		56.2	45.8	66.0	64.5	*	*	-	-	-	-	-	-
X-Decoder (L) [11]	COCO (0.12M)		56.9	46.7	67.5	64.6	*	*	-	-	-	-	-	-
UNINEXT (T) [48]	Image+Video (3M)		-	44.9	-	70.0	*	*	-	-	-	-	-	-
UNINEXT (L) [48]	Image+Video (3M)		-	49.6	-	73.4	*	*	-	-	-	-	-	-
Painter (L) [57]	COCO+ADE+NYUv2 (0.16M)	Generalist	43.4	*	*	-	-	-	-	-	-	-	-	-
#SegGPT (L) [50]	COCO+ADE+NYUv2 (0.16M)		34.4	*	*	-	-	-	-	-	-	-	-	-
#SAM (B) [36]	SAM (11M)		-	-	-	-	-	-	2.47	2.65	3.28	2.23	3.13	4.12
#SAM (L) [36]	SAM (11M)		-	-	-	-	-	-	1.85	2.15	2.60	2.01	2.46	3.12
#SAM (H) [36]	SAM (11M)		-	-	-	-	-	-	1.82	2.13	2.55	1.98	2.43	3.11
SEEM (T)	COCO+LVIS (0.12M)		50.8	39.7	62.2	60.9	65.7	74.8	1.72	2.30	3.37	1.97	2.83	4.41
SEEM (B)	COCO+LVIS (0.12M)		56.1	46.4	66.3	65.0	69.6	78.2	1.56	2.04	2.93	1.77	2.47	3.79
SEEM (L)	COCO+LVIS (0.12M)		57.5	47.7	67.6	65.6	70.3	78.9	1.51	1.95	2.77	1.71	2.36	3.61
SEEM (T)	COCO+LVIS (0.12M)		-	-	-	70.4	71.7	82.1	1.72	2.28	3.32	1.97	2.82	4.37
SEEM (B)	COCO+LVIS (0.12M)	Composition	-	-	-	76.2	77.8	87.8	1.56	2.03	2.91	1.77	2.46	3.76
SEEM (L)	COCO+LVIS (0.12M)		-	-	-	75.1	76.9	86.8	1.52	1.97	2.81	1.72	2.38	3.64

Florence 2

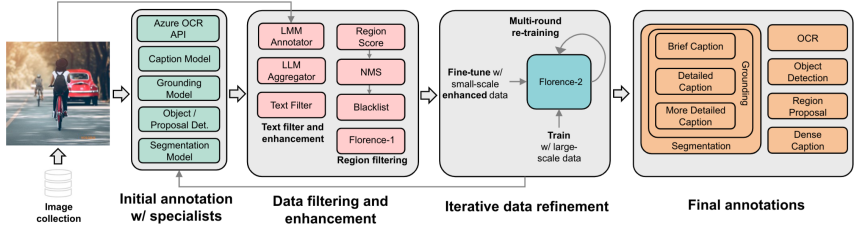
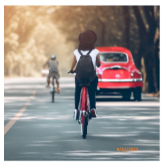


Figure 3. ***Florence-2* data engine** consists of three essential phrases: (1) initial annotation employing specialist models, (2) data filtering to correct errors and remove irrelevant annotations, and (3) an iterative process for data refinement. Our final dataset (***FLD-5B***) of over **5B** annotations contains **126M** images, **500M** text annotations, **1.3B** region-text annotations, and **3.6B** text-phrase-region annotations.

Dataset	Rep. Model	#Images	#Annotations	Spatial hierarchy	Semantics granularity
JFT300M [21]	ViT	300M	300M	Image-level	Coarse
WIT [64]	CLIP	400M	400M	Image-level	Coarse
SA-1B [32]	SAM	11M	1B	Region-level	Non-semantic
GrIT [60]	Kosmos-2	91M	137M	Image & Region-level	Fine-grained
M3W [2]	Flamingo	185M	43.3M*	Multi-image-level	Fine-grained
FLD-5B (ours)	*Florence-2* (ours)	126M	5B	Image & Region-level	Coarse to fine-grained

Xiao et al. Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks. arXiv:2311.06242

Florence 2



person, car, road

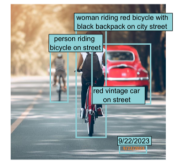
A woman riding a bike down a street next to a red car.

