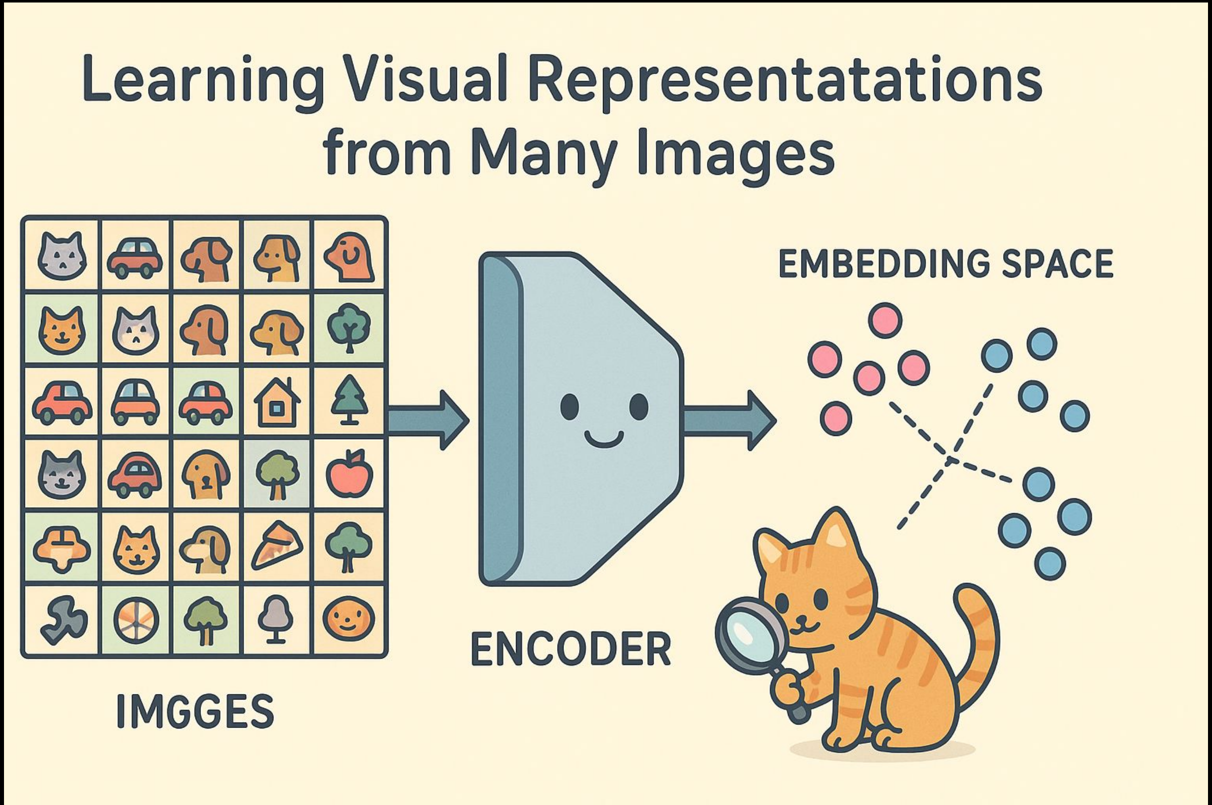
Visual Representation in the Multimodal Era

Shengbang Tong,

Visual Representation Learning



Visual Representation Learning

Self-Supervision

Language-Supervision

Visual Representation Learning

Self-Supervision
- MoCo, MAE, DINO

Language-Supervision:
- CLIP, SigLIP, MetaCLIP

Visual Representation Learning

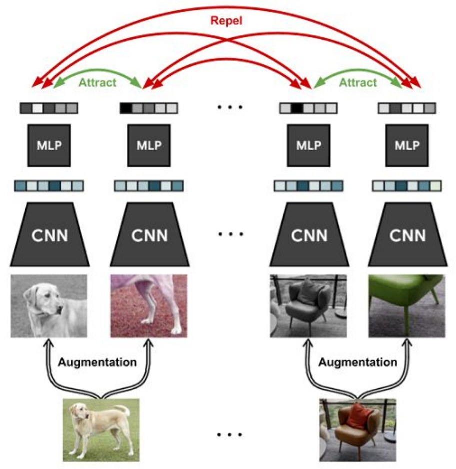
Self-Supervision
- MoCo, MAE, DINO
- Learn from images itself (augmentation, masking)

Language-Supervision:
- CLIP, SigLIP, MetaCLIP
- Learn from language that “describe the image”

Visual Representation Learning

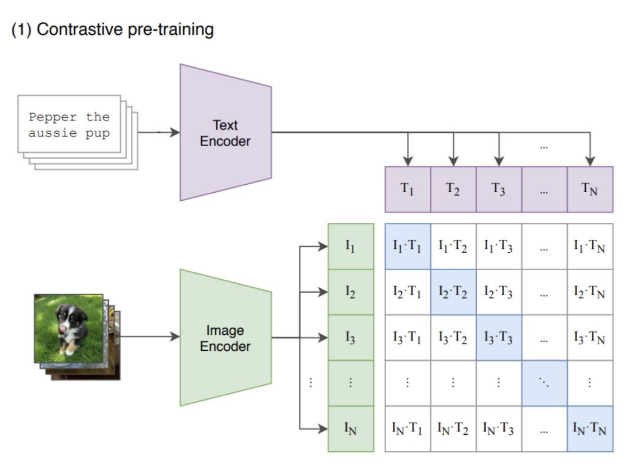
Self-Supervision

- MoCo, MAE, DINO
- Learn from images itself (augmentation, masking)



Language-Supervision:

- CLIP, SigLIP, MetaCLIP
- Learn from language that "describe the image"



Visual Representation Learning

Self-Supervision
- MoCo, MAE, DINO
- Learn from images itself (augmentation, masking)
- Train on **ImageNet-Like Data** (million scale to hundred million scale)

Language-Supervision:
- CLIP, SigLIP, MetaCLIP
- Learn from language that “describe the text”
- Train on **Image-Text pairs crawled from the internet** (400 million to 100 billion)

Visual Representation Learning

Self-Supervision
- MoCo, MAE, DINO
- Learn from images itself (augmentation, masking)
- Train on ImageNet-Like Data (million scale to hundred million scale)
- Good at <u>classification</u>, segmentation, depth estimation, etc

Language-Supervision:
- CLIP, SigLIP, MetaCLIP
- Learn from language that “describe the text”
- Train on Image-Text pairs crawled from the internet (400 million to 100 billion)
- Good at <u>classification</u>, and widely used at backbone for multimodal models

Eyes Wide Shut?

Exploring the Visual Shortcomings of Multimodal LLMs

Shengbang Tong[1], Zhuang Liu[2], Yuexiang Zhai[3], Yi Ma[3], Yann LeCun[1], Saining Xie[1]

[1]NYU,[2]FAIR, Meta AI, [3]UC Berkeley

Recap on the MLLM Architecture

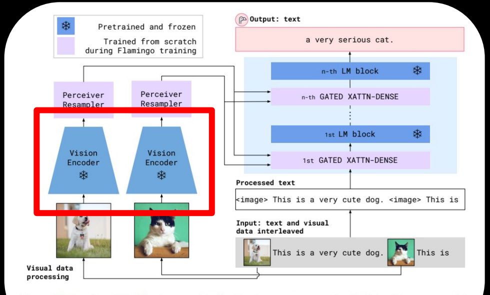
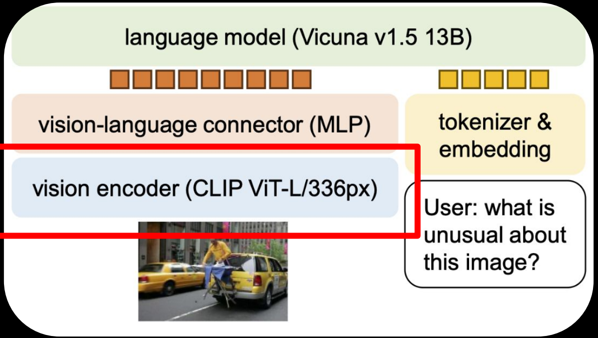
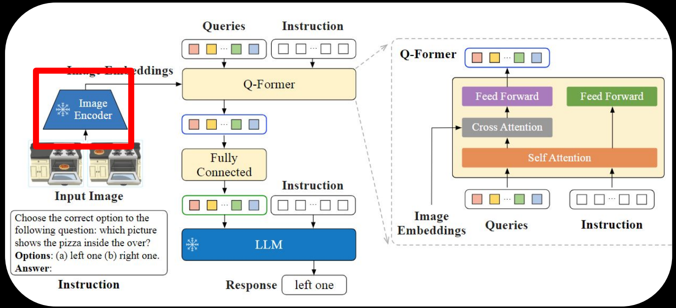


Figure 3 | **Overview of the Flamingo model.** The Flamingo models are a family of visual language model (VLM) that can take as input visual data interleaved with text and can produce free-form text as output. Key to ts performance are novel architectural components and pretraining strategies described in Section 3.

Flamingo
[Alayrac, Jean-Baptiste, et al. 2022]



LLaVA
[Liu, Haotian, et al. 2023]



InstructBLIP
[Wenliang, Dai, et al. 2023]

Recap on the MLLM Architecture

CLIP!

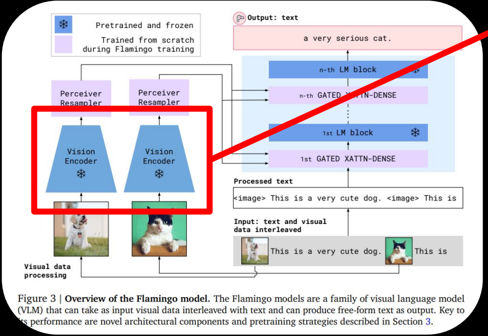
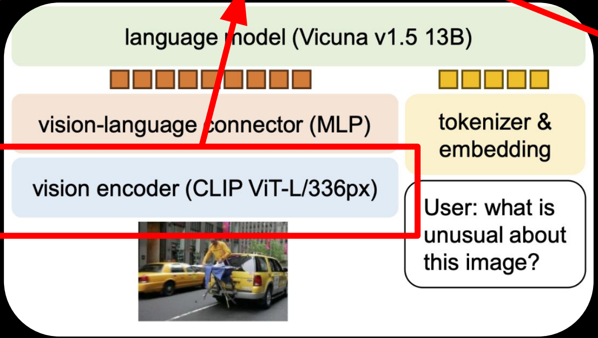
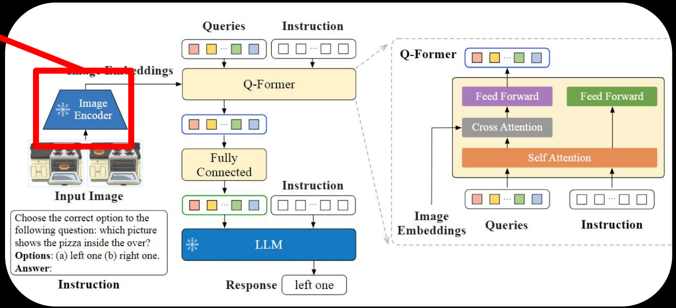


Figure 3 | **Overview of the Flamingo model.** The Flamingo models are a family of visual language model (VLM) that can take as input visual data interleaved with text and can produce free-form text as output. Key to its performance are novel architectural components and pretraining strategies described in Section 3.

Flamingo
[Alayrac, Jean-Baptiste, et al. 2022]



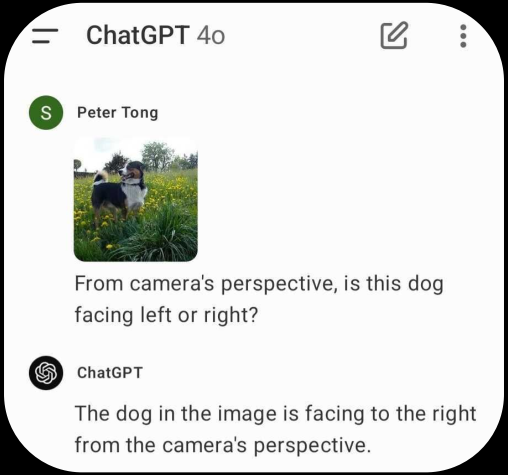
LLaVA
[Liu, Haotian, et al. 2023]

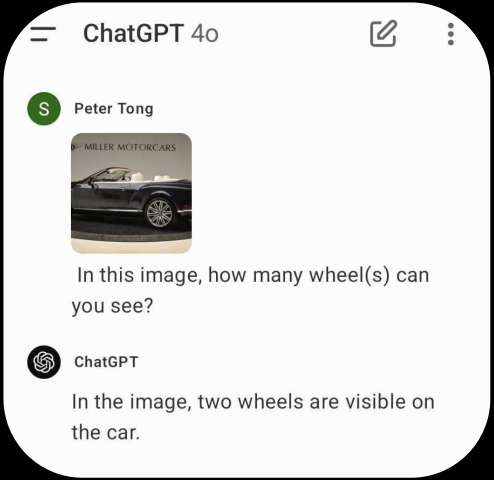


InstructBLIP
[Wenliang, Dai, et al. 2023]

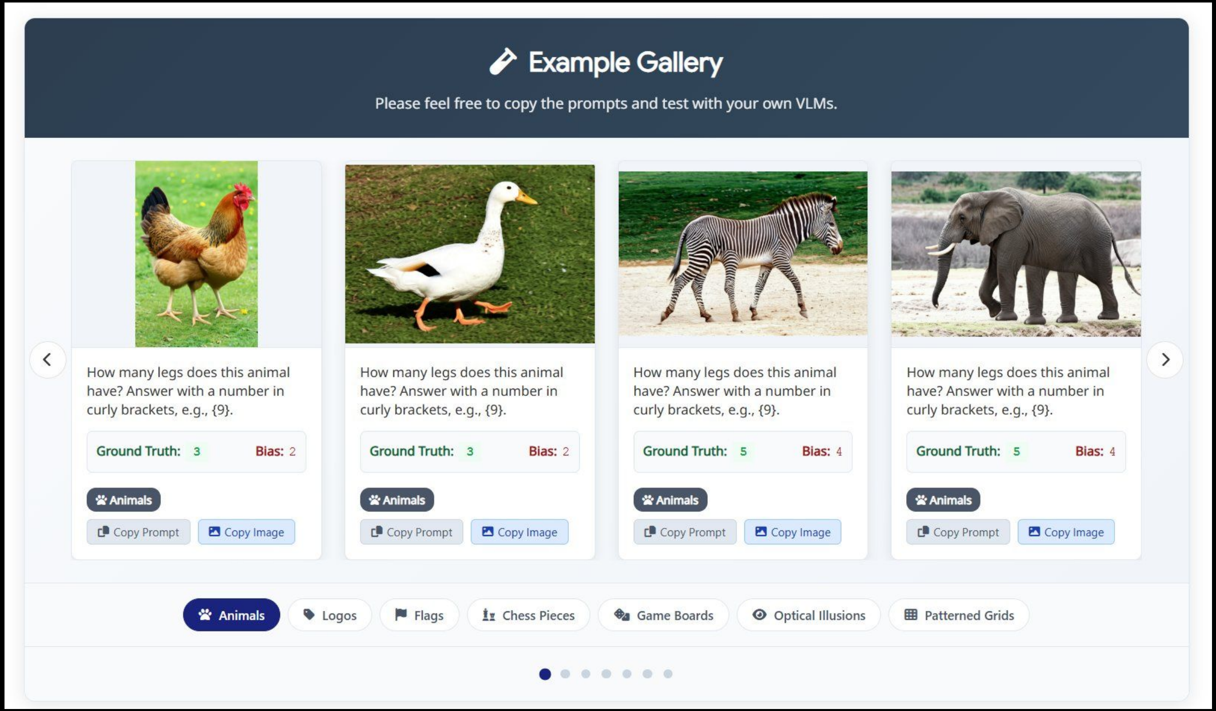
They all use a pretrained Vision Encoder, CLIP!

But MLLMs also make unexpected mistakes





But MLLMs also make unexpected mistakes



Agenda

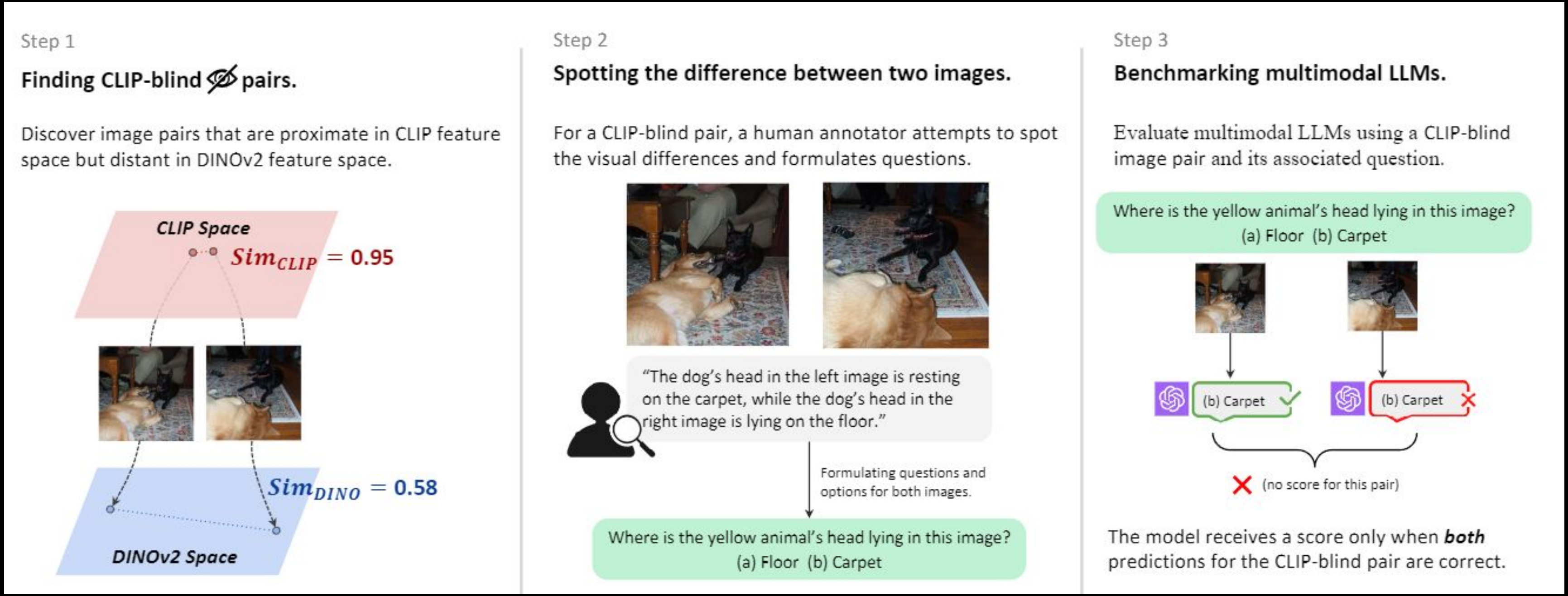
- How do we find these mistakes?

- Why do models make these mistakes?

How do we find these mistakes?

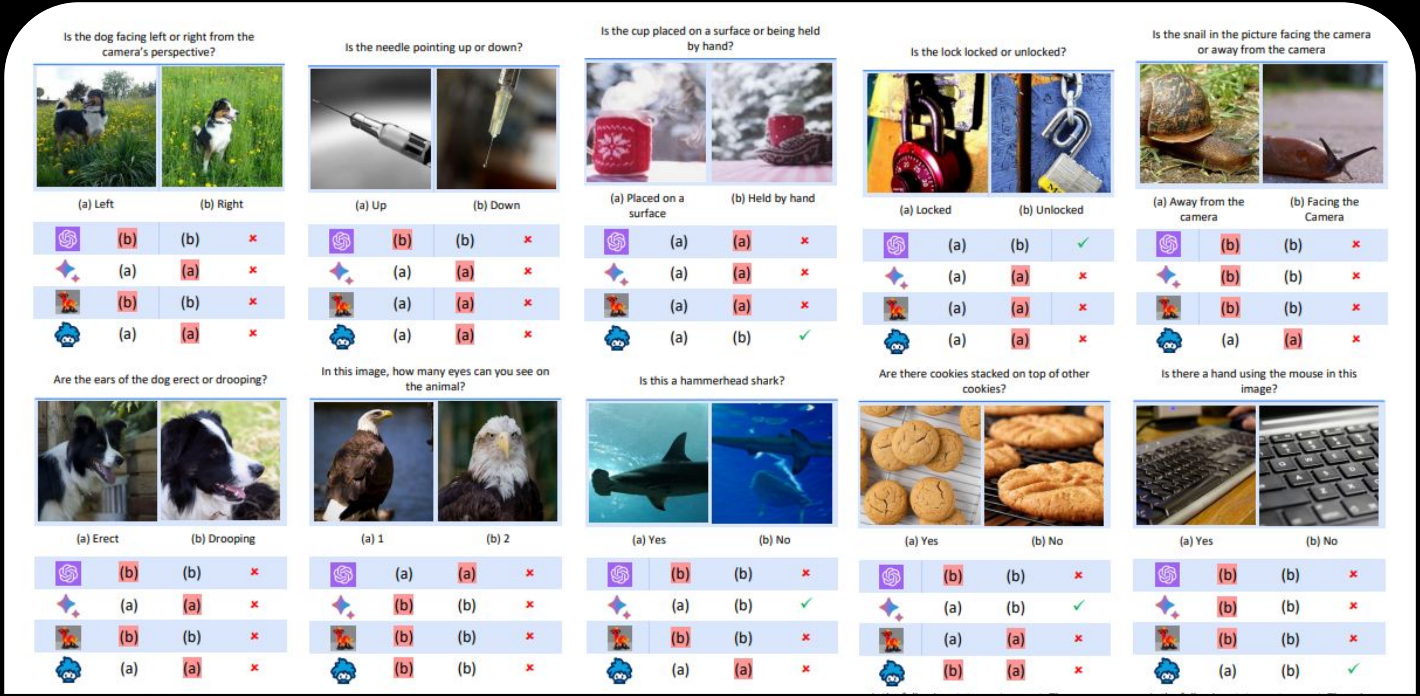
Finding CLIP-blind pairs

CLIP-blind pairs: If two images are encoded similarly by the CLIP model yet very different in visual appearance, then at least one of them has been inaccurately encoded.