The image shows a person riding a red bicycle on a road with a red car in the background. The person is wearing a white t-shirt, black pants, and a black hat. She has a backpack on her back and is pedaling with their feet on the pedals. The road is lined with trees on both sides and there is another person riding another bicycle in front of her. The date "9/22/2023" is visible in the bottom right corner of the image.

Image level

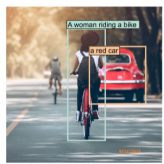
Less granular (image level) — More granular (image level)

Text annotations

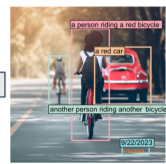


None semantic — Rich semantic

Region-text pairs annotations



A woman riding a bikedown a street next to a red car.



The image shows a person riding a red bicycle on a road with a red car in the background. The person is wearing a white t-shirt, black pants, and a black hat. She has a backpack on her back and is pedaling with their feet on the pedals. The road is lined with trees on both sides and there is another person riding another bicycle in front of her. The date '9/22/2023' is visible in the bottom right corner of the image.

Region level

Less granular (region level) — More granular (region level)

Text-phrase-region annotations

Florence 2

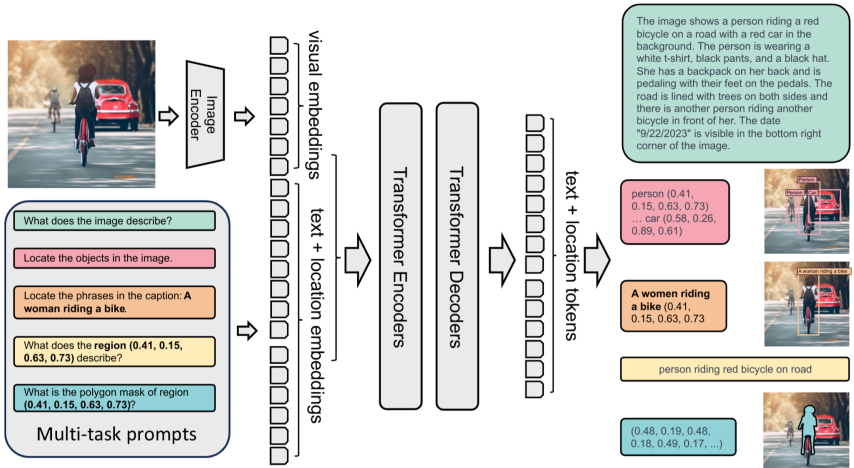
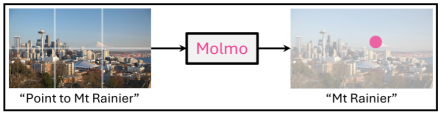
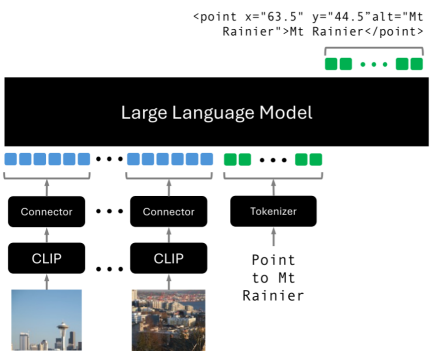


Image Encoder
visual embeddings
text + location embeddings
What does the image describe?
Locate the objects in the image.
Locate the phrases in the caption: A woman riding a bike.
What does the region (0.41, 0.15, 0.63, 0.73) describe?
What is the polygon mask of region (0.41, 0.15, 0.63, 0.73)?
Multi-task prompts
Transformer Encoders
Transformer Decoders
text + location tokens
The image shows a person riding a red bicycle on a road with a red car in the background. The person is wearing a white t-shirt, black pants, and a black hat. She has a backpack on her back and is pedaling with their feet on the pedals. The road is lined with trees on both sides and there is another person riding another bicycle in front of her. The date "9/22/2023" is visible in the bottom right corner of the image.
person (0.41, 0.15, 0.63, 0.73) … car (0.58, 0.26, 0.89, 0.61)
A women riding a bike (0.41, 0.15, 0.63, 0.73
person riding red bicycle on road
(0.48, 0.19, 0.48, 0.18, 0.49, 0.17, ...)

Molmo





OpenAI ViT-L/14 336px CLIP model
Various LLMs

Training:

1. Pretrain on PixMo
2. Finetune on academic datasets

Deitke et al. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Multimodal Models. arXiv:2409.17146

PixMo (Pixels for Molmo)

1. **PixMo-Cap for pretraining:**
   3 labellers speak for 60 seconds → transcribe → improve with LLM → summarize with LLM; 712k images, 1.3M captions
2. **PixMo-AskModelAnything:**
   labellers use language-only LLMs to semi-automatically generate question; 73k images, 162k question-answer pairs
3. **PixMo-Points:**
   428k images, 2.3M question-point pairs
   Augment prev dataset with points, 29k images and 79k question-answer pairs
4. **PixMo-CapQA, PixMo-Docs, PixMo-Clocks:** generated using an LLM

Openness

		VLM		LLM Backbone		Vision Encoder	
Category	Model	Open Weights	Open Data + Code	Open Weights	Open Data + Code	Open Weights	Open Data + Code
Molmo	Molmo-72B	Open	Open	Open	Closed	Open	Closed
	Molmo-7B-D	Open	Open	Open	Closed	Open	Closed
	Molmo-7B-O	Open	Open	Open	Open	Open	Closed
	MolmoE-1B	Open	Open	Open	Open	Open	Closed
API Models	GPT-4o	Closed	Closed	Closed	Closed	Closed	Closed
	GPT-4V	Closed	Closed	Closed	Closed	Closed	Closed
	Gemini 1.5 Pro	Closed	Closed	Closed	Closed	Closed	Closed
	Gemini 1.5 Flash	Closed	Closed	Closed	Closed	Closed	Closed
	Claude 3.5 Sonnet	Closed	Closed	Closed	Closed	Closed	Closed
	Claude 3 Opus	Closed	Closed	Closed	Closed	Closed	Closed
	Claude 3 Haiku	Closed	Closed	Closed	Closed	Closed	Closed
Open Weights	Qwen VL2 72B	Open	Closed	Open	Closed	Open	Closed
	Qwen VL2 7B	Open	Closed	Open	Closed	Open	Closed
	Intern VL2 LLAMA 76B	Open	Closed	Open	Closed	Open	Closed
	Intern VL2 8B	Open	Closed	Open	Closed	Open	Closed
	Pixtral 12B	Open	Closed	Open	Closed	Open	Closed
	Phi3.5-Vision 4B	Open	Closed	Open	Closed	Open	Closed
	PaliGemma 3B	Open	Closed	Open	Closed	Open	Closed
Open Weights & Data	LLAVA OneVision 72B	Open	Distilled	Open	Closed	Open	Closed
	LLAVA OneVision 7B	Open	Distilled	Open	Closed	Open	Closed
	Cambrian-1 34B	Open	Distilled	Open	Closed	Open	Closed
	Cambrian-1 8B	Open	Distilled	Open	Closed	Open	Closed
	xGen - MM - Interleave 4B	Open	Distilled	Open	Closed	Open	Closed
	LLAVA-1.5 13B	Open	Open	Open	Closed	Open	Closed
	LLAVA-1.5 7B	Open	Open	Open	Closed	Open	Closed

Evaluation

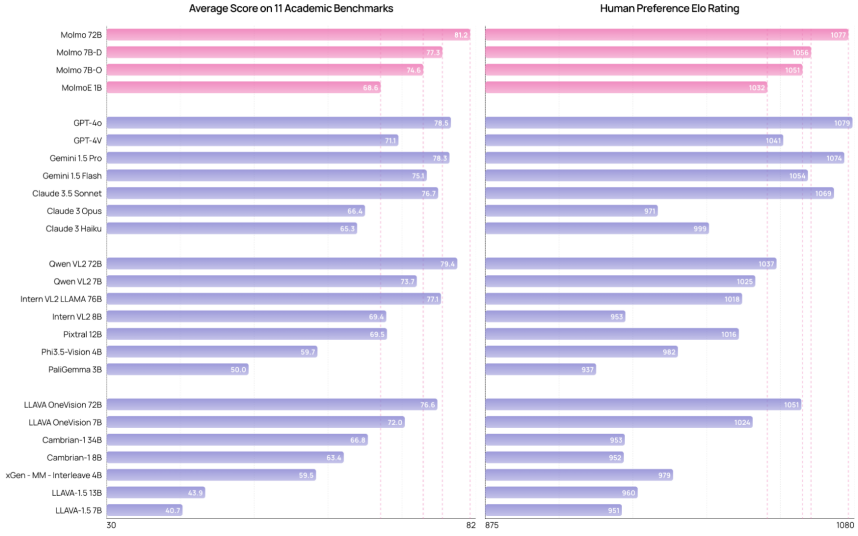


Figure 2. (Left) Average scores on the 11 academic benchmarks. See Table 1 for per-benchmark results. (Right) Elo ratings from our human preference evaluation.