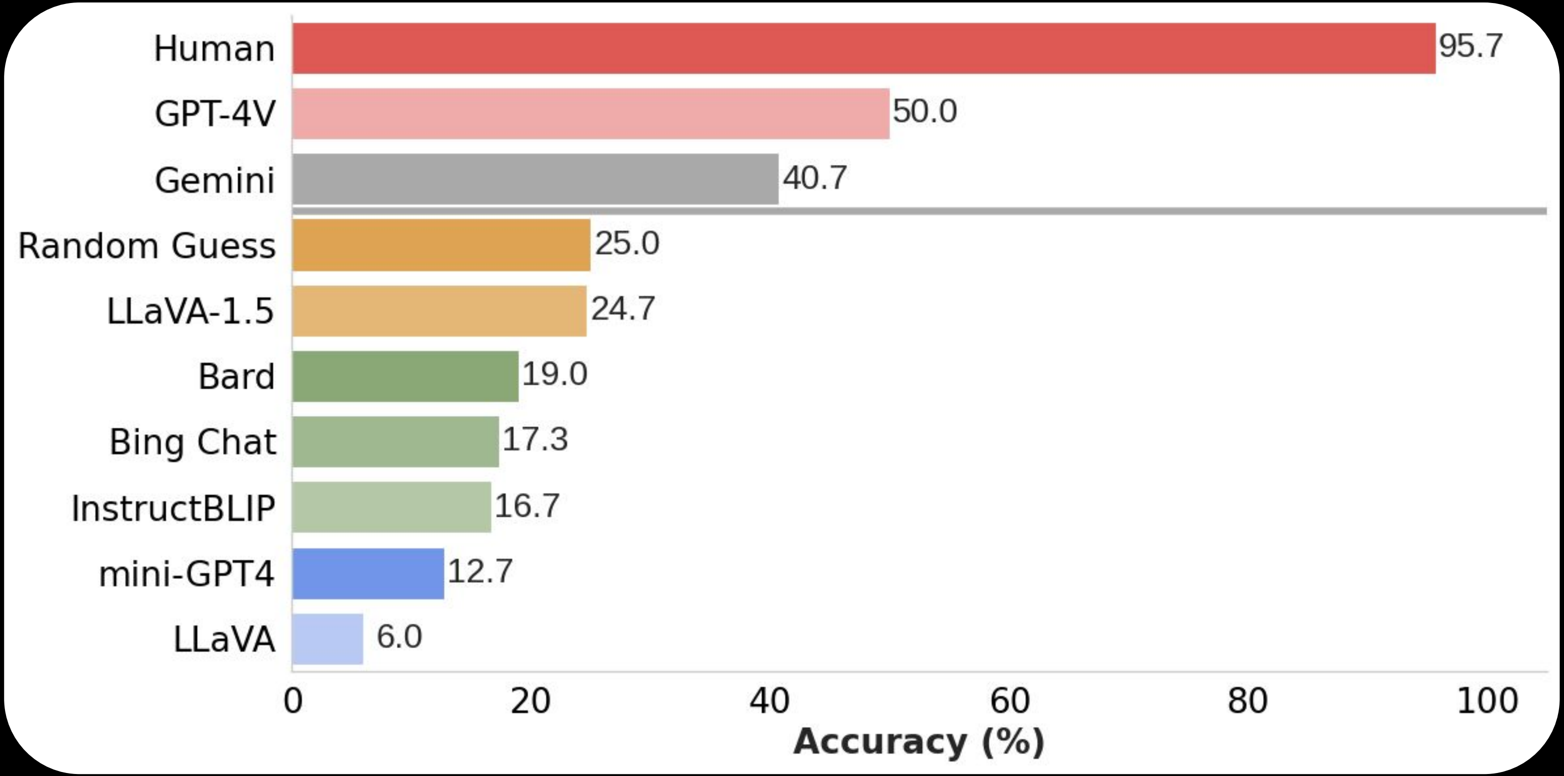


MMVP (MultiModal Visual Patterns) Benchmark

MMVP Benchmark: 150 CLIP-blind pairs & handcrafted questions



MMVP Benchmark Results



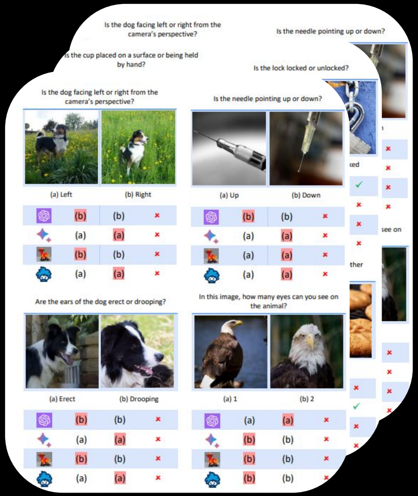
MMVP Benchmark Results

Capability	Benchmark	Seed 1.5-VL thinking	Seed 1.5-VL non-thinking	Gemini 2.5 Pro thinking	OpenAI o1 thinking	Claude 3.7 Sonnet thinking	OpenAI GPT-4o non-thinking	Qwen 2.5-VL 72B non-thinking
Multimodal reasoning	MMMU	77.9	73.6	**81.7**	77.6	75.2*	70.7*	70.2
	MMMU-Pro	67.6	59.9	**68.8***	66.4*	50.1*	54.5*	51.1
	MathVision	68.7	65.5	**73.3***	63.2*	58.6*	31.2*	38.1
	OlympiadBench	65.0	60.4	**69.8***	48.5*	54.2*	25.9*	35.9
	MathVista	**85.6**	83.0	82.7*	71.8	74.5*	63.8*	74.8
	V^*	89.0	**89.5**	79.1*	69.7*	86.4*	73.9*	86.4
	VLM are Blind	**92.1**	90.8	84.3*	57.0*	69.0*	50.4*	69
	ZeroBench (main)	2	0	**3***	0*	**3***	0*	0
	ZeroBench (sub)	**30.8**	29.0	26.0*	20.2*	20.4*	19.6*	13.0
	VisuLogic	**35.0**	33.0	31.0*	29.0*	24.8*	26.3*	28.0
General visual question answering	RealWorldQA	**78.4**	77.0	78.0*	77.1*	67.8*	76.2*	75.7
	SimpleVQA	**63.4**	63.1	62.0*	58.8*	50.1*	52.4*	52.4
	MMStar	**77.8**	76.2	77.5*	67.5*	68.8*	65.1*	70.8
	MMBench-en	89.9	88.0	**90.1***	83.8*	82.0*	84.3*	88.6
	MMBench-cn	89.1	88.1	**89.7***	81.3*	82.7*	82.0*	87.9
	MMVP	69.3	**70.7**	**70.7***	–†	–†	**70.7***	66.7
	HallusionBench	60.3	60.0	**63.7***	55.6*	58.3*	56.2*	55.2

Why do models make these mistakes?

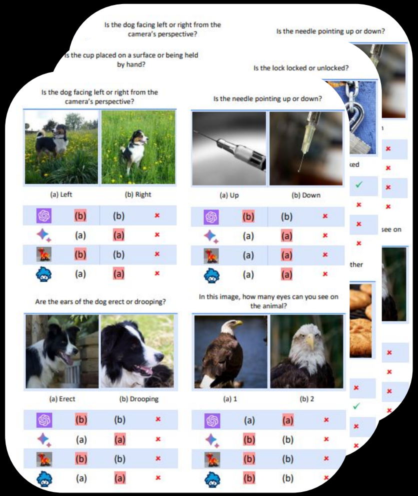
Finding Patterns in CLIP-blind Pairs

Questions in MMVP:



Finding Patterns in CLIP-blind Pairs

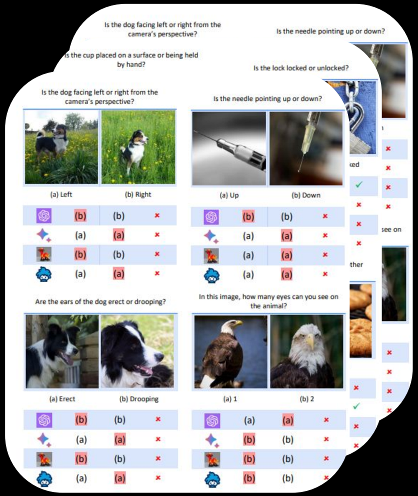
Questions in MMVP:



: Summarize Patterns

Finding Patterns in CLIP-blind Pairs

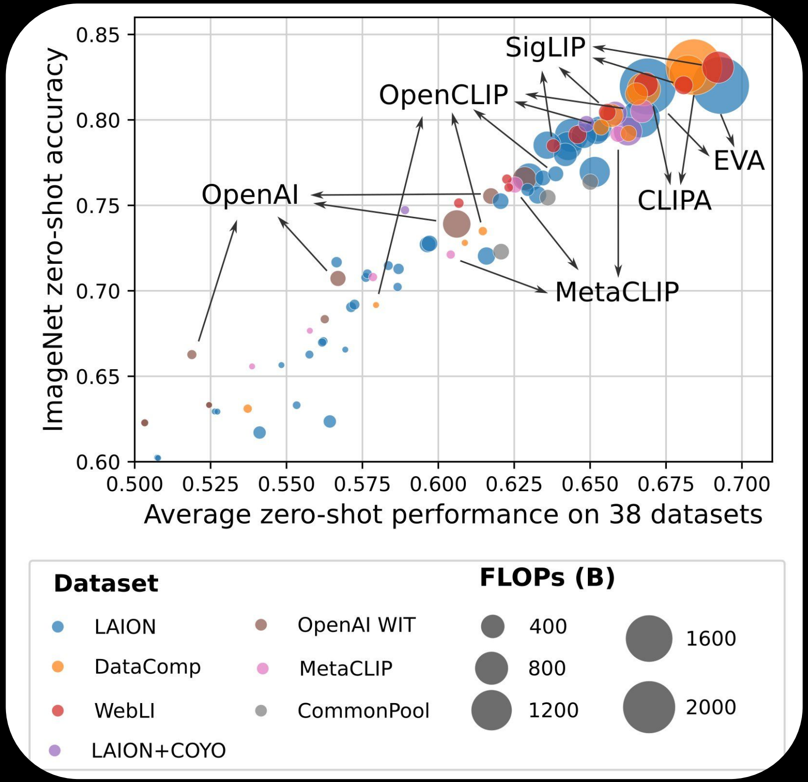
Questions in MMVP:



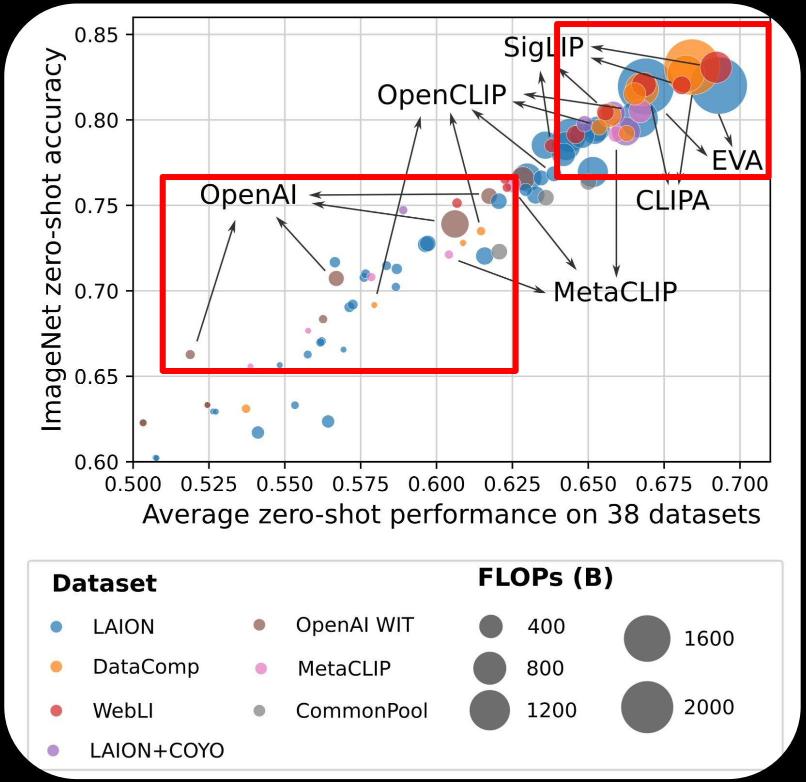
: Summarize Patterns

Visual Patterns:

- Orientation and Direction
- Presence of Specific Features
- State and Condition
- Quantity and Count
- Positional and Relational Context
- Color and Appearance
- Structural and Physical Characteristics
- Texts
- Viewpoint and Perspective

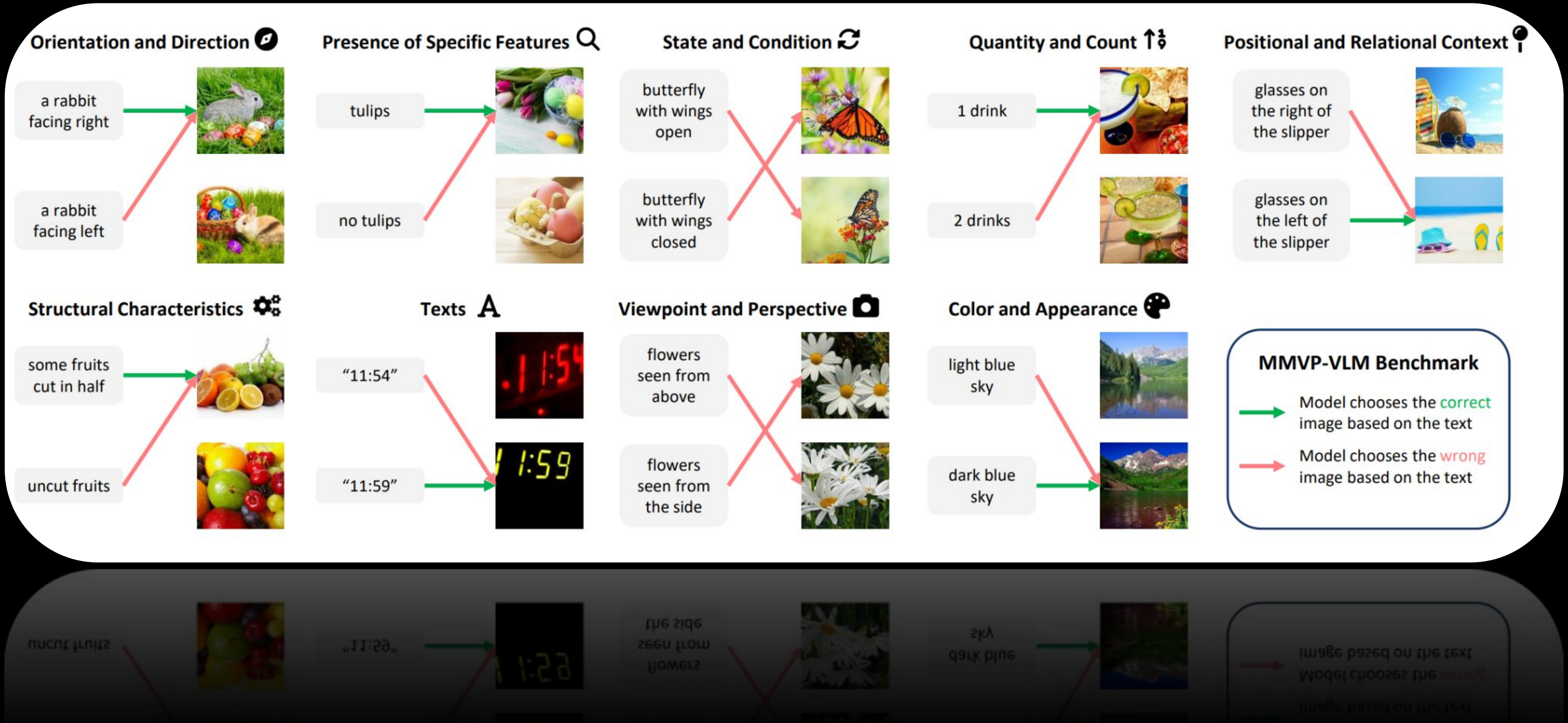


Does scaling up CLIP solves the problem?



Does scaling up CLIP solves the problem?

MMVP-VLM Benchmark



CLIP models struggle

	Image Size	Params (M)	IN-1k ZeroShot										MMVP Average
OpenAI ViT-L-14 [35]	224^2	427.6	75.5	13.3	13.3	20.0	20.0	13.3	53.3	20.0	6.7	13.3	19.3
OpenAI ViT-L-14 [35]	336^2	427.9	76.6	0.0	20.0	40.0	20.0	6.7	20.0	33.3	6.7	33.3	20.0
SigLIP ViT-SO-14 [53]	224^2	877.4	82.0	26.7	20.0	53.3	40.0	20.0	66.7	40.0	20.0	53.3	37.8
SigLIP ViT-SO-14 [53]	384^2	878.0	83.1	20.0	26.7	60.0	33.3	13.3	66.7	33.3	26.7	53.3	37.0
DFN ViT-H-14 [9]	224^2	986.1	83.4	20.0	26.7	73.3	26.7	26.7	66.7	46.7	13.3	53.3	39.3
DFN ViT-H-14 [9]	378^2	986.7	84.4	13.3	20.0	53.3	33.3	26.7	66.7	40.0	20.0	40.0	34.8
MetaCLIP ViT-L-14 [49]	224^2	427.6	79.2	13.3	6.7	66.7	6.7	33.3	46.7	20.0	6.7	13.3	23.7
MetaCLIP ViT-H-14 [49]	224^2	986.1	80.6	6.7	13.3	60.0	13.3	6.7	53.3	26.7	13.3	33.3	25.2
EVA01 ViT-g-14 [43]	224^2	1136.4	78.5	6.7	26.7	40.0	6.7	13.3	66.7	13.3	13.3	20.0	23.0
EVA02 ViT-bigE-14+ [43]	224^2	5044.9	82.0	13.3	20.0	66.7	26.7	26.7	66.7	26.7	20.0	33.3	33.3

CLIP models struggle

#1: Scaling up resolution **does not help**

	Image Size	Params (M)	IN-1k ZeroShot										MMVP Average
OpenAI ViT-L-14 [35]	224^2	427.6	75.5	13.3	13.3	20.0	20.0	13.3	53.3	20.0	6.7	13.3	19.3
OpenAI ViT-L-14 [35]	336^2	427.9	76.6	0.0	20.0	40.0	20.0	6.7	20.0	33.3	6.7	33.3	20.0
SigLIP ViT-SO-14 [53]	224^2	877.4	82.0	26.7	20.0	53.3	40.0	20.0	66.7	40.0	20.0	53.3	37.8
SigLIP ViT-SO-14 [53]	384^2	878.0	83.1	20.0	26.7	60.0	33.3	13.3	66.7	33.3	26.7	53.3	37.0
DFN ViT-H-14 [9]	224^2	986.1	83.4	20.0	26.7	73.3	26.7	26.7	66.7	46.7	13.3	53.3	39.3
DFN ViT-H-14 [9]	378^2	986.7	84.4	13.3	20.0	53.3	33.3	26.7	66.7	40.0	20.0	40.0	34.8
MetaCLIP ViT-L-14 [49]	224^2	427.6	79.2	13.3	6.7	66.7	6.7	33.3	46.7	20.0	6.7	13.3	23.7
MetaCLIP ViT-H-14 [49]	224^2	986.1	80.6	6.7	13.3	60.0	13.3	6.7	53.3	26.7	13.3	33.3	25.2
EVA01 ViT-g-14 [43]	224^2	1136.4	78.5	6.7	26.7	40.0	6.7	13.3	66.7	13.3	13.3	20.0	23.0
EVA02 ViT-bigE-14+ [43]	224^2	5044.9	82.0	13.3	20.0	66.7	26.7	26.7	66.7	26.7	20.0	33.3	33.3

CLIP models struggle

#1: Scaling up resolution **does not help**

#2: Scaling up network **helps a little**

	Image Size	Params (M)	IN-1k ZeroShot										MMVP Average
OpenAI ViT-L-14 [35]	224^2	427.6	75.5	13.3	13.3	20.0	20.0	13.3	53.3	20.0	6.7	13.3	19.3
OpenAI ViT-L-14 [35]	336^2	427.9	76.6	0.0	20.0	40.0	20.0	6.7	20.0	33.3	6.7	33.3	20.0
SigLIP ViT-SO-14 [53]	224^2	877.4	82.0	26.7	20.0	53.3	40.0	20.0	66.7	40.0	20.0	53.3	37.8
SigLIP ViT-SO-14 [53]	384^2	878.0	83.1	20.0	26.7	60.0	33.3	13.3	66.7	33.3	26.7	53.3	37.0
DFN ViT-H-14 [9]	224^2	986.1	83.4	20.0	26.7	73.3	26.7	26.7	66.7	46.7	13.3	53.3	39.3
DFN ViT-H-14 [9]	378^2	986.7	84.4	13.3	20.0	53.3	33.3	26.7	66.7	40.0	20.0	40.0	34.8
MetaCLIP ViT-L-14 [49]	224^2	427.6	79.2	13.3	6.7	66.7	6.7	33.3	46.7	20.0	6.7	13.3	23.7
MetaCLIP ViT-H-14 [49]	224^2	986.1	80.6	6.7	13.3	60.0	13.3	6.7	53.3	26.7	13.3	33.3	25.2
EVA01 ViT-g-14 [43]	224^2	1136.4	78.5	6.7	26.7	40.0	6.7	13.3	66.7	13.3	13.3	20.0	23.0
EVA02 ViT-bigE-14+ [43]	224^2	5044.9	82.0	13.3	20.0	66.7	26.7	26.7	66.7	26.7	20.0	33.3	33.3