Outline

1. Intro
2. Pretext tasks
3. Contrastive and masked learning
4. Foundation models
5. Benchmarking foundation models

WildVision-Arena

Figure 1: WILDVISION-ARENA (WV-ARENA) supports multi-round multimodal chats with 20+ models, enabling the comparison of VLMs in real-world scenarios. We curate WILDVISION-BENCH (WV-BENCH) by selecting 500 samples from $20k+$ in-the-wild chats and $8k+$ user ratings. Automatic model scorings on WV-BENCH closely correlate with the Elo ratings on WV-ARENA.

Lu et al. WildVision: Evaluating Vision-Language Models in the Wild with Human Preferences. arXiv:2406.11069

UI

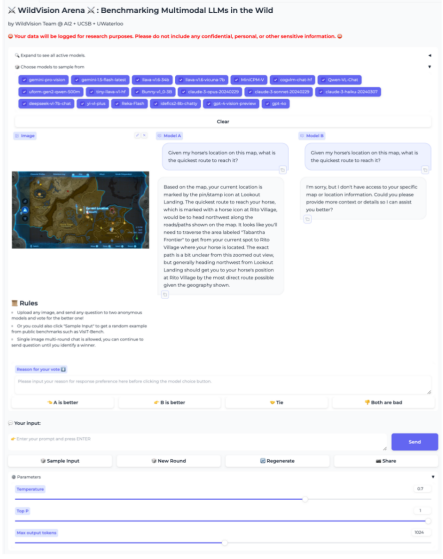


Figure 8: User Interface of WILDVISION-ARENA.

Question distribution

Statistic	Number
Total Votes	8,076
Anonymous	6,636
Non-anonymous	1,440
Left Vote	2,932
Right Vote	2,839
Tie Vote	979
Bad Vote	1,326
Days	102
Total Round	10,884
Avg Round	1.34
Avg Token Input	31.00
Avg Token Output	108.87

Table 1: Statistics of votings in WV-ARENA.

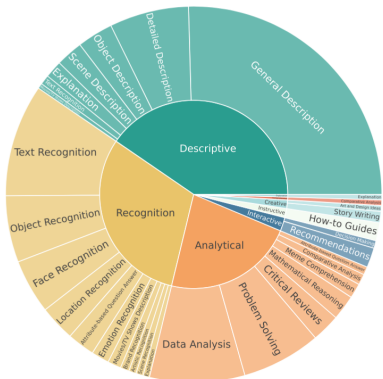


Figure 2: Question Category

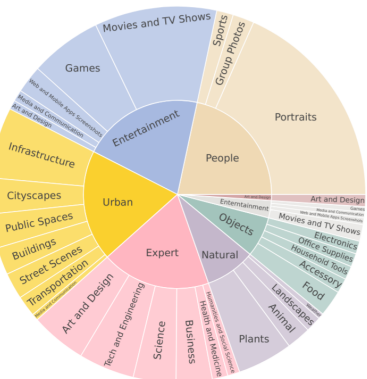


Figure 3: Image Domain

Battles

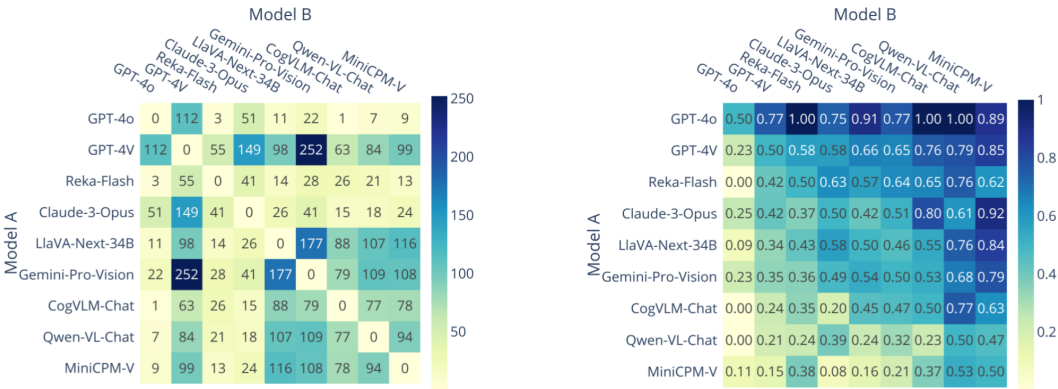


Figure 4: Battle Count Heatmap (Left): the number of voted comparisons between models. Win Fraction Heatmap (Right): the winning rate of Model A over Model B in voted comparisons.