CLIP models struggle

#1: Scaling up resolution **does not help**

#2: Scaling up network **helps a little**

#3: Scaling up data **helps a little**

	Image Size	Params (M)	IN-1k ZeroShot										MMVP Average
OpenAI ViT-L-14 [35]	224^2	427.6	75.5	13.3	13.3	20.0	20.0	13.3	53.3	20.0	6.7	13.3	19.3
OpenAI ViT-L-14 [35]	336^2	427.9	76.6	0.0	20.0	40.0	20.0	6.7	20.0	33.3	6.7	33.3	20.0
SigLIP ViT-SO-14 [53]	224^2	877.4	82.0	26.7	20.0	53.3	40.0	20.0	66.7	40.0	20.0	53.3	37.8
SigLIP ViT-SO-14 [53]	384^2	878.0	83.1	20.0	26.7	60.0	33.3	13.3	66.7	33.3	26.7	53.3	37.0
DFN ViT-H-14 [9]	224^2	986.1	83.4	20.0	26.7	73.3	26.7	26.7	66.7	46.7	13.3	53.3	39.3
DFN ViT-H-14 [9]	378^2	986.7	84.4	13.3	20.0	53.3	33.3	26.7	66.7	40.0	20.0	40.0	34.8
MetaCLIP ViT-L-14 [49]	224^2	427.6	79.2	13.3	6.7	66.7	6.7	33.3	46.7	20.0	6.7	13.3	23.7
MetaCLIP ViT-H-14 [49]	224^2	986.1	80.6	6.7	13.3	60.0	13.3	6.7	53.3	26.7	13.3	33.3	25.2
EVA01 ViT-g-14 [43]	224^2	1136.4	78.5	6.7	26.7	40.0	6.7	13.3	66.7	13.3	13.3	20.0	23.0
EVA02 ViT-bigE-14+ [43]	224^2	5044.9	82.0	13.3	20.0	66.7	26.7	26.7	66.7	26.7	20.0	33.3	33.3

CLIP models struggle

#1: Scaling up resolution **does not help**

#2: Scaling up network **helps a little**

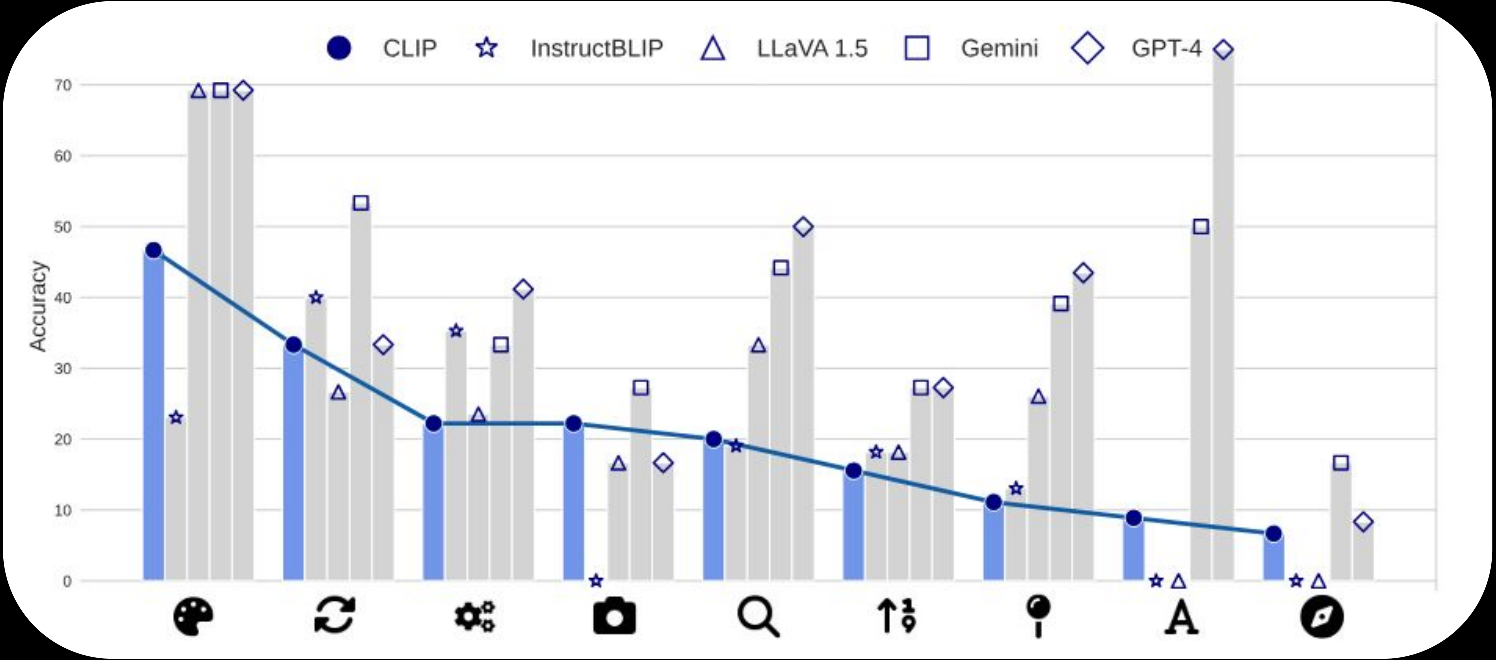
#3: Scaling up data **helps a little**

#4: All CLIP-variants **struggle**

	Image Size	Params (M)	IN-1k ZeroShot										MMVP Average
OpenAI ViT-L-14 [35]	224^2	427.6	75.5	13.3	13.3	20.0	20.0	13.3	53.3	20.0	6.7	13.3	19.3
OpenAI ViT-L-14 [35]	336^2	427.9	76.6	0.0	20.0	40.0	20.0	6.7	20.0	33.3	6.7	33.3	20.0
SigLIP ViT-SO-14 [53]	224^2	877.4	82.0	26.7	20.0	53.3	40.0	20.0	66.7	40.0	20.0	53.3	37.8
SigLIP ViT-SO-14 [53]	384^2	878.0	83.1	20.0	26.7	60.0	33.3	13.3	66.7	33.3	26.7	53.3	37.0
DFN ViT-H-14 [9]	224^2	986.1	83.4	20.0	26.7	73.3	26.7	26.7	66.7	46.7	13.3	53.3	39.3
DFN ViT-H-14 [9]	378^2	986.7	84.4	13.3	20.0	53.3	33.3	26.7	66.7	40.0	20.0	40.0	34.8
MetaCLIP ViT-L-14 [49]	224^2	427.6	79.2	13.3	6.7	66.7	6.7	33.3	46.7	20.0	6.7	13.3	23.7
MetaCLIP ViT-H-14 [49]	224^2	986.1	80.6	6.7	13.3	60.0	13.3	6.7	53.3	26.7	13.3	33.3	25.2
EVA01 ViT-g-14 [43]	224^2	1136.4	78.5	6.7	26.7	40.0	6.7	13.3	66.7	13.3	13.3	20.0	23.0
EVA02 ViT-bigE-14+ [43]	224^2	5044.9	82.0	13.3	20.0	66.7	26.7	26.7	66.7	26.7	20.0	33.3	33.3

Mistakes in CLIP and MLLM are correlated

The worse CLIP models are, the worse MLLMs are.

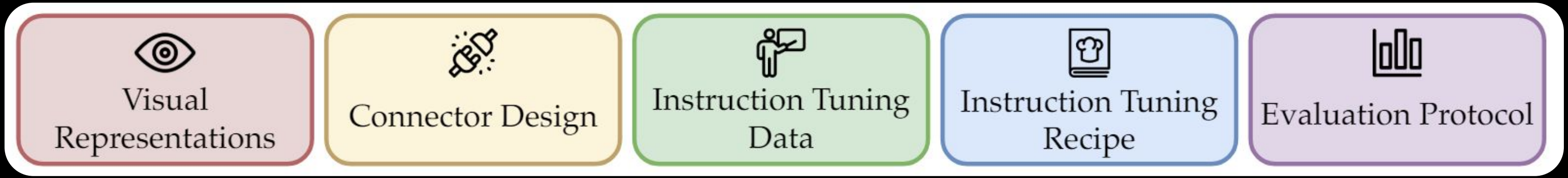


What’s next?

What's next for more *vision-centric* MLLMs?

- What visual representations to use?
- How to align modalities?
- What data do we train the model?
- How to train the model?
- How do we evaluate and interpret results?

…



Cambrian-1
A Fully Open, *Vision-Centric* Exploration of Multimodal LLMs

Shengbang Tong*, Ellis Brown*, Penghao Wu*,
Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang,
Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang,
Rob Fergus, Yann LeCun, Saining Xie

New York University

Some Trivia

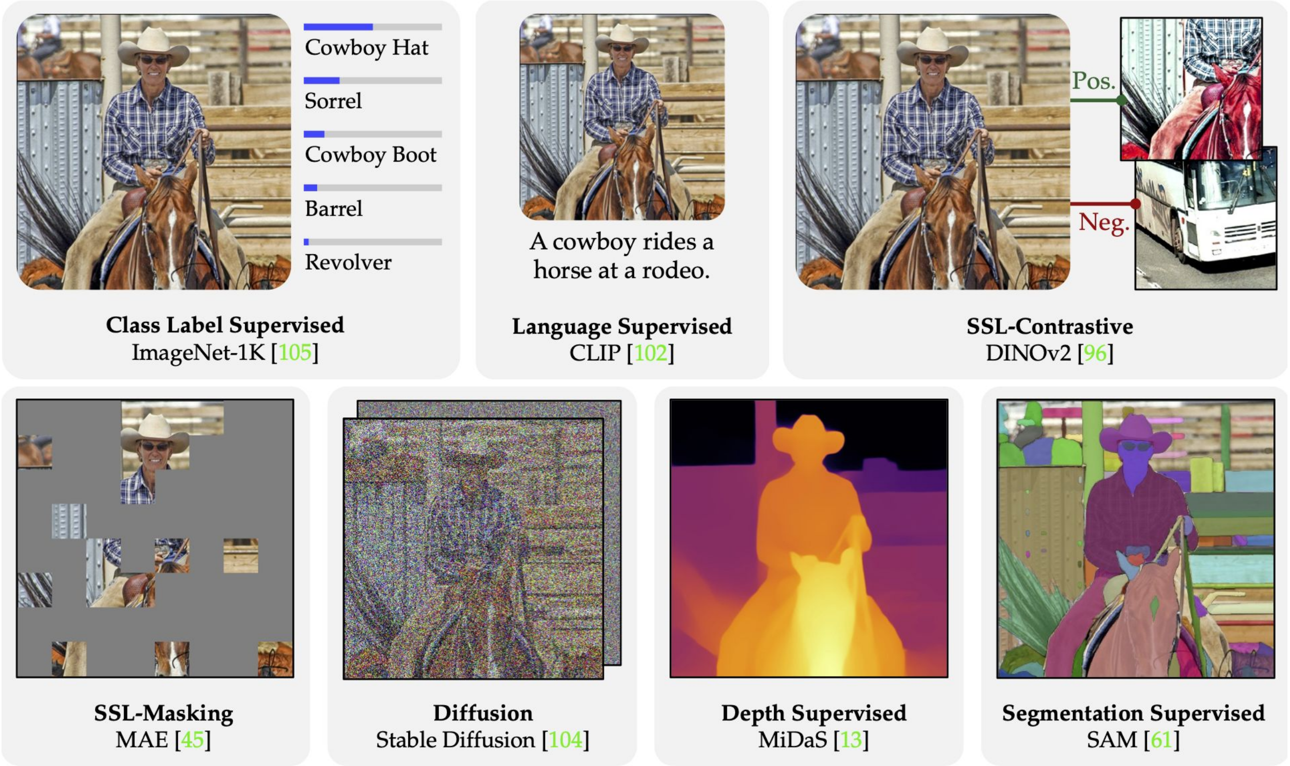
when did vision develop in animals?

Vision in animals is believed to have developed around 541 million years ago during the Cambrian Explosion a period marked by a rapid diversification of life forms. During this time, the first complex eyes appeared in early arthropods, which likely provided a significant evolutionary advantage by allowing these organisms to better navigate their environments, find food, and avoid predators. This development of vision played a crucial role in the predator-prey dynamics and the overall evolutionary arms race that characterized the Cambrian period.

Visual Representations for MLLMs

How to evaluate visual reprs.?

What visual reprs. to use?

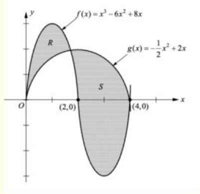


Visual Representations for MLLMs

Supervision Type	Method	Architecture	Patch Size	Res.	# Tok.	Hidden Size
Language-Supervised						
Language	OpenAI CLIP	ViT-L	14	336	576	768
	DFN-CLIP	ViT-L	14	224	256	1024
	DFN-CLIP	ViT-H	14	378	729	1280
	EVA-CLIP-02	ViT-L	14	336	576	1024
	SigLIP	ViT-L	16	384	576	1024
	SigLIP	ViT-SO400M	14	384	729	1152
	OpenCLIP	ConvNeXT-L	-	512	[I]576	1536
	OpenCLIP	ConvNeXT-L	-	1024	[I]576	1536
	OpenCLIP	ConvNeXT-XXL	-	1024	[I]576	3072
Self-Supervised						
Contrastive	DINOv2	ViT-L	14	336	576	1024
	DINOv2	ViT-L	14	518	[I]576	1024
	MoCo v3	ViT-B	16	224	196	768
	MoCo v3	ViT-L	16	224	196	1024
Masked	MAE	ViT-L	16	224	196	1024
	MAE	ViT-H	14	224	256	1280
JEPA	I-JEPA	ViT-H	14	224	256	1280
Other						
Segmentation	SAM	ViT-L	16	1024	[I]576	1024
	SAM	ViT-L	16	1024	[I]576	1280
Depth	MiDaS 3.0	ViT-L	16	384	576	1024
	MiDaS 3.1	ViT-L	16	518	1024	1024
Diffusion	Stable Diffusion 2.1	VAE+UNet	16	512	1024	3520
Class Labels	SupViT	ViT-L	16	224	196	1024
	SupViT	ViT-H	14	224	256	1280

23 models!

Evaluation Protocol



Question: <*image 1*> The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R.

Options:

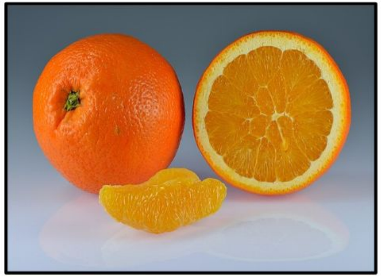
(A) $\int_0^{1.5}[f(x) - g(x)]dx$

(B) $\int_0^{1.5}[g(x) - f(x)]dx$

(C) $\int_0^{2}[f(x) - g(x)]dx$

(D) $\int_0^{2}[g(x) - x(x)]dx$

MMMU [Yue, et al. 2024]



Q: what is the color of this object?

A. Purple

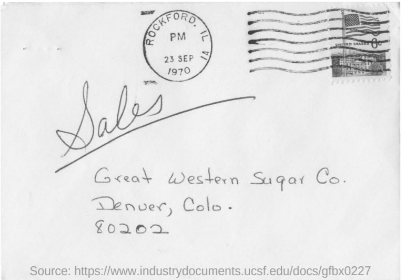
B. Pink

C. Gray

D. Orange

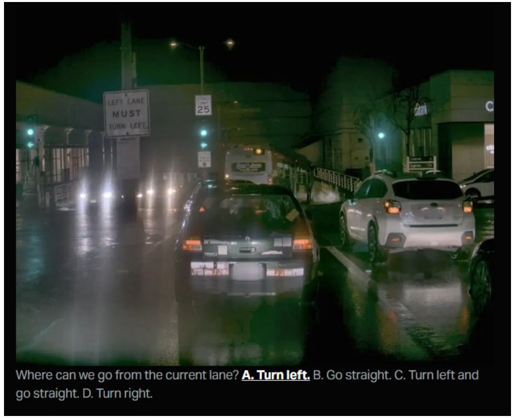
GT: D

MM-Bench [Liu, et al. 2024]

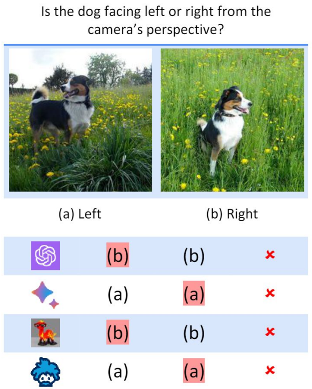


Q: Mention the ZIP code written?
A: 80202
Q: What date is seen on the seal at the top of the letter?
A: 23 sep 1970
Q: Which company address is mentioned on the letter?
A: Great western sugar Co.

DocVQA [Mathew, et al. 2020]



RealWorldQA [Grok, et al. 2024]



MMVP [Tong, et al. 2024]

and a lot more…

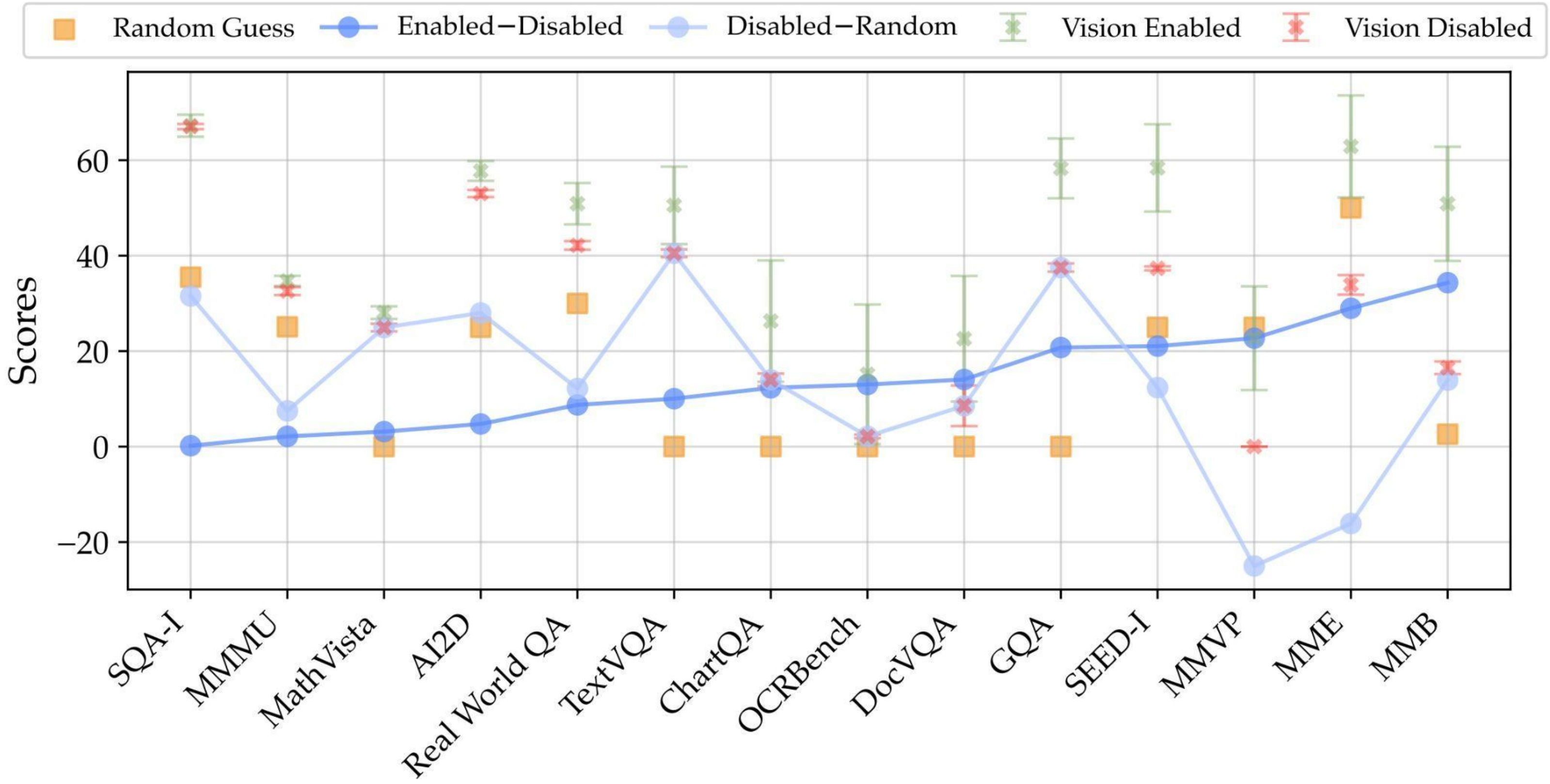
How should we systematically evaluate an MLLM and interpret the evaluation results?

Benchmark Analysis

1. Assess the “Multimodality” of the Benchmarks

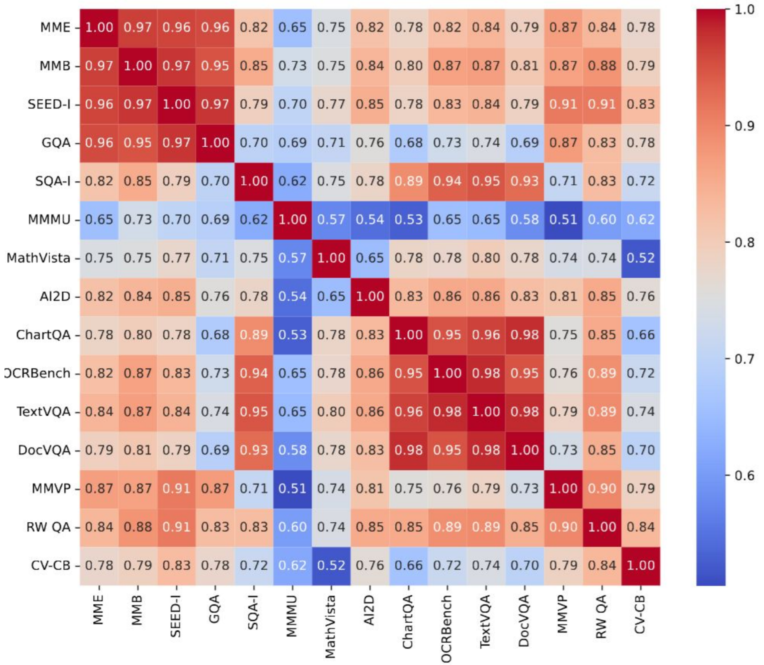
2. Group Benchmarks into Clusters

Who's answering the question: the **LLM** or **MLLM**?

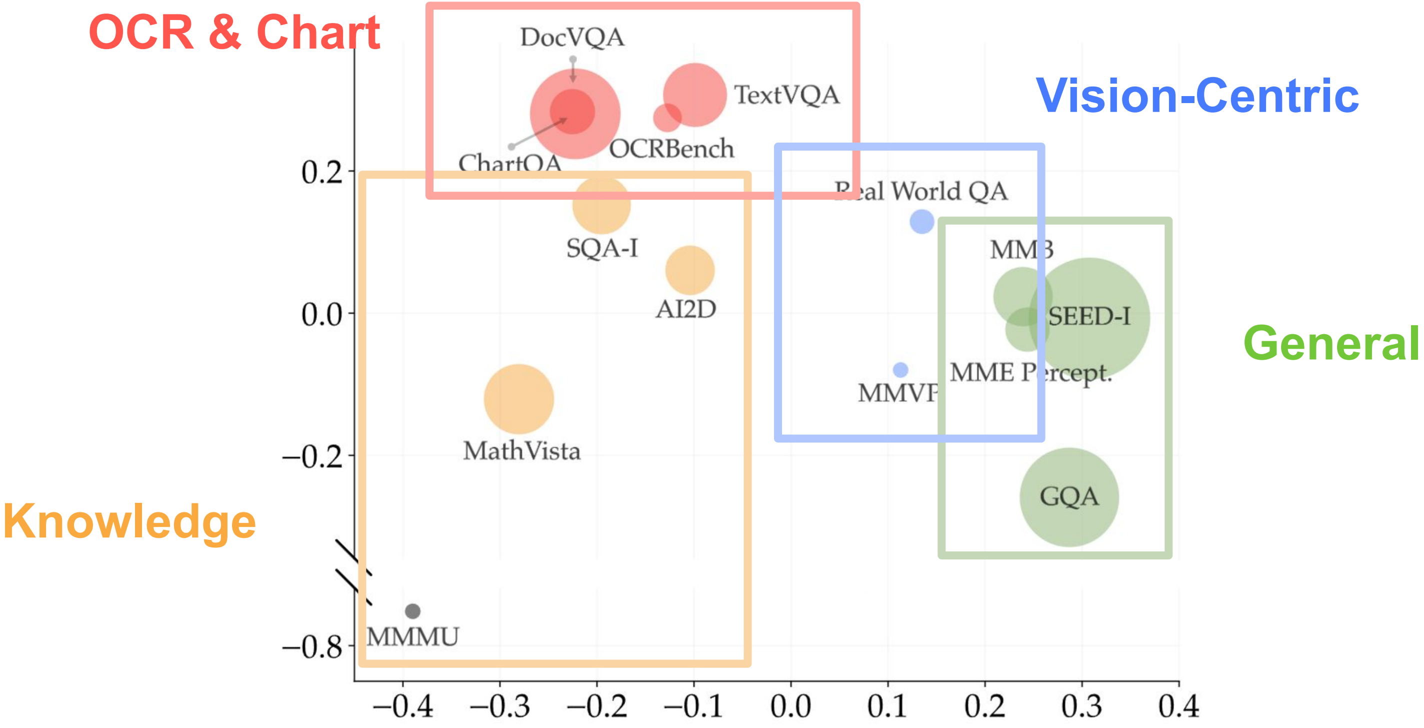


Group Benchmarks by Correlation

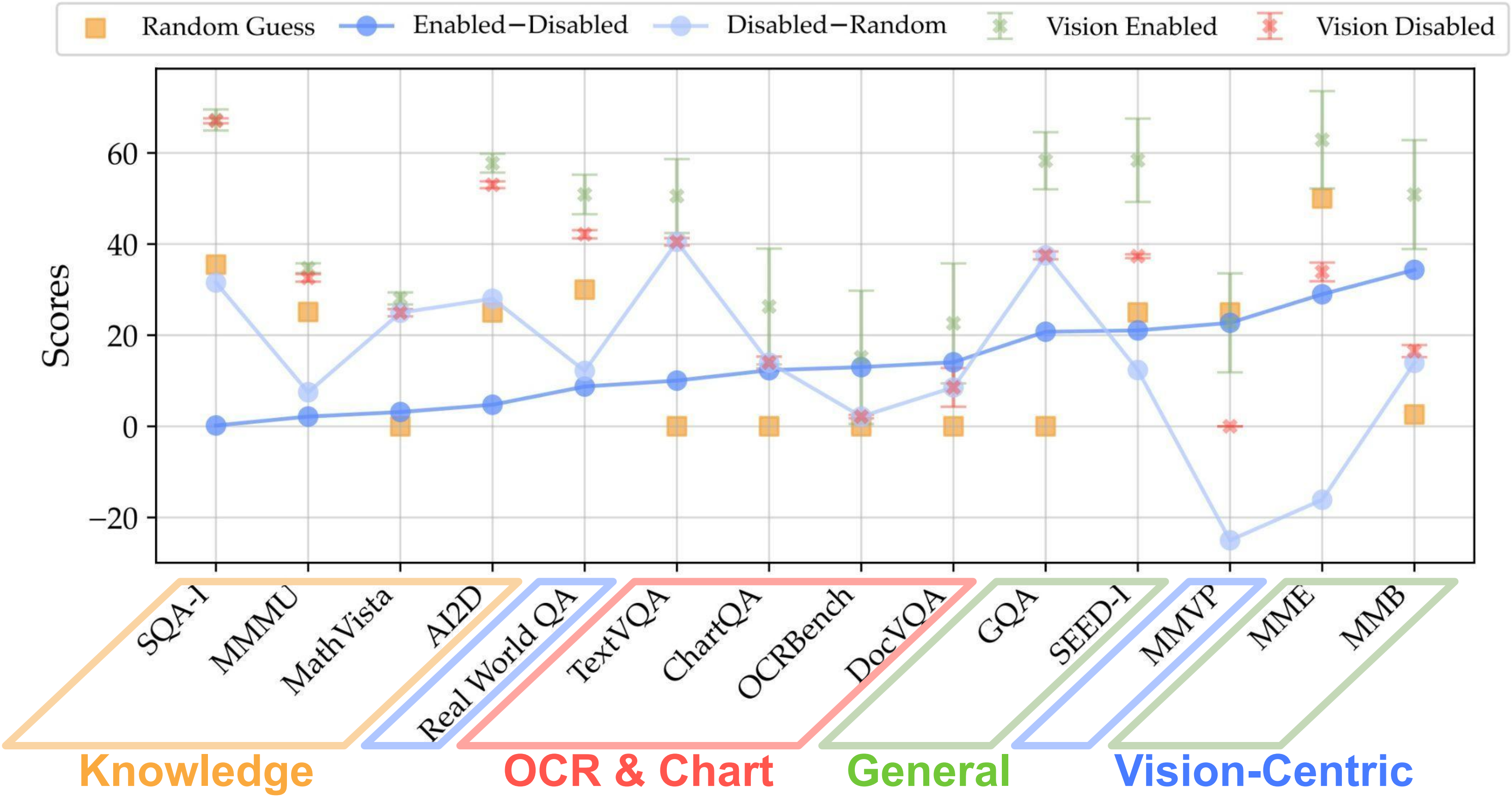
If two benchmarks evaluate on similar domains, they should have a strong correlation



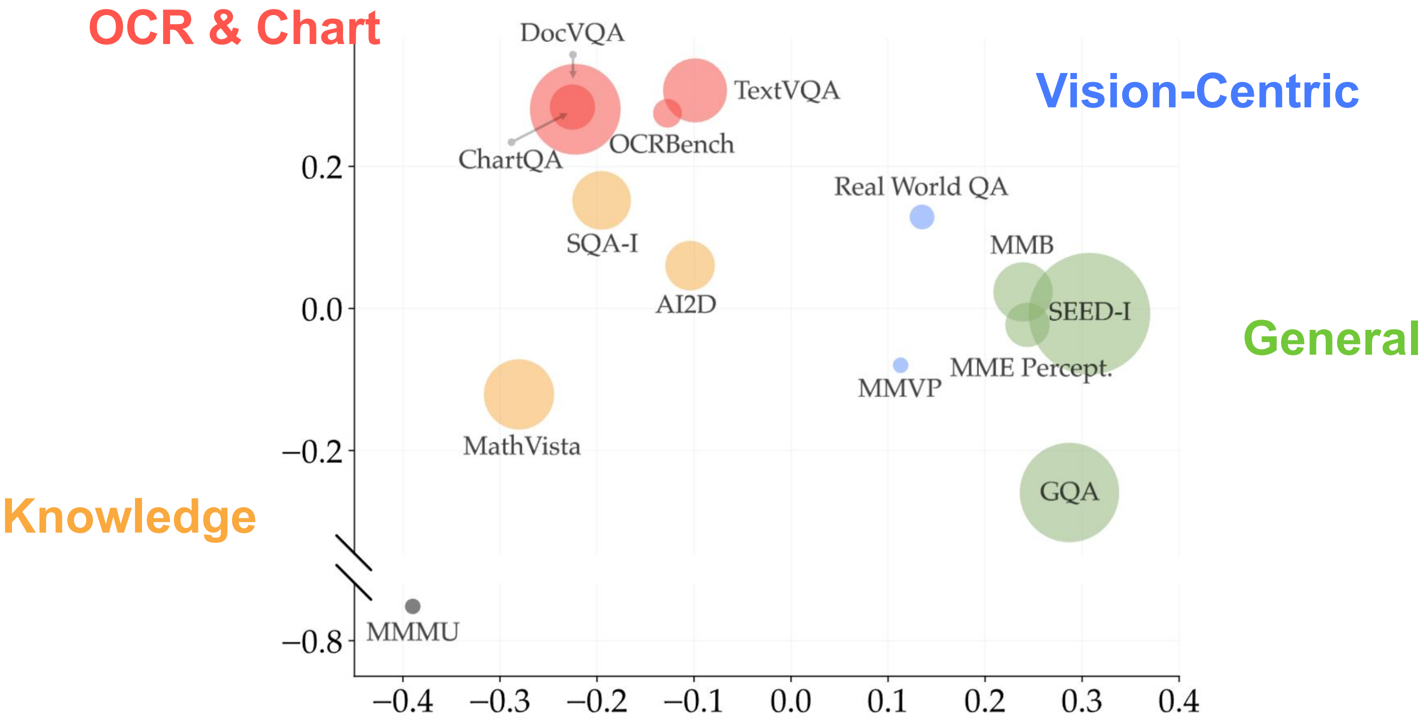
Group Benchmarks by Correlation



Who's answering the question: the **LLM** or **MLLM**?

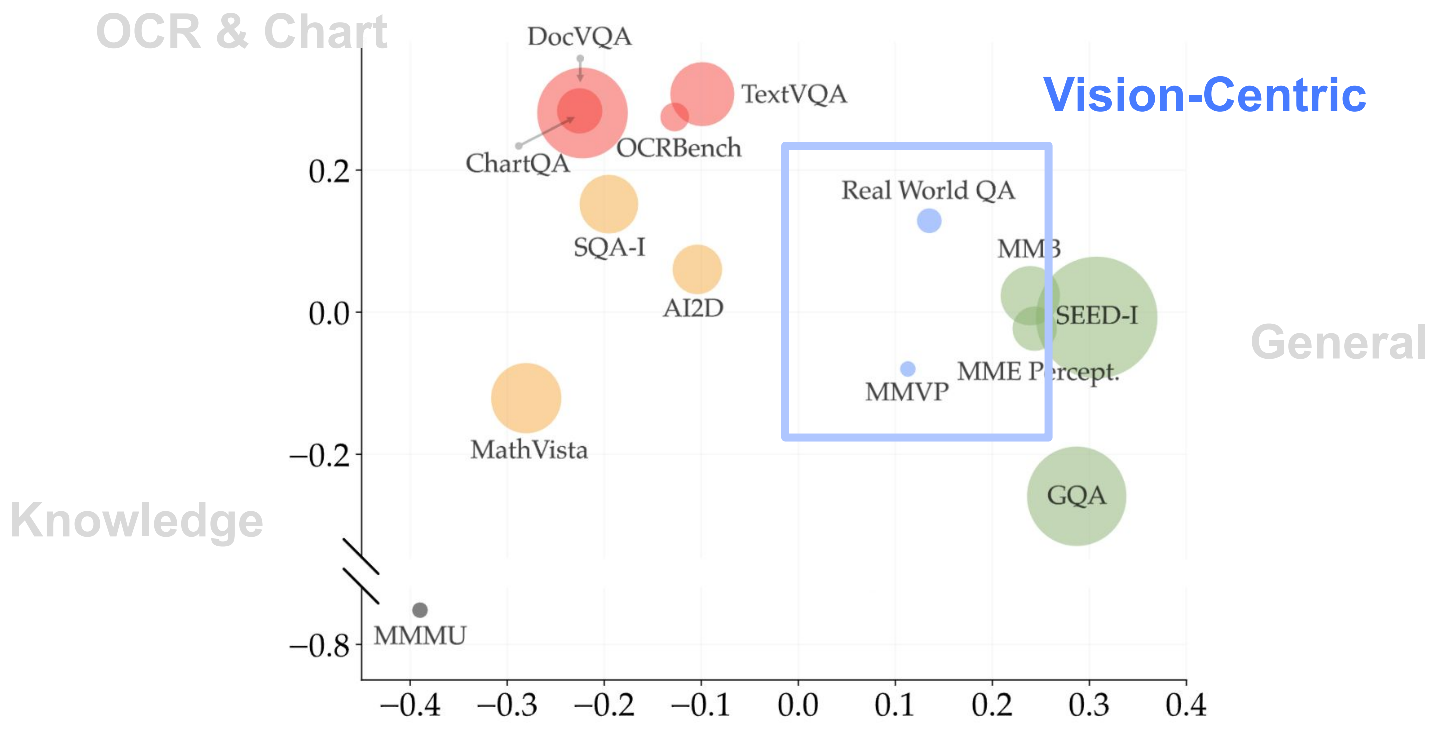


Group Benchmarks by Correlation



Group Benchmarks by Correlation

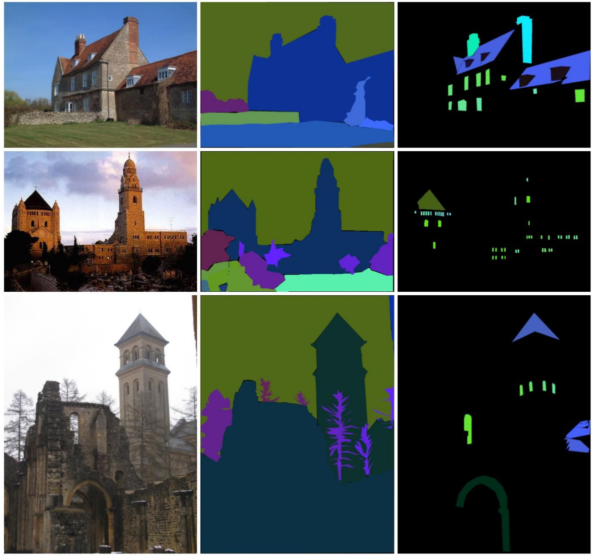
Tiny compared to others!



Q: How can we scalably generate ***vision-centric*** MLLM evaluations?

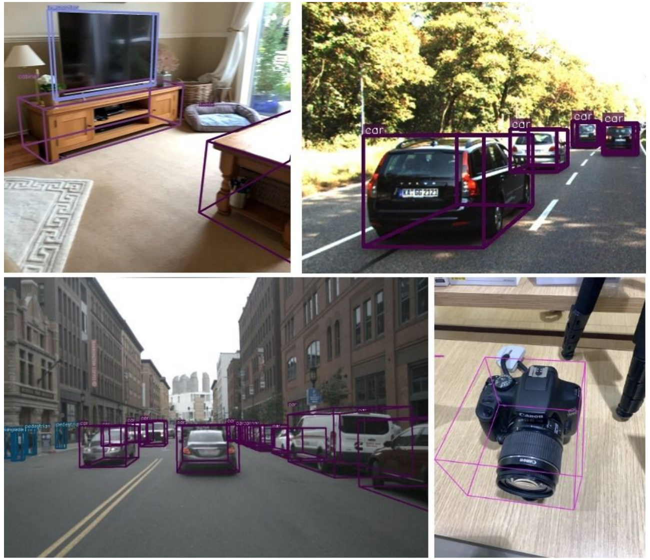
💡 Repurpose existing vision datasets!