Leaderboard

Table 2: WILDVISION-ARENA Leaderboard. We show the full elo score and within three question categories (`Analytical`, `Descriptive`, `Recognition`) and three image domains (`Entertainment`, `Objects`, `Expert`) of 22 models with a time cutoff at May 29, 2024. **Best** Second Best Best among proprietary models Best among open-source models.

Models	Size	Elo	Battles	MMMU	Question Category			Image Domain		
					Analyt.	Descri.	Recogn.	Entert.	Objects	Expert
GPT-4O [69]	–	**1235**	434	**62.8**	**1290**	**1250**	**1236**	**1362**	**1203**	**1293**
GPT-4-Vision [68]	–	1132	2288	56.8	1154	1169	1099	1177	1109	1178
Reka-Flash [83]	–	1107	513	56.3	1093	1141	1067	1069	1101	1191
Claude-3-OPUS [2]	–	1100	908	59.4	1117	1096	1092	1111	1127	1128
Gemini-Pro-Vision [82]	–	1061	2229	47.9	1099	1041	1090	1088	1077	1041
Yi-VL-PLUS [1]	–	1061	283	–	1084	1040	1078	1001	1119	1101
LLaVA-NEXT [48]	$34B$	1059	1826	51.1	1068	1104	1021	1074	1015	1052
Gemini-1.5-Flash [81]	–	1055	132	–	1090	1018	1085	1190	990	1127
Claude-3-Sonnet [2]	–	1044	496	53.1	1063	1056	1041	1033	1023	1119
CogVLM-Chat-HF [89]	$13B$	1016	1024	32.1	950	947	1006	955	930	950
Claude-3-Haiku [2]	–	1002	419	50.2	964	1008	996	1033	1014	1005
LLaVA-NEXT [48]	$7B$	992	1367	35.1	963	1032	977	992	1023	1001
DeepSeek-VL [51]	$7B$	979	646	36.6	988	984	953	956	1026	962
Idefics2 [37]	$8B$	965	100	36.6	818	1003	1011	909	1071	1020
LLaVA-NEXT [48]	$13B$	956	201	35.9	965	974	1006	975	971	987
Qwen-VL-Chat [5]	$10B$	930	1328	35.9	898	937	940	923	942	902
Bunny-V1 [23]	$3B$	921	389	38.2	897	922	878	884	823	823
MiniCPM-V [26]	$3B$	910	1349	34.7	895	911	925	888	890	840
LLaVA-v1.5 [47]	$13B$	891	299	36.4	952	838	920	887	827	914
Tiny-LLaVA-v1-HF [111]	$3B$	879	288	33.1	901	828	821	808	853	894
InstructBLIP [14]	$7B$	862	807	30.6	834	856	891	840	902	763
UFORM-Gen2-Qwen [86]	$500M$	827	452	–	911	785	853	768	937	830