CV-Bench



ADE20K

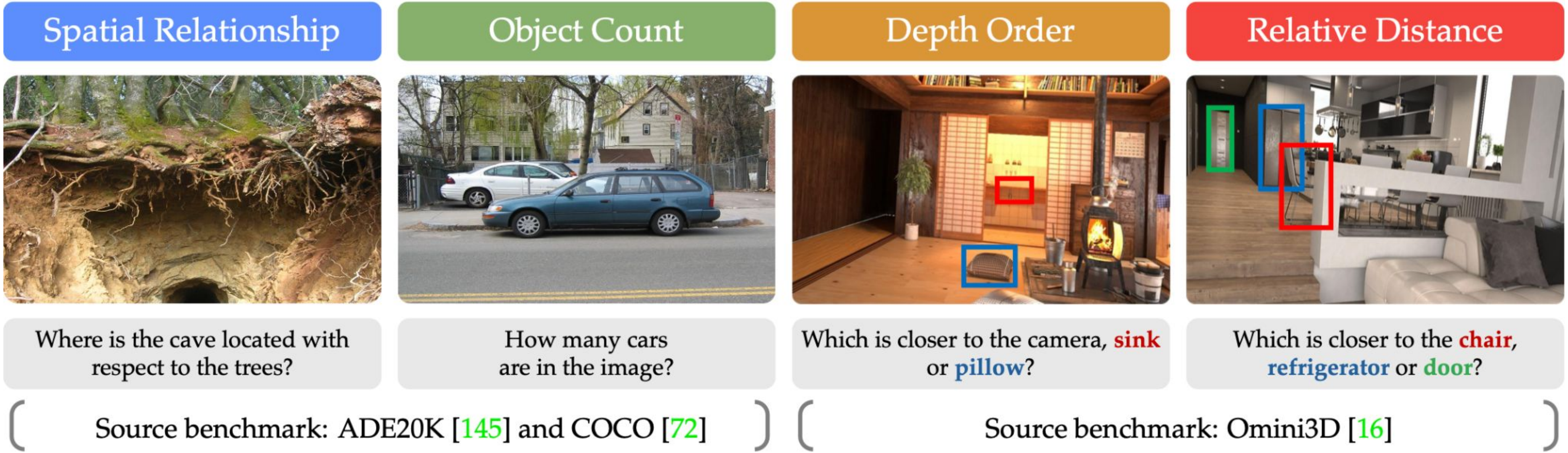
MSCOCO



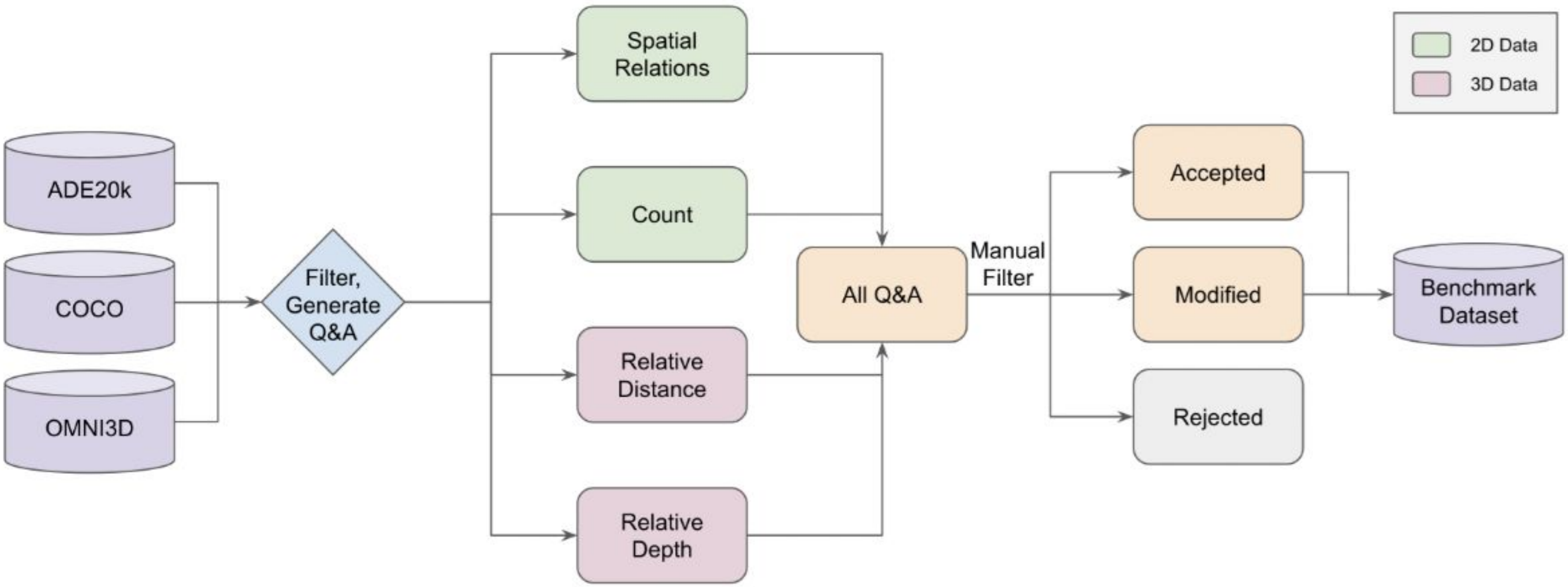
Omni3D

CV-Bench

2D ***3D***



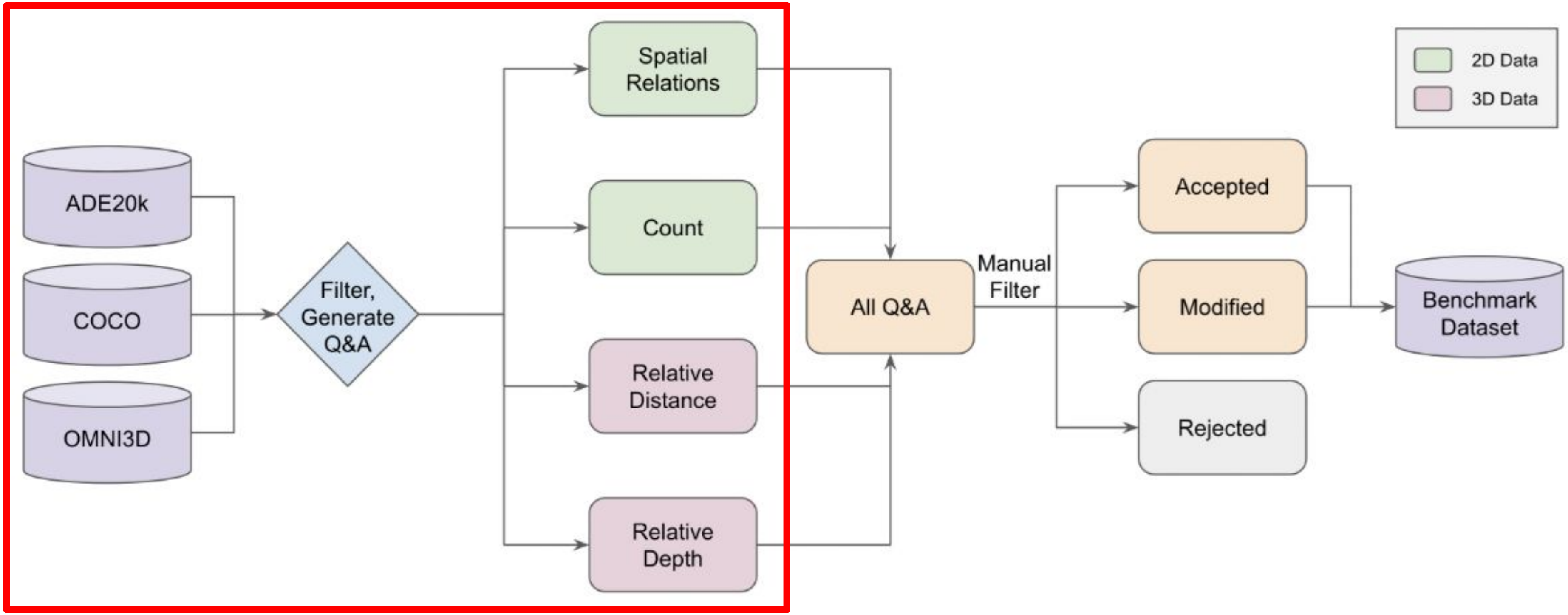
CV-Bench



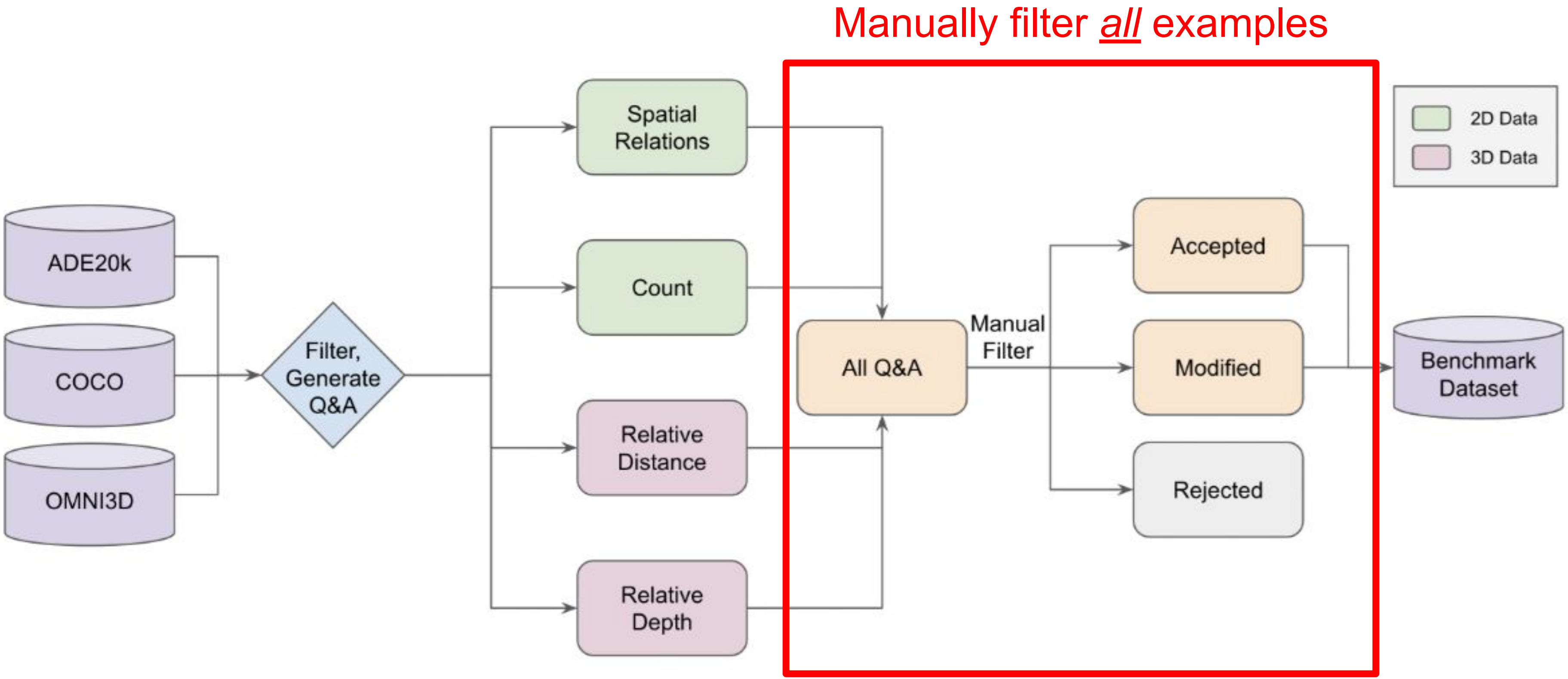
2D Data
3D Data
ADE20k
COCO
OMNI3D
Filter, Generate Q&A
Spatial Relations
Count
Relative Distance
Relative Depth
All Q&A
Manual Filter
Accepted
Modified
Rejected
Benchmark Dataset

CV-Bench

Programmatically construct VQA questions using GT annos



CV-Bench

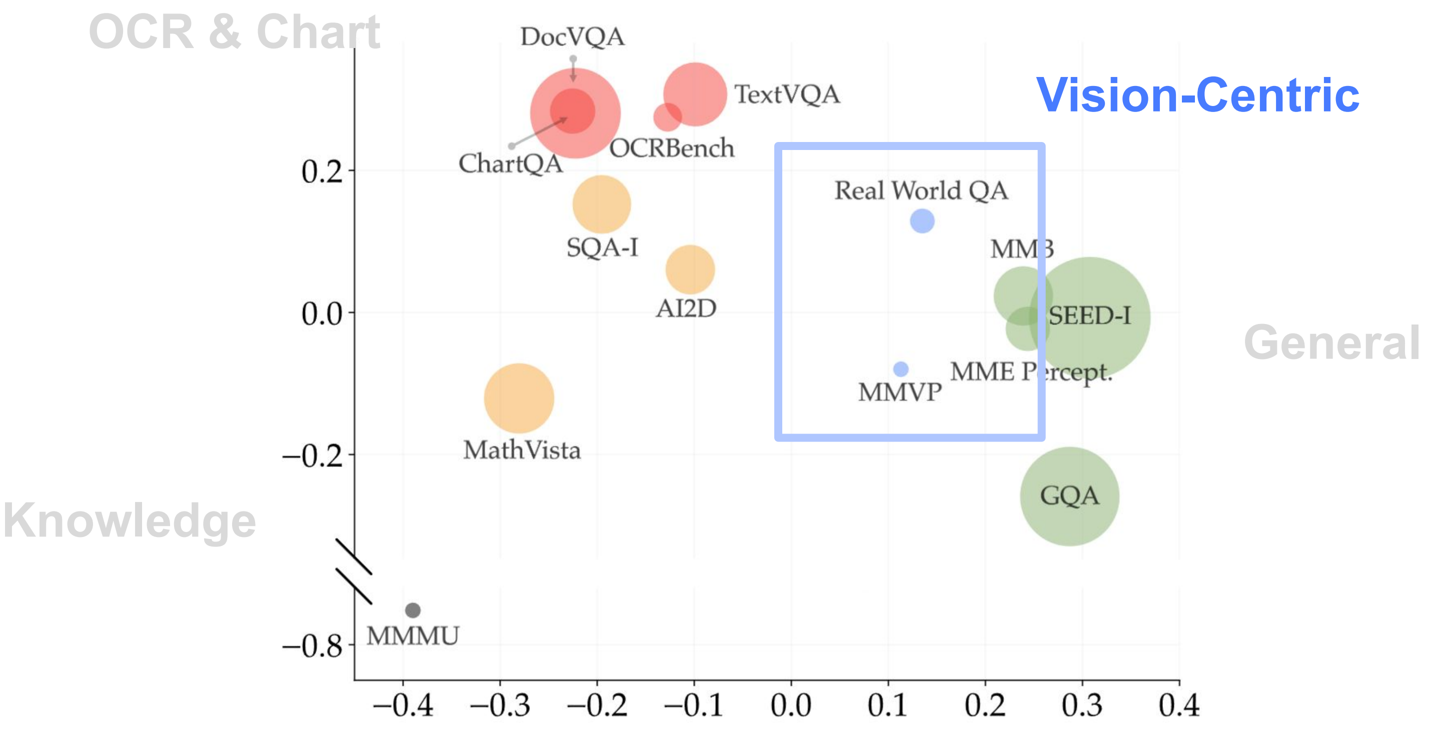

Manually filter *all* examples
ADE20k
COCO
OMNI3D
Filter, Generate Q&A
Spatial Relations
Count
Relative Distance
Relative Depth
All Q&A
Manual Filter
Accepted
Modified
Rejected
Benchmark Dataset
2D Data
3D Data

CV-Bench

2,638 manually-inspected examples

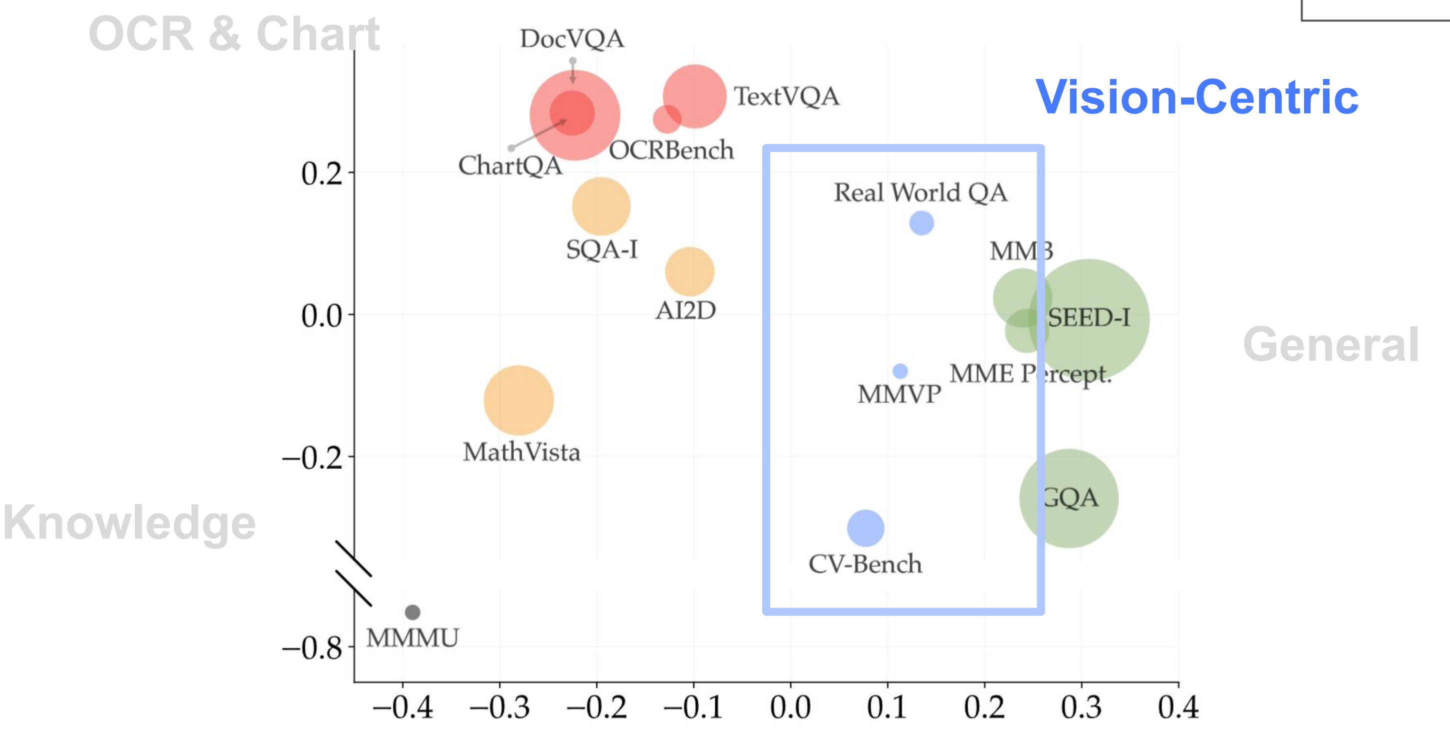
Type	Task	Description	Sources	# Samples
2D	Spatial Relationship	Determine the relative position of an object w.r.t. the anchor object. Consider left-right or top-bottom relationship.	ADE20K COCO	650
	Object Count	Determine the number of instances present in the image.	ADE20K COCO	788
3D	Depth Order	Determine which of the two distinct objects is closer to the camera.	Omni3D	600
	Relative Distance	Determine which of the two distinct objects is closer to the anchor object.	Omni3D	600

Group Benchmarks by Correlation



Group Benchmarks by Correlation

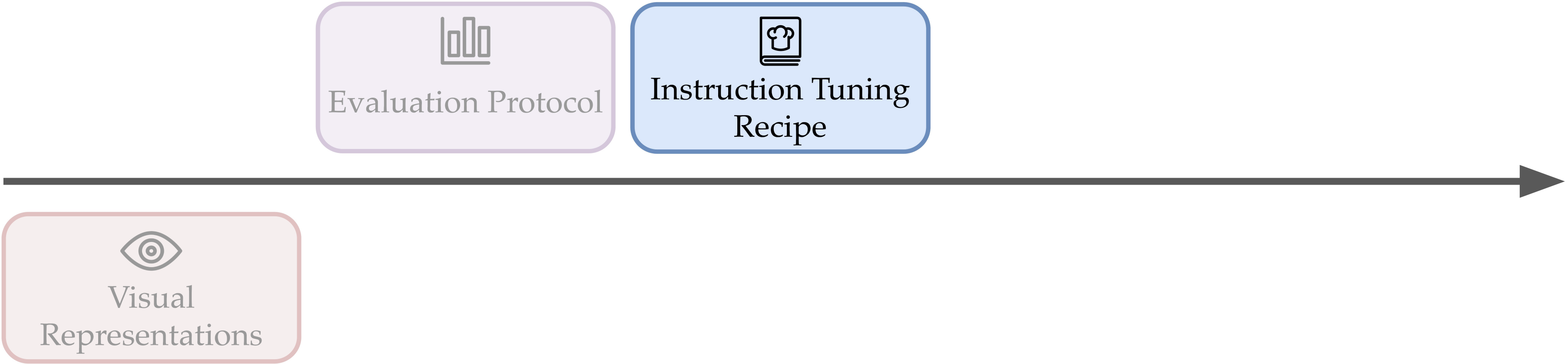
3.5x more vision-centric examples!



Overview


Evaluation Protocol
Instruction Tuning Recipe
Visual Representations

Overview

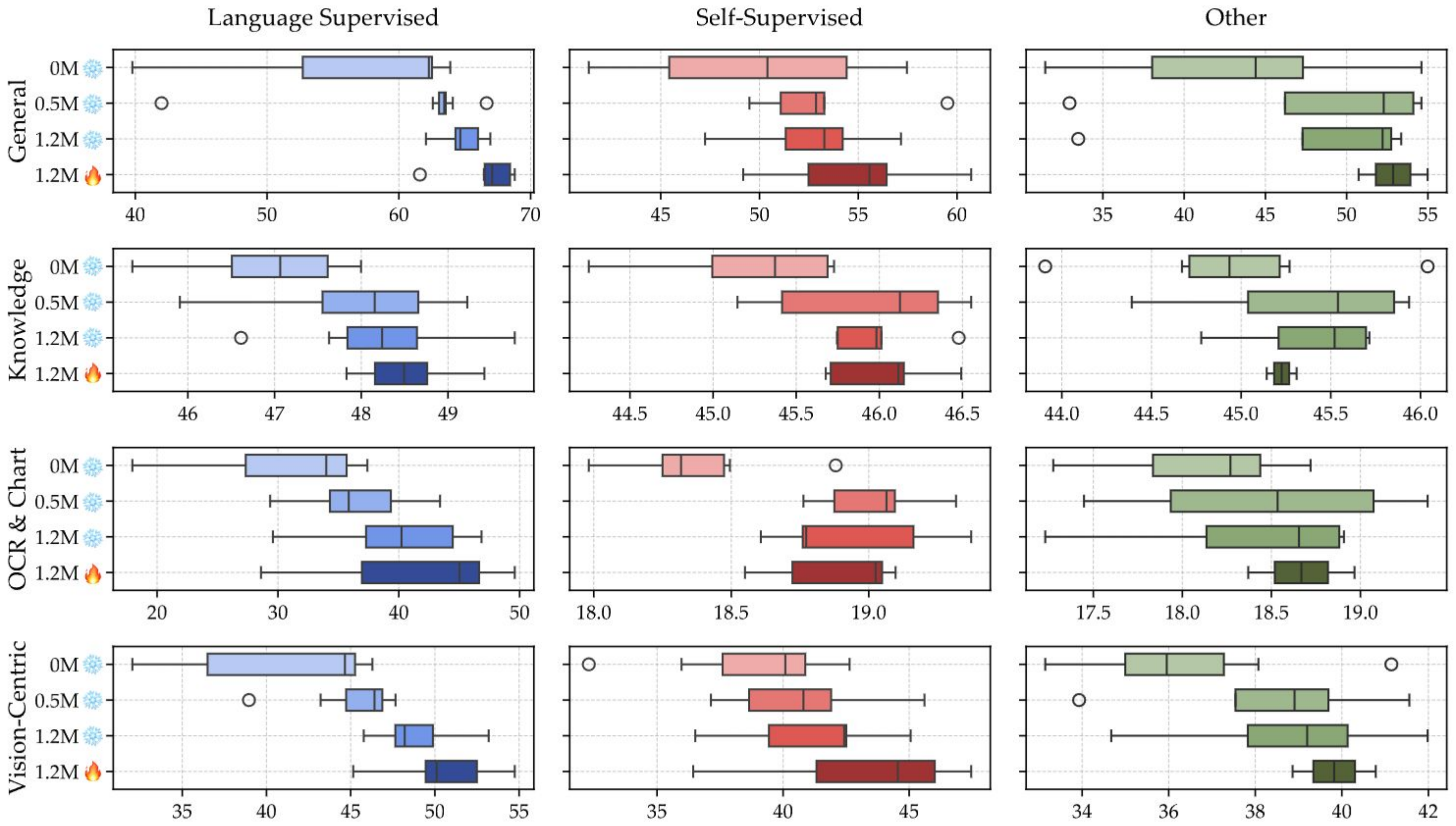

Evaluation Protocol
Instruction Tuning Recipe
Visual Representations

Instruction Tuning Recipe

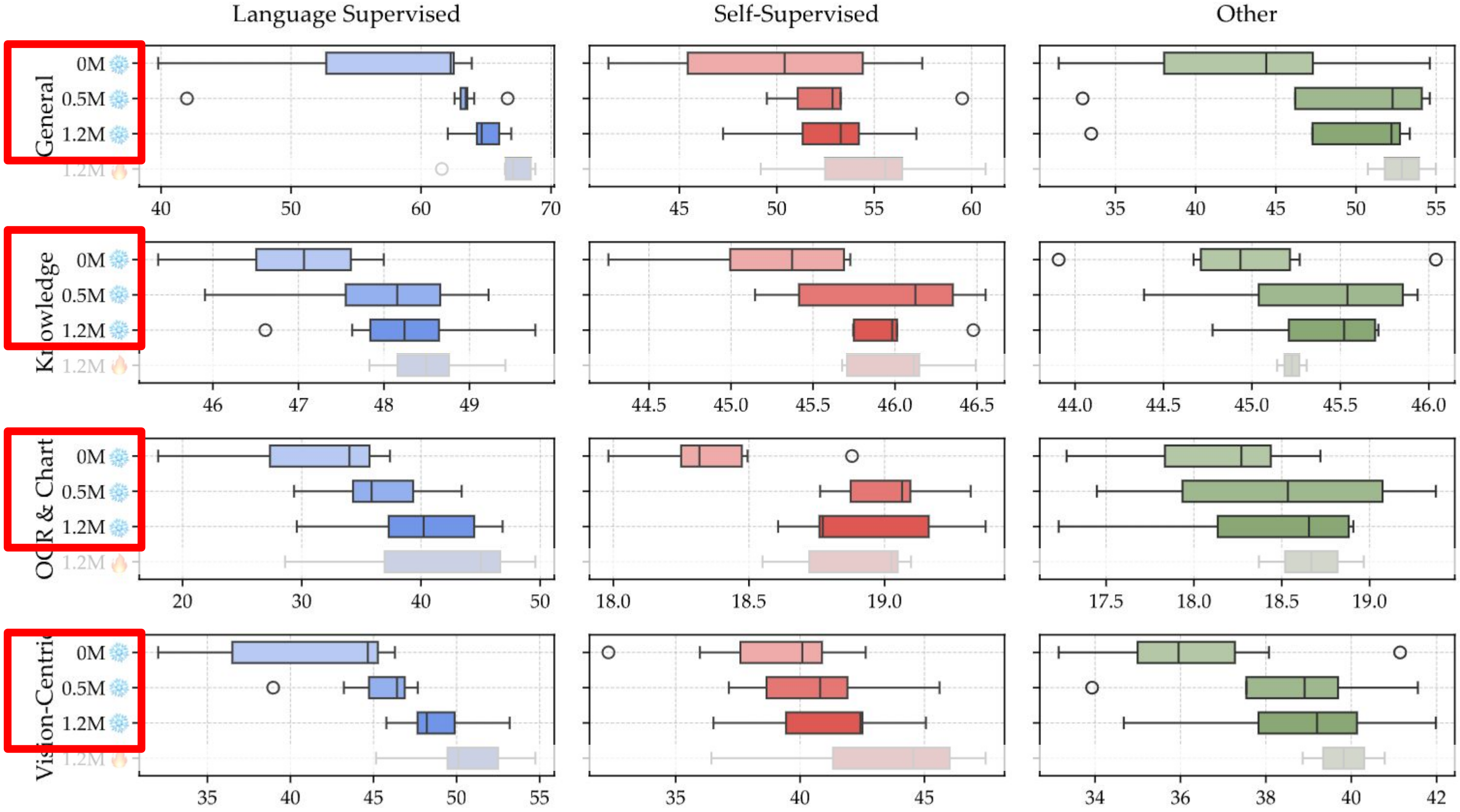
- 1 Stage or 2-Stage Training
    - Training Connector first with Alignment Data?

- Freeze or Unfreeze Vision Backbone

Instruction Tuning Recipe



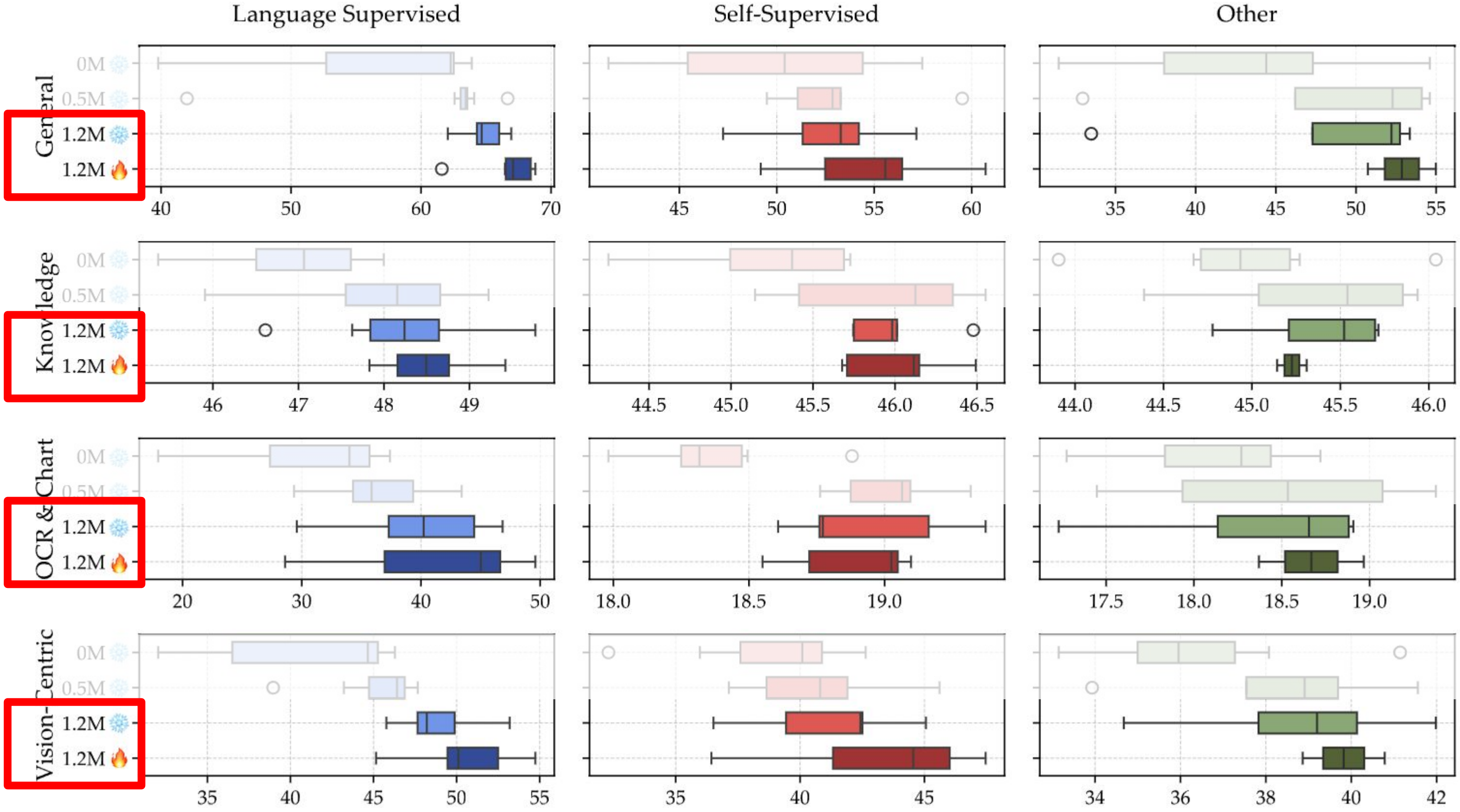
Instruction Tuning Recipe



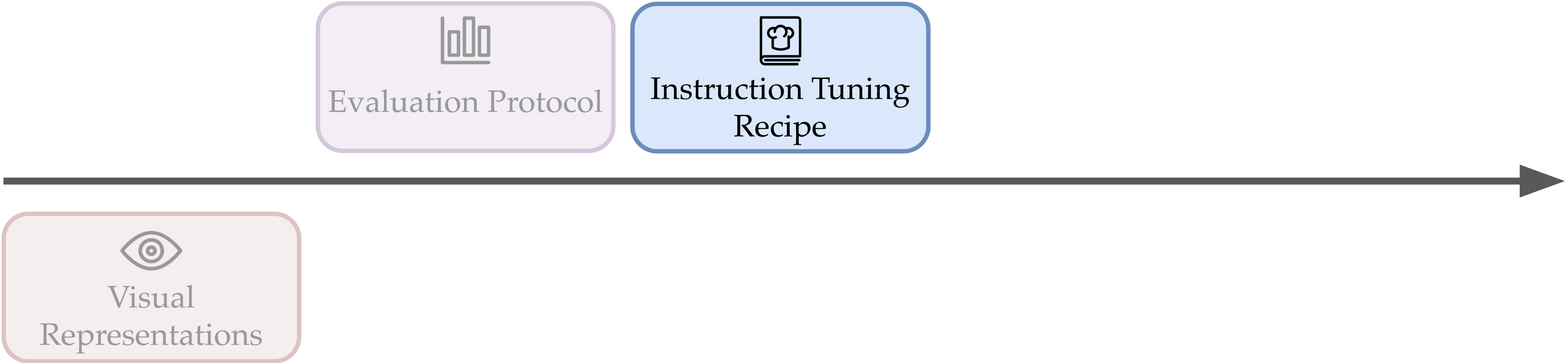
More Alignment Data helps!

Instruction Tuning Recipe

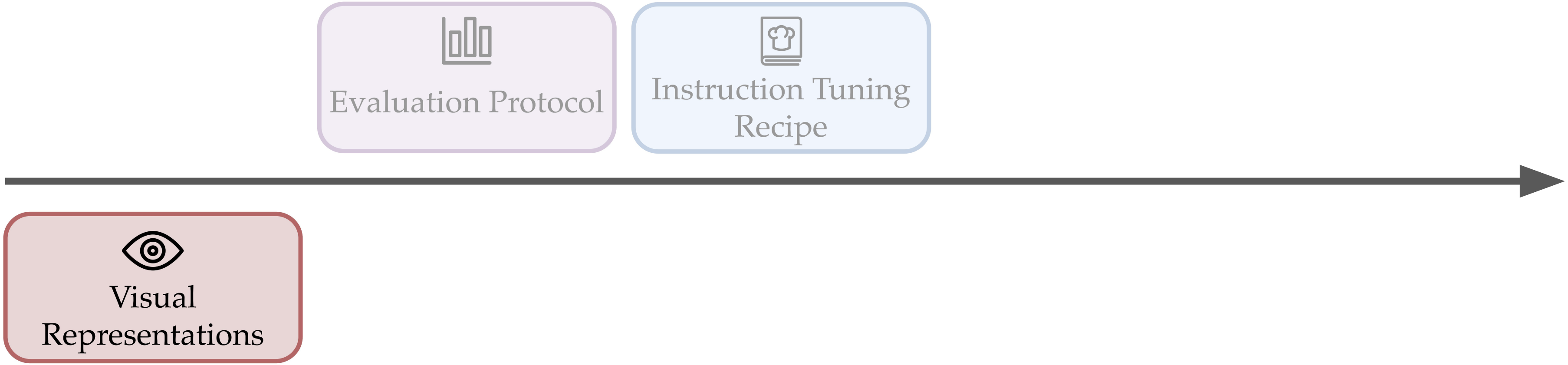
Unfreezing Vision Encoder Helps 🔥



Overview


Evaluation Protocol
Instruction Tuning Recipe
Visual Representations

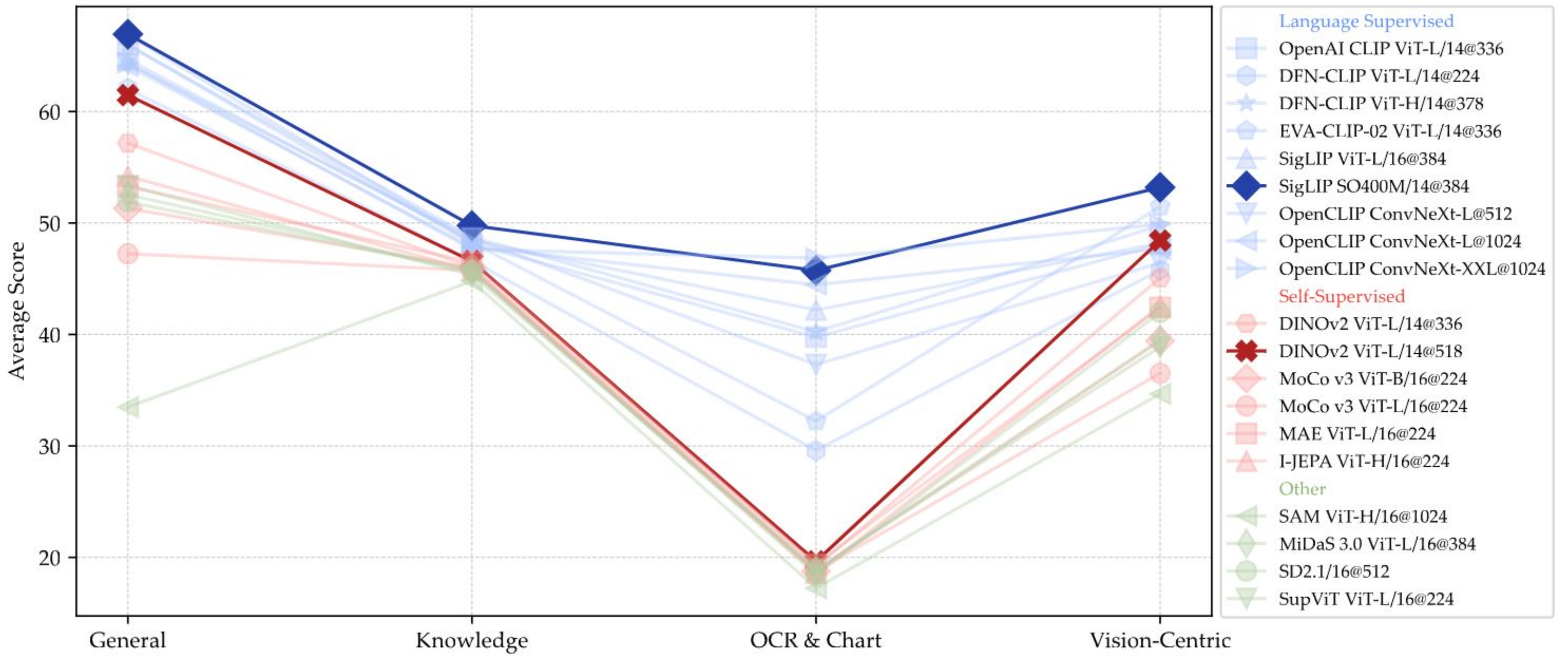
Overview



Evaluation Protocol
Instruction Tuning Recipe
Visual Representations

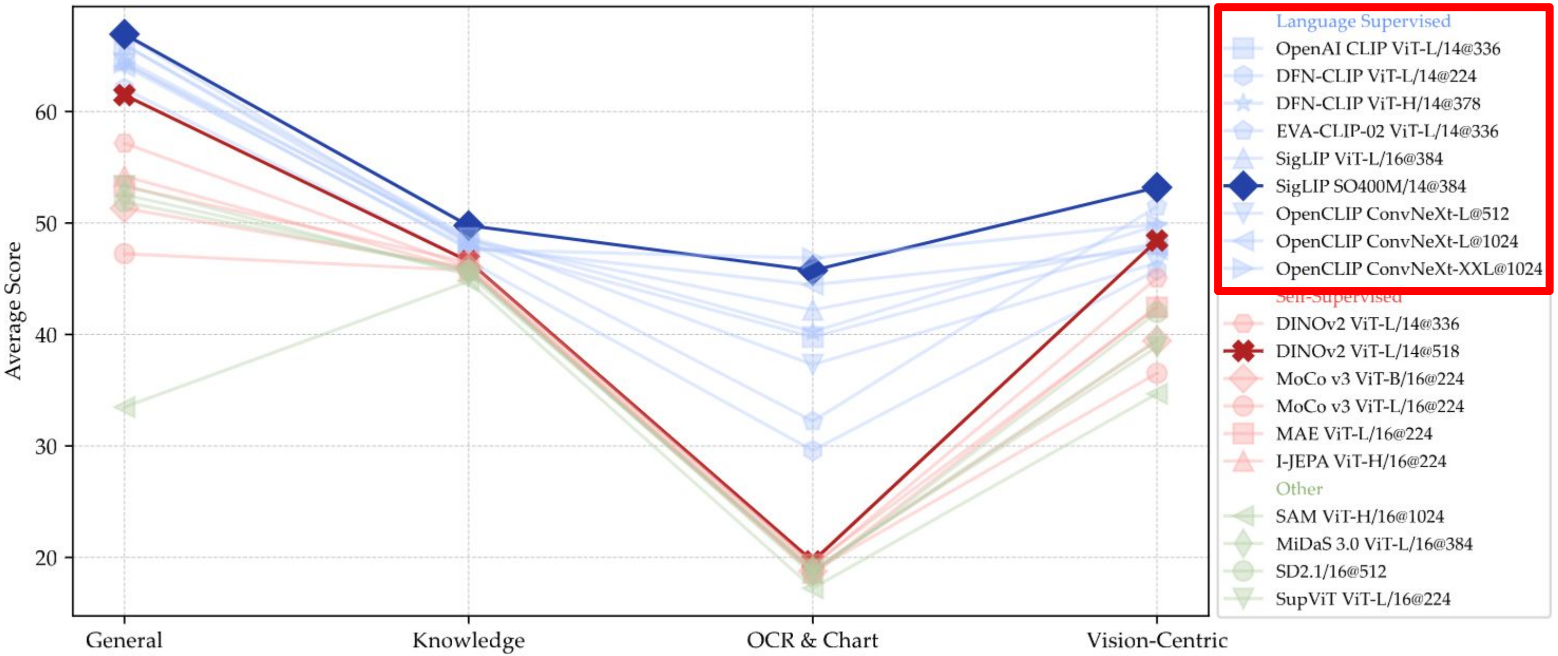
Visual Representation

❄️ **1.2M**



Visual Representation

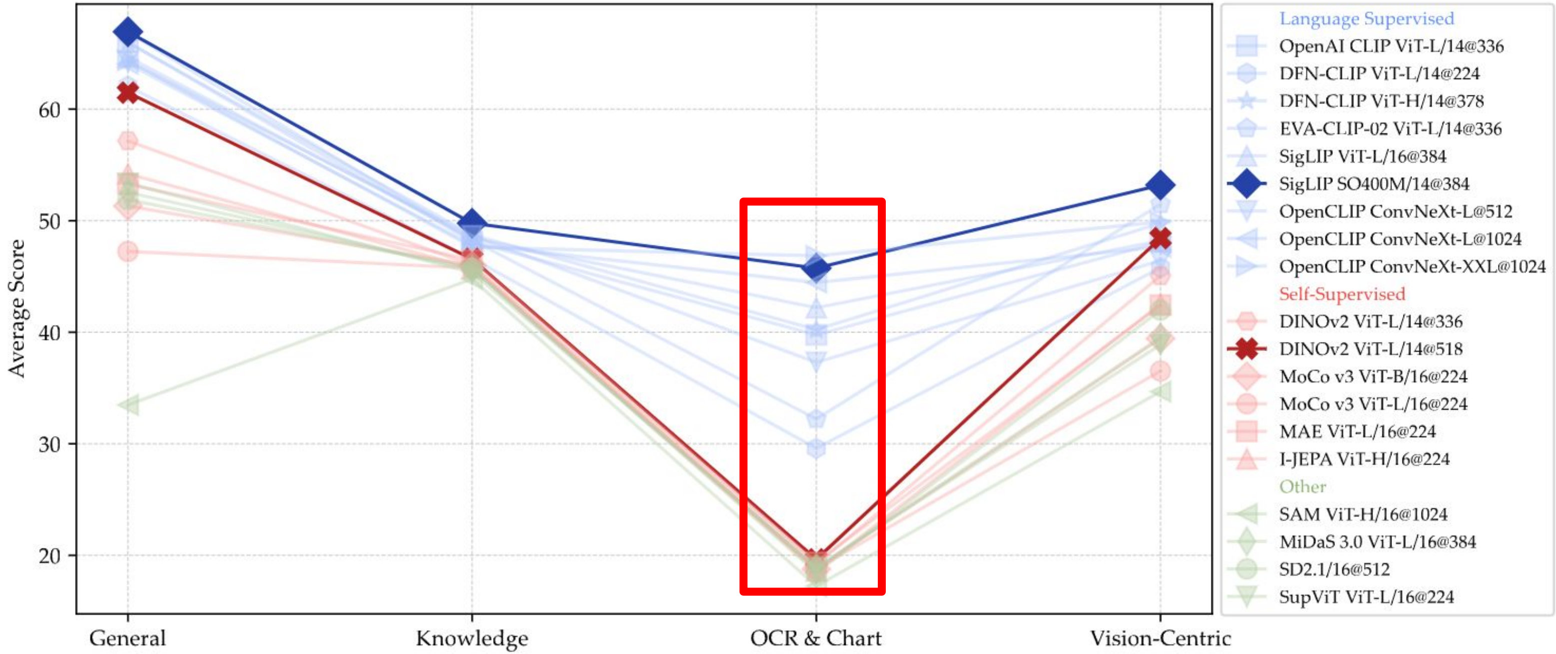
#1 Language Supervised Models are better