Per-domain quality

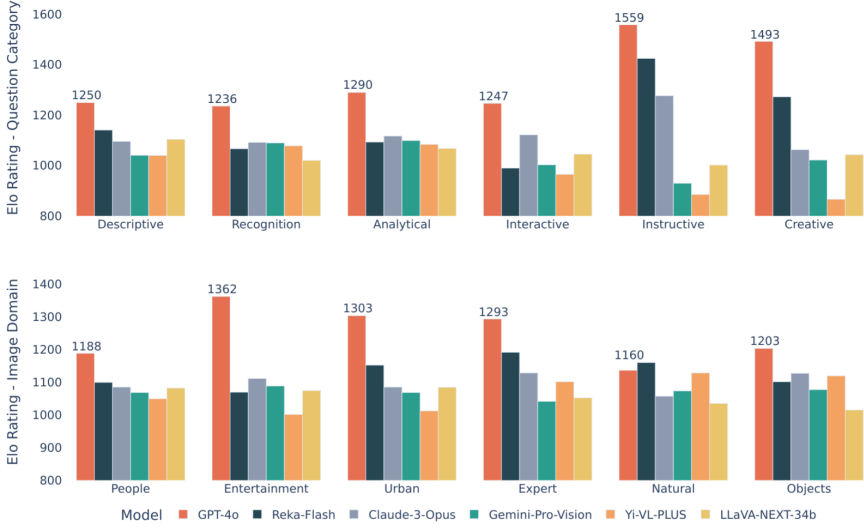


Figure 5: Elo ratings of six models across question categories (Top) and image domains (Bottom).

WV-Bench

500 data samples from Arena with expert annotation

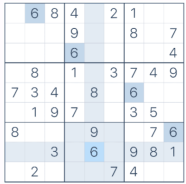
Table 3: VLMs' responses on two cases from WILDVISION-BENCH expert annotated samples. The example #61 is a hard case that all models fall short at.

Image *#44* in WildVision Bench

Text Prompt: Answer shortly what is he doing? And what do we call him?

GPT-4V: He is performing a breakdance move. We call him a breakdancer or b-boy.
LLaVA-v1.6-34B: He is performing a dance move that involves spinning or flipping while balancing on one hand ..., he may be referred to as a dancer or acrobat.
Gemini Pro Vision: He is breakdancing. He is called a b-boy or breaker.

Image *#61* in WildVision Bench



Text Prompt: Convert the first row of this sudoku matrix into the text format, each cell separated by a space. Use 0 to represent empty cells.

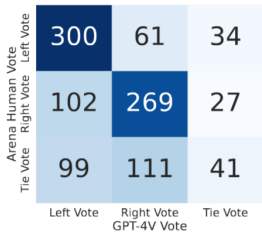
GPT-4V: 6 8 4 2 1 0 0 0 0

LLaVA-v1.6-34B: 6 8 4 0 2 1 0 7 3 4 0 6

Gemini Pro Vision: 6 8 4 2 1 9 8 7 7

WV-Bench

Evaluation is done using GPT-4o as judge and Claude-3-Sonnet as reference



Metric vs Human	GPT-4v		
	4-way	3-way	Binary
F1 Score (Macro)	0.4245	0.5143	0.7792
F1 Score (Micro)	0.5747	0.5842	0.7796
F1 Score (Weighted)	0.5407	0.5536	0.7798
Cohen's Kappa Score	0.3404	0.3442	0.5585
Pearson Correlation	0.2906	0.2880	0.5587

Figure 6: Left: GPT-4V vs. Arena Human Voting. Right: Agreement; 4-way: left/right/tie/bad vote. 3-way: left/right/other. Binary: left/right vote

WV-Bench

Evaluation is done using GPT-4o as judge and Claude-3-Sonnet as reference

Table 4: Estimated model scores of VLMs on WILDVISION-BENCHtest split of 500 samples.