Visual Representation

#1 Language Supervised Models are better

#2 Gap is largest in OCR & Chart

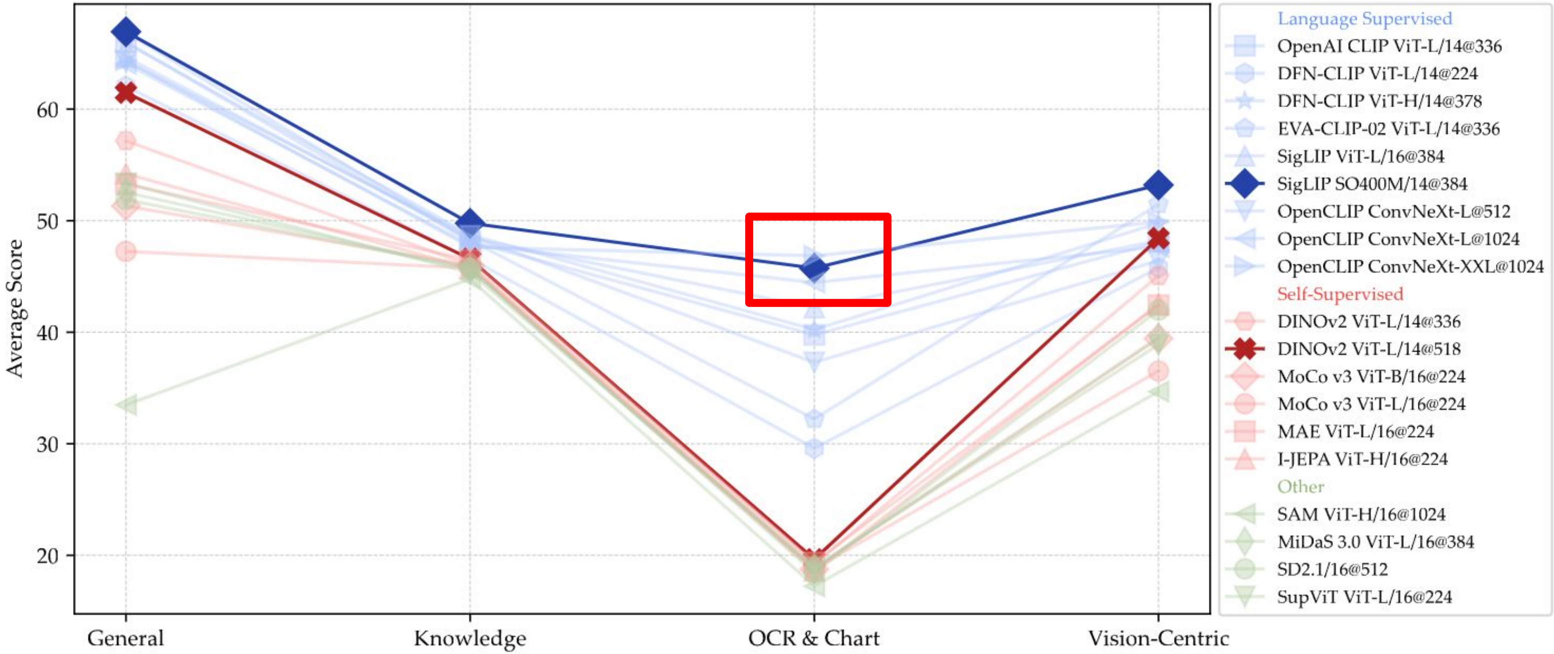


Visual Representation

#1 Language Supervised Models are better

#2 Gap is largest in OCR & Chart

#3 ConvNets are good at OCR



Visual Representation

#1 Language Supervised Models are better

#2 Gap is largest in OCR & Chart

#3 ConvNets are good at OCR

Model	Architecture	All	G	K	O	V
SigLIP	ViT-SO400M/14@384	1	1	1	2	1
OpenCLIP	ConvNeXt-XXL@1024	2	6	8	1	3
DFN-CLIP	ViT-H/14@378	3	4	2	5	4
OpenCLIP	ConvNeXt-L@1024	4	8	7	3	8
SigLIP	ViT-L/16@384	5	5	4	4	6
OpenAI CLIP	ViT-L/14@336	6	3	6	6	7
EVA-CLIP-02	ViT-L/14@336	7	2	5	8	2
OpenCLIP	ConvNeXt-L@512	8	7	3	7	9
DFN-CLIP	ViT-L/14@224	9	9	9	9	10
DINOv2*	ViT-L/14@518	10	10	10	10	5

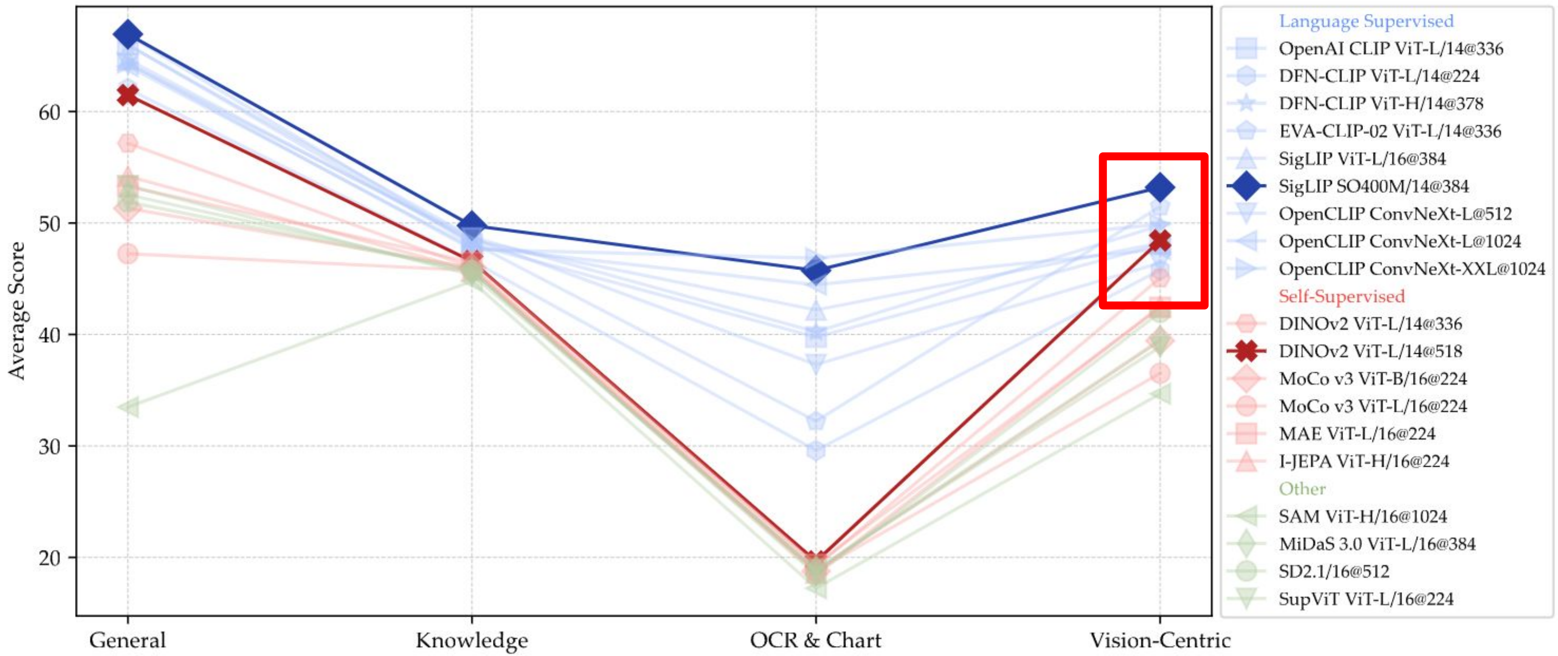
Visual Representation

#1 Language Supervised Models are better

#2 Gap is largest in OCR & Chart

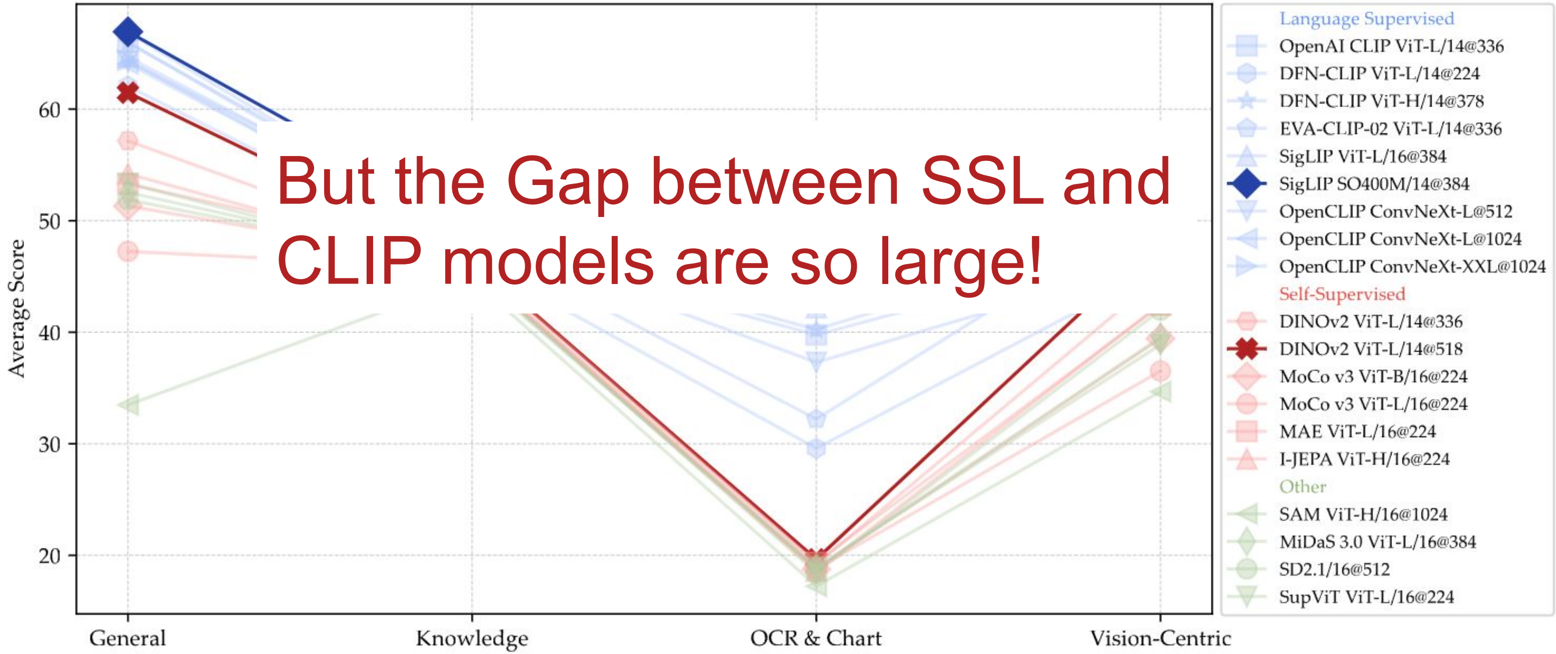
#3 ConvNets are good at OCR

#4 Best SSL Model good at vision-centric



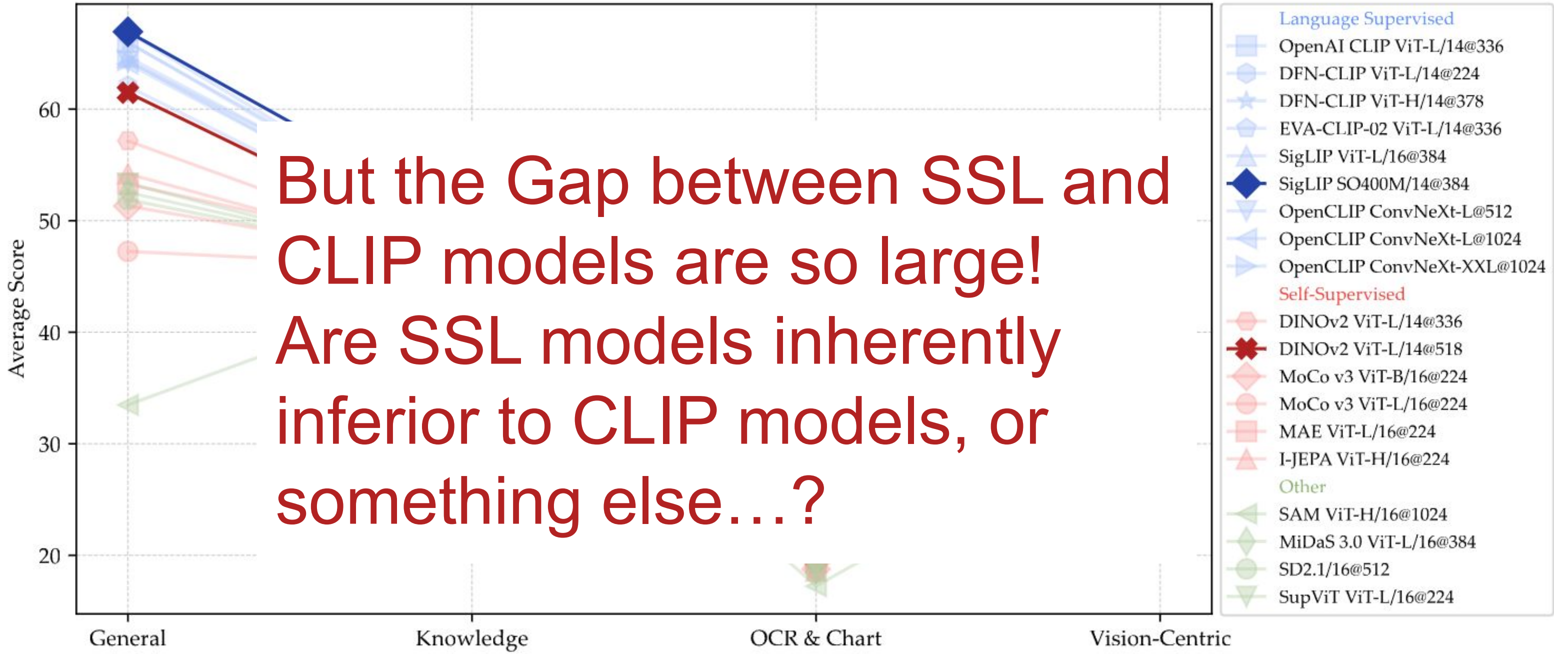
Visual Representation

#1 Language Supervised Models are better
#2 Gap is largest in OCR & Chart
#3 ConvNets are good at OCR
#4 Best SSL Model good at vision-centric



Visual Representation

#1 Language Supervised Models are better
#2 Gap is largest in OCR & Chart
#3 ConvNets are good at OCR
#4 Best SSL Model good at vision-centric



But the Gap between SSL and CLIP models are so large! Are SSL models inherently inferior to CLIP models, or something else…?

Scaling **Language-Free** Visual Representation Learning

David Fan*, Shengbang Tong*, Jiachen Zhu, Koustuv Sinha, Zhuang Liu,
Xinlei Chen, Michael Rabbat, Nicolas Ballas, Yann LeCun, Amir Bar†, Saining Xie†

FAIR, Meta, New York University, Princeton University

Visual Representation Learning

Self-Supervision

- MoCo, MAE, DINO
- Learn from images itself (augmentation, masking)
- Train on ImageNet-Like Data **(million scale to hundred million scale)**
- Good at <u>classification</u>, segmentation, depth estimation, etc

Language-Supervision:

- CLIP, SigLIP, MetaCLIP
- Learn from language that “describe the text”
- Train on Image-Text pairs crawled from the internet **(400 million to 100 billion)**
- Good at <u>classification</u>, and widely used at backbone for multimodal models

Visual Representation Learning

Self-Supervision

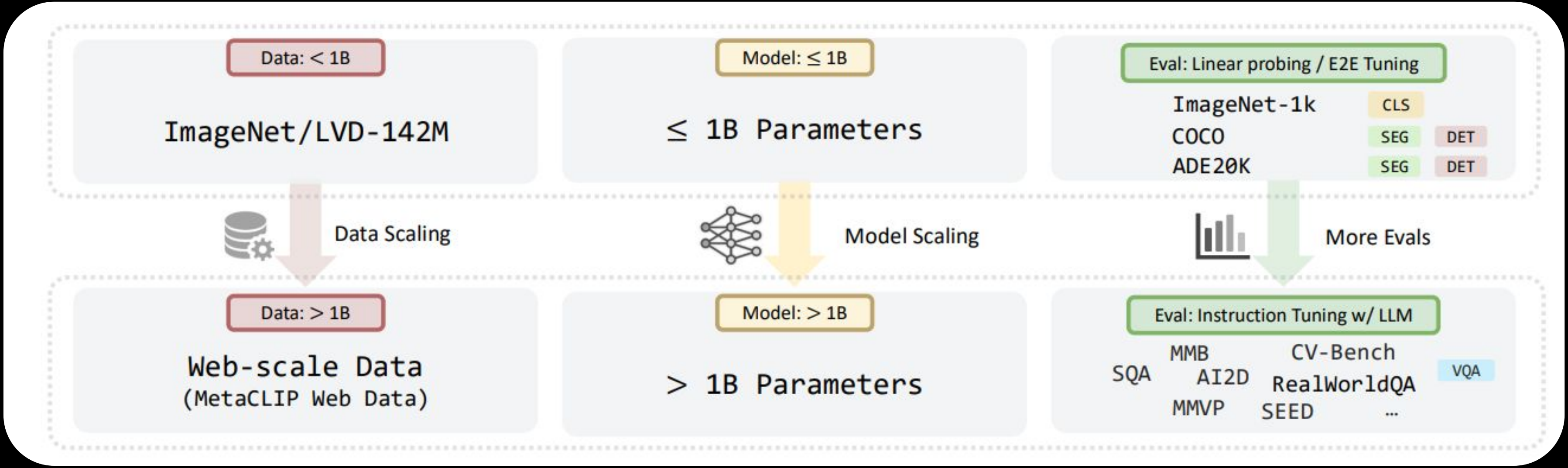
- MoCo, MAE, DINO
- Learn from images itself (augmentation, masking)
- Train on ImageNet-Like Data **(million scale to hundred million scale)**
- Good at classification, segmentation, depth estimation, etc

Language-Supervision:

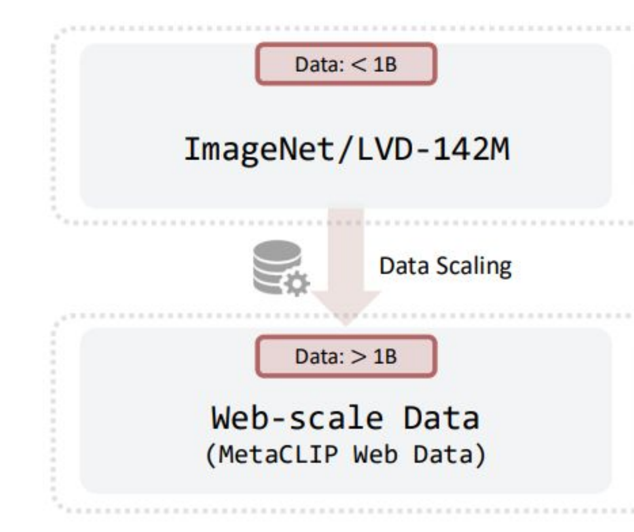
- CLIP, SigLIP, MetaCLIP
- Learn from language that “describe the text”
- Train on Image-Text pairs crawled from the internet **(400 million to 100 billion)**
- Good at classification, and widely used at backbone for multimodal models

Datasize is at least 10x smaller!