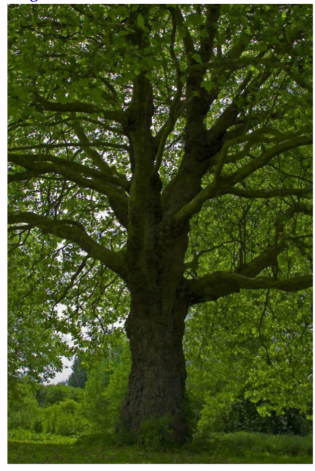
Model	Score	95% CI	Win Rate	Reward	Much Better	Better	Tie	Worse	Much Worse	Avg Tokens
GPT-4o [69]	89.41	(−1.7, 2.0)	80.6%	56.4	255.0	148.0	14.0	72.0	11.0	157
GPT-4-Vision [68]	80.01	(−1.9, 2.8)	71.8%	39.4	182.0	177.0	22.0	91.0	28.0	140
Reka-Flash [83]	64.79	(−2.9, 3.0)	58.8%	18.9	135.0	159.0	28.0	116.0	62.0	181
Claude-3-Opus [2]	62.15	(−2.8, 3.4)	53.0%	13.5	103.0	162.0	48.0	141.0	46.0	120
Yi-VL-PLUS [1]	55.09	(−2.9, 3.0)	52.8%	7.2	98.0	166.0	29.0	124.0	83.0	150
LLaVA-NEXT-34B [48]	51.91	(−3.1, 2.4)	49.2%	2.5	90.0	156.0	26.0	145.0	83.0	165
Claude-3-Sonnet [2]	50.00	–	–	–	–	–	–	–	–	120
Claude-3-Haiku [2]	37.70	(−3.2, 4.2)	30.6%	−16.5	54.0	99.0	47.0	228.0	72.0	97
Gemini-Pro-Vision [82]	35.45	(−2.6, 3.2)	32.6%	−21.0	80.0	83.0	27.0	167.0	143.0	66
LLaVA-NEXT-13B [48]	33.69	(−3.8, 2.7)	33.8%	−21.4	62.0	107.0	25.0	167.0	139.0	138
DeepSeek-VL-7B [51]	33.48	(−2.2, 3.0)	35.6%	−21.2	59.0	119.0	17.0	161.0	144.0	119
CogVLM-Chat-HF [89]	31.88	(−2.7, 2.4)	30.6%	−26.4	75.0	78.0	15.0	172.0	160.0	63
LLaVA-NEXT-7B [48]	26.15	(−2.7, 2.3)	27.0%	−31.4	45.0	90.0	36.0	164.0	165.0	139
Idefics2 [37]	23.71	(−2.4, 2.5)	26.4%	−35.8	44.0	88.0	19.0	164.0	185.0	128
Qwen-VL-Chat [5]	17.87	(−2.6, 2.2)	19.6%	−47.9	42.0	56.0	15.0	155.0	232.0	70
LLaVA-v1.5-13B [47]	14.15	(−2.2, 2.2)	16.8%	−52.5	28.0	56.0	19.0	157.0	240.0	87
Bunny-3B [23]	12.70	(−1.8, 1.9)	16.6%	−54.4	23.0	60.0	10.0	164.0	243.0	76
MiniCPM-V [26]	11.66	(−1.8, 2.1)	13.6%	−57.5	25.0	43.0	16.0	164.0	252.0	89
Tiny-LLaVA [111]	8.01	(−1.4, 1.4)	11.0%	−66.2	16.0	39.0	15.0	127.0	303.0	74
UFORM-Gen2-Qwen [86]	7.55	(−1.6, 1.1)	10.8%	−68.5	16.0	38.0	11.0	115.0	320.0	92
InstructBLIP-7B [14]	5.54	(−1.3, 1.5)	7.8%	−72.5	11.0	28.0	15.0	117.0	329.0	47

WV-Bench samples

Image [Entertainment-Movies/TV Shows]

[Descriptive-Movies/TV Shows] **Text Prompt:** What are the two giraffe characters on this movie poster doing?

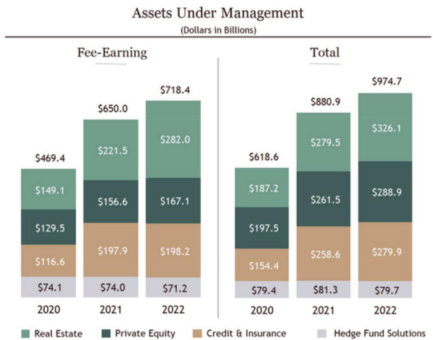
Image [Natural-Plants]



[Analytical-Problem Solving] **Text Prompt:** How likely is it to snow after this picture was taken? What would change with this type of tree before it's likely to snow?

WV-Bench samples

Image `[Expert-Business]`



`[Analytical-Data Analysis]` **Text Prompt:** Which of the companies featured in the dashboard are headquartered outside the US?

Image `[Urban-Infrastructure]`

`[Recognition-Text]` **Text Prompt:** Can you tell me the potential risks and the unreasonale parts in the image?

WV-Bench samples

Image [Urban-Buildings]

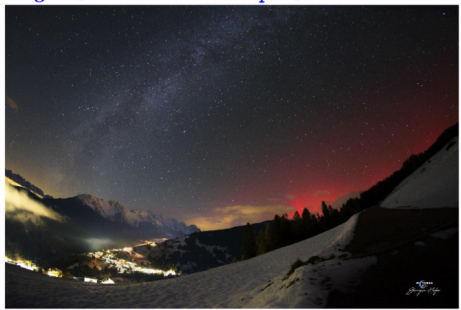
[Recognition-Location] **Text Prompt:** where is this?

Image [Expert-Science]

[Analytical-Safety Procedures] **Text Prompt:** Can you tell me the potential risks and the unreasonale parts in the image?

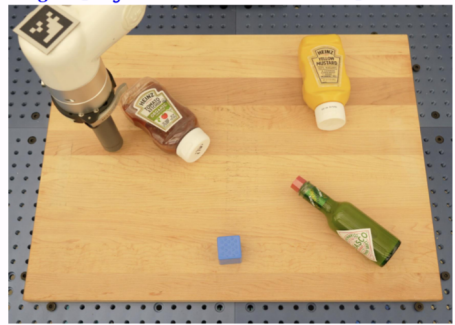
WV-Bench samples

Image `[Natural-Landscapes]`



`[Recognition-Location]` **Text Prompt:** where was this photo taken?

Image `[Objects-Household Tools]`



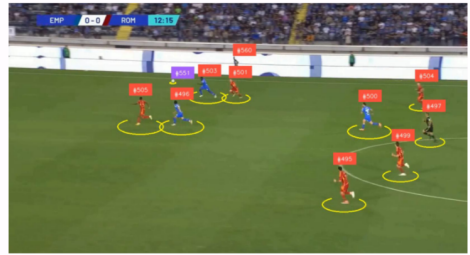
`[Descriptive-Object Description]` **Text Prompt:** describe the scene and objects

WV-Bench samples

Image `[Entertainment-Web and Mobile Apps Screenshots]`

`[Interactive-Web Navigation]` **Text Prompt:** I need to download flyer, you will be given screenshot from browser with elements marked with number. give next action to take on web page to download the flyersngive me response in below format example 1 action:[click,scroll,wait], box:1 format action:, box:

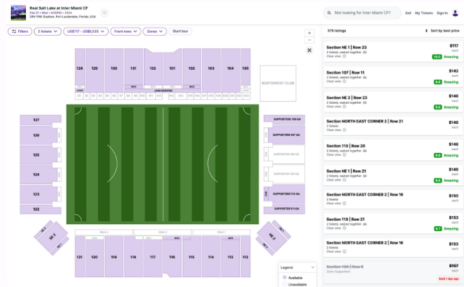
Image `[Event-Sports]`



`[Descriptive-Scene Description]` **Text Prompt:** this is a football match , every player has an identifier , describe every player action (example : player #501 is running)

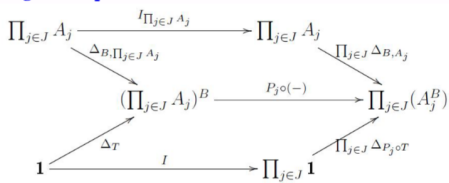
WV-Bench samples

Image `[Urban-Infrastructure]`



`[Interactive-Recommendations]` **Text Prompt:**
Which section's ticket would you recommend I purchase?

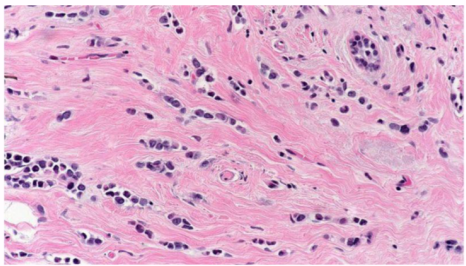
Image `[Expert-Science]`



`[Interactive-Code Generation]` **Text Prompt:**
Give me Latex code to create this diagram

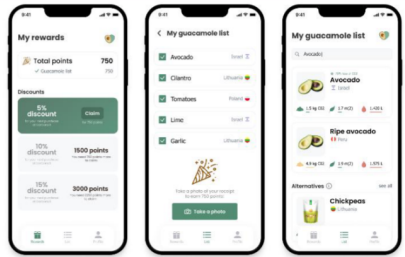
WV-Bench samples

Image [Expert-Health and Medcine]



[Recognition-Object] **Text Prompt:** what type of tumor is this?

Image [Entertainment-Web and Mobile Apps Screenshots]



[Analytical-Critical Reviews] **Text Prompt:** Review each screenshot carefully, focusing on different aspects of usability...

Conclusion

We reviewed following topics:

- motivation for self-supervised methods
- various proxy tasks for self-supervised learning
- contrastive learning and masked learning
- several definitions and examples of foundation models
- benchmarking foundation models using arenas