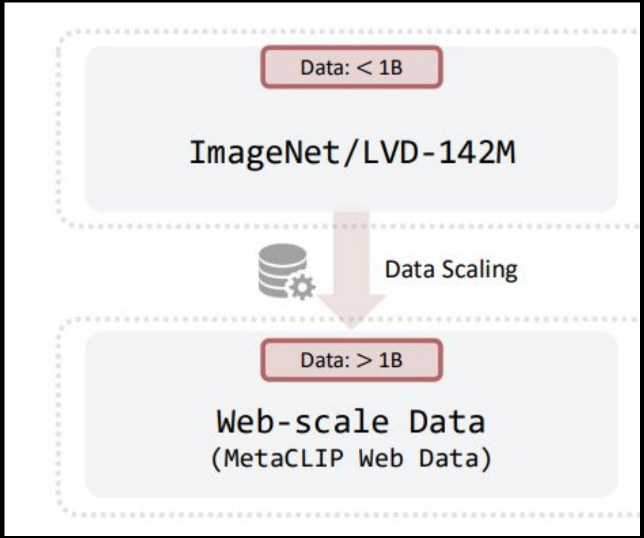
Visual SSL in This New Era



Visual SSL in This New Era

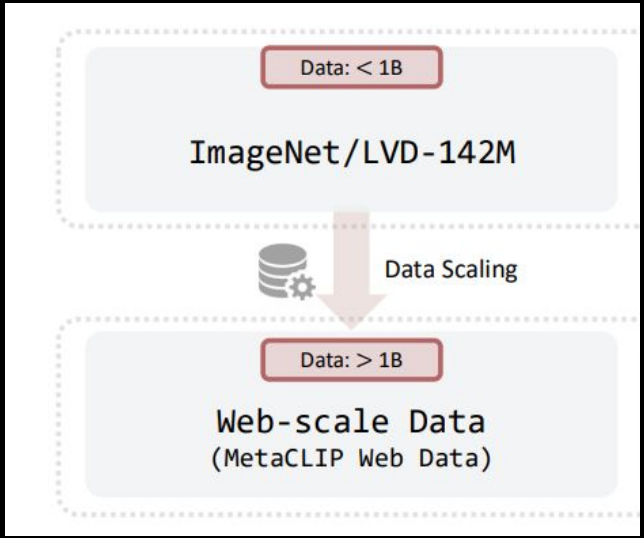


Visual SSL in This New Era



ImageNet/LVD-142M: **Million scale** ImageNet or ImageNet-like distribution, mostly natural images

Visual SSL in This New Era



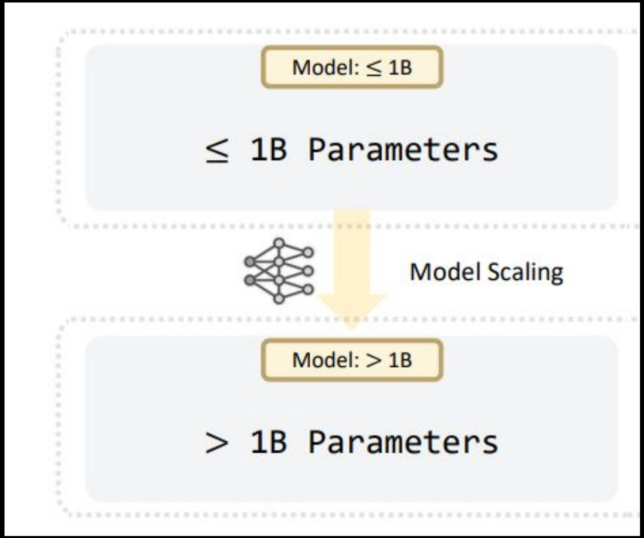
ImageNet/LVD-142M: **Million scale** ImageNet or ImageNet-like distribution, mostly natural images

——>

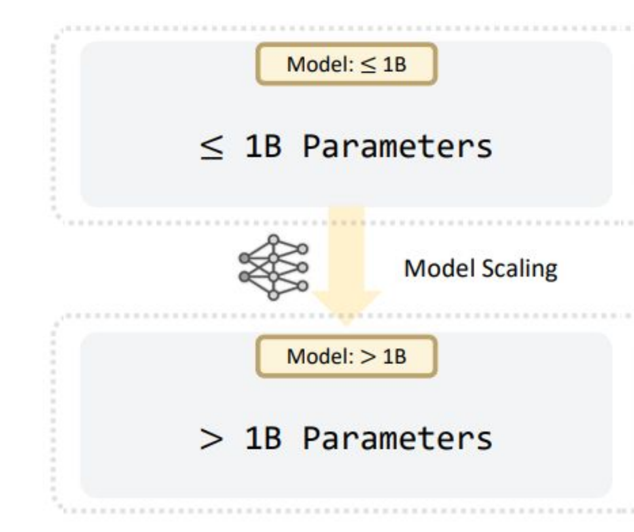
Web-Scale Images (e.g. MetaCLIP): **Billion scale** diverse “random” images from the internet

MetaCLIP-2B: *MC-2B*

Visual SSL in This New Era

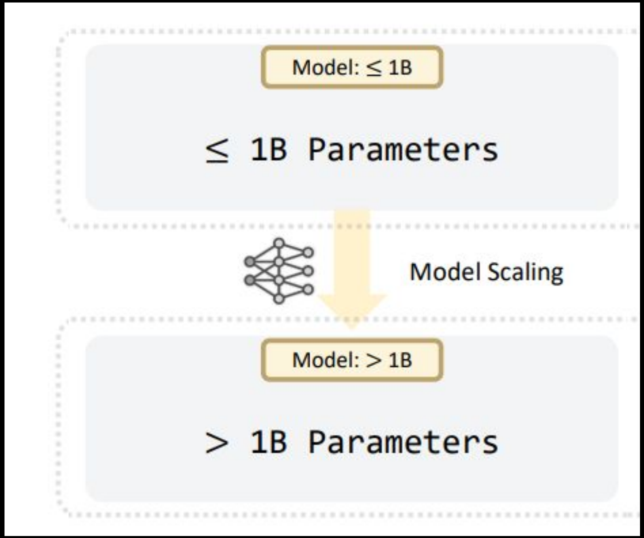


Visual SSL in This New Era



Less than 1B params: ViT-Base, ViT-Large, ViT-Huge, …

Visual SSL in This New Era

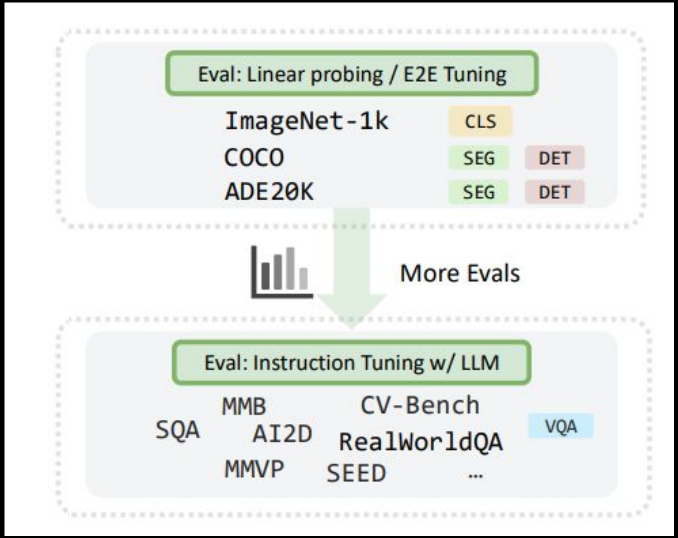


Less than 1B params: ViT-Base, ViT-Large, ViT-Huge, …

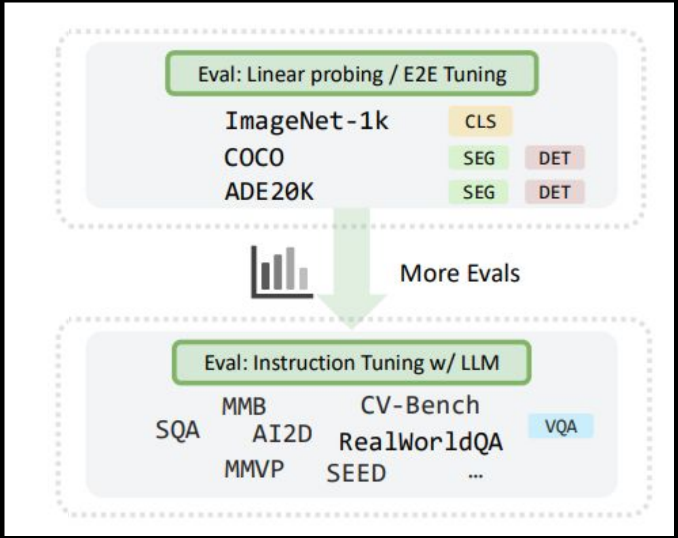
——>

More than 1B params: ViT-1B, ViT-2B, ViT-3B, ViT-5b, …

Visual SSL in This New Era

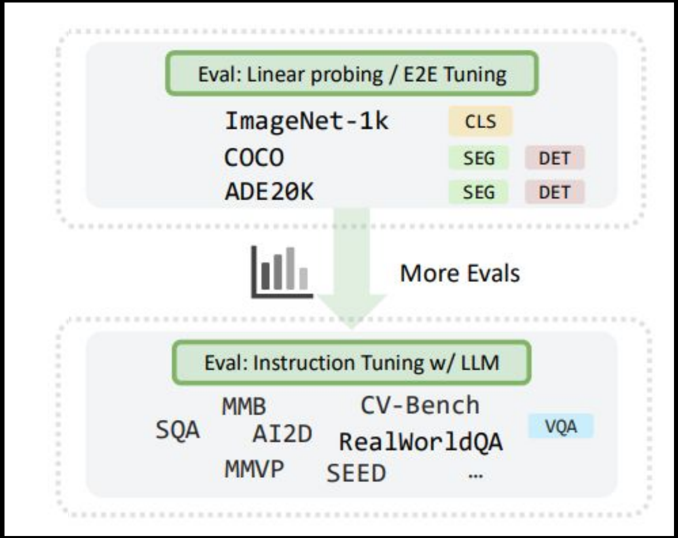


Visual SSL in This New Era



Classic vision eval: classification, segmentation, depth estimation, etc

Visual SSL in This New Era

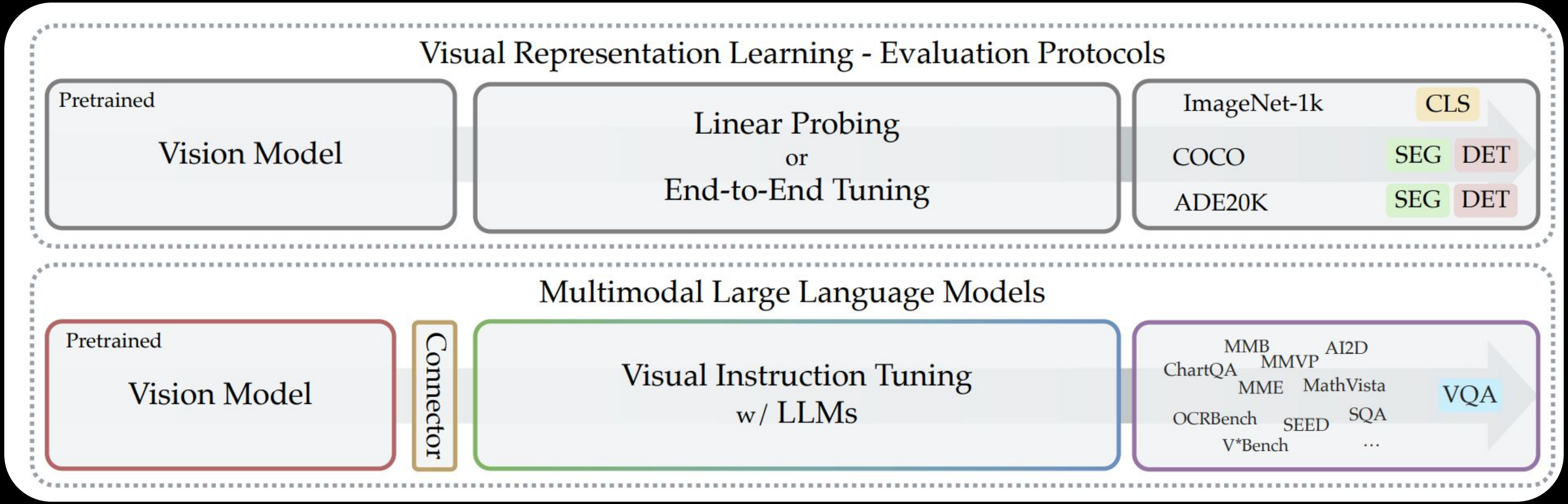


Classic vision eval: classification, segmentation, depth estimation, etc

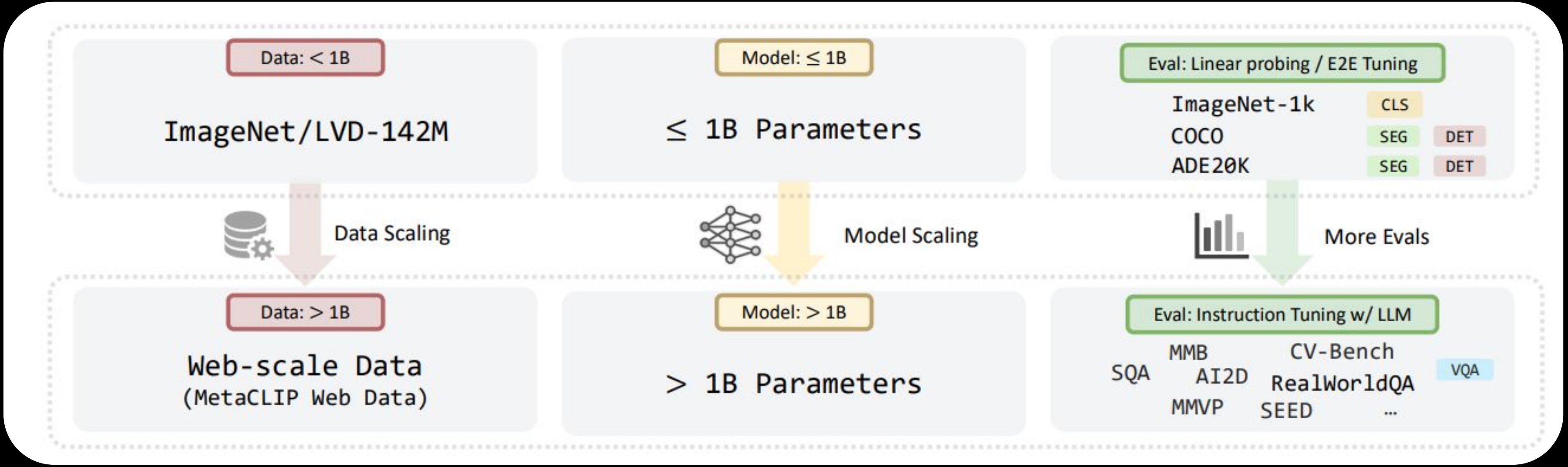
——>

Using VQA as eval: diverser question, more than classic vision tasks

Evaluation Setup



We use Cambrian with *frozen* vision encoder (but finetuned adapter + LLM) to evaluate on VQA tasks: **General, Knowledge, OCR&Chart, Vision-Centric**


Data: < 1B
ImageNet/LVD-142M
Model: ≤ 1B
≤ 1B Parameters
Eval: Linear probing / E2E Tuning
ImageNet-1k CLS
COCO SEG DET
ADE20K SEG DET
Data Scaling
Model Scaling
More Evals
Data: > 1B
Web-scale Data
(MetaCLIP Web Data)
Model: > 1B
> 1B Parameters
Eval: Instruction Tuning w/ LLM
MMB CV-Bench
SQA AI2D RealWorldQA VQA
MMVP SEED ...

Let's Scale!

Scaling Up Model

Scaling Up Model

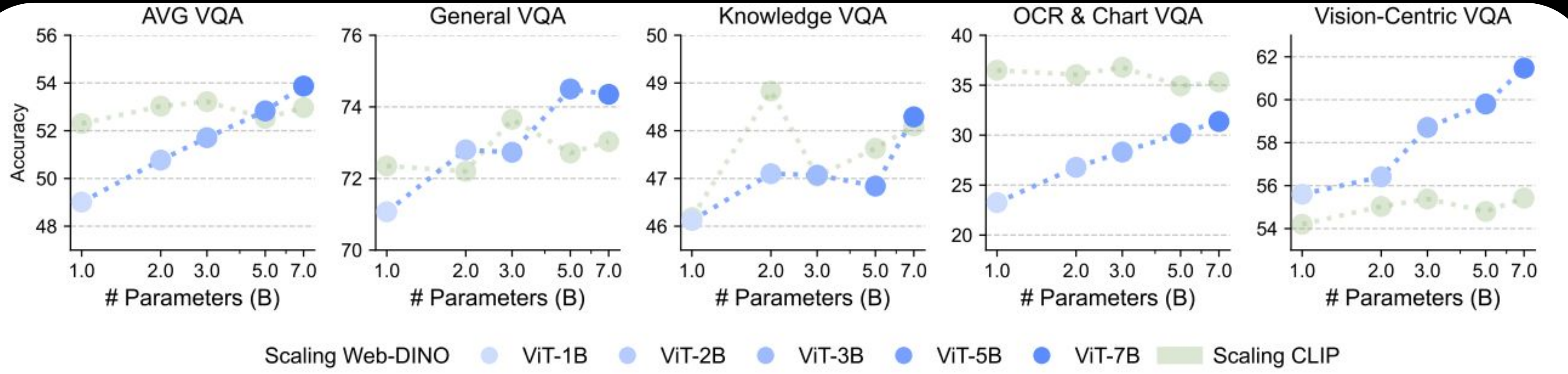
Data: *MC-2B*, 2 billion samples seen

Model: ViT-1B, ViT-2B, ViT-3B, ViT-5B, ViT-7B

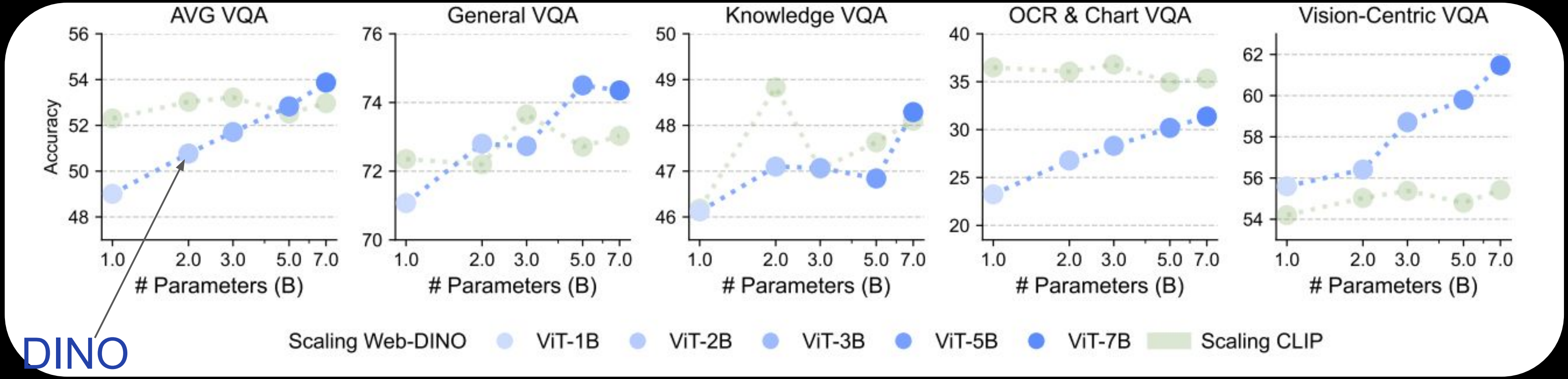
Method: DINOv2

Eval: Use VQA as evaluation.

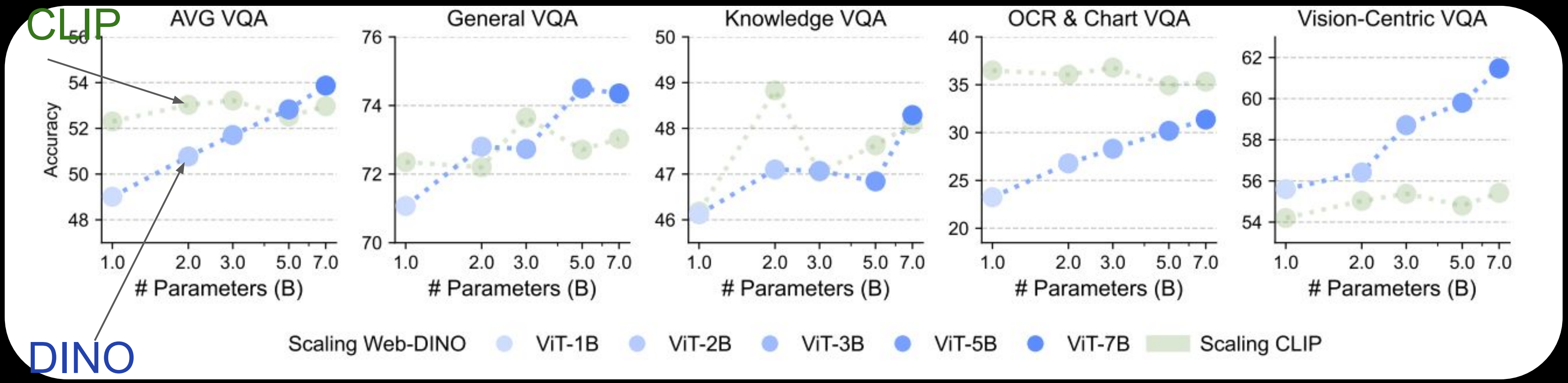
Scaling Up Model



Scaling Up Model

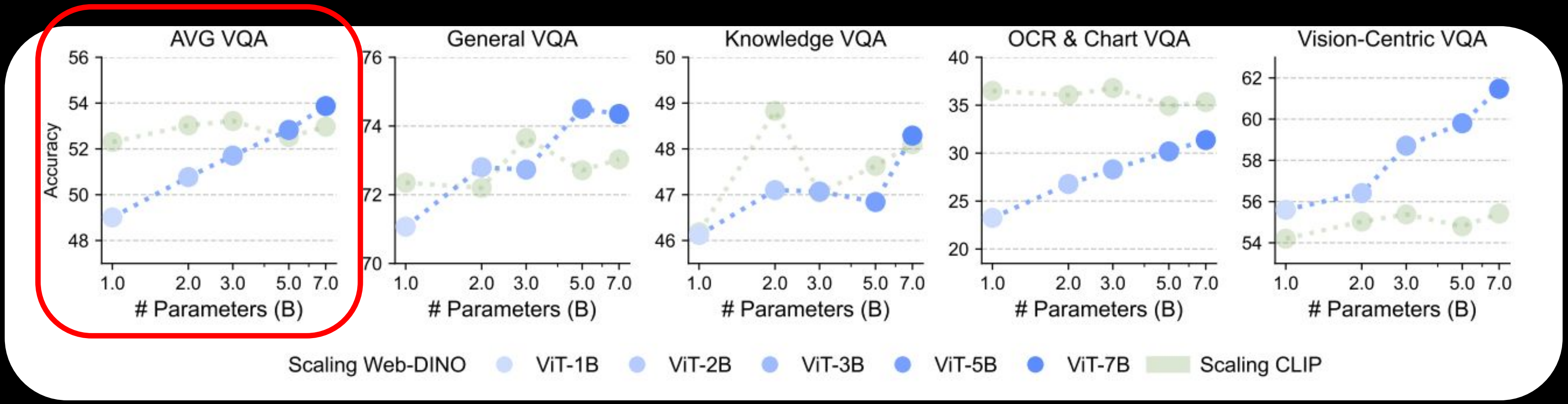


Scaling Up Model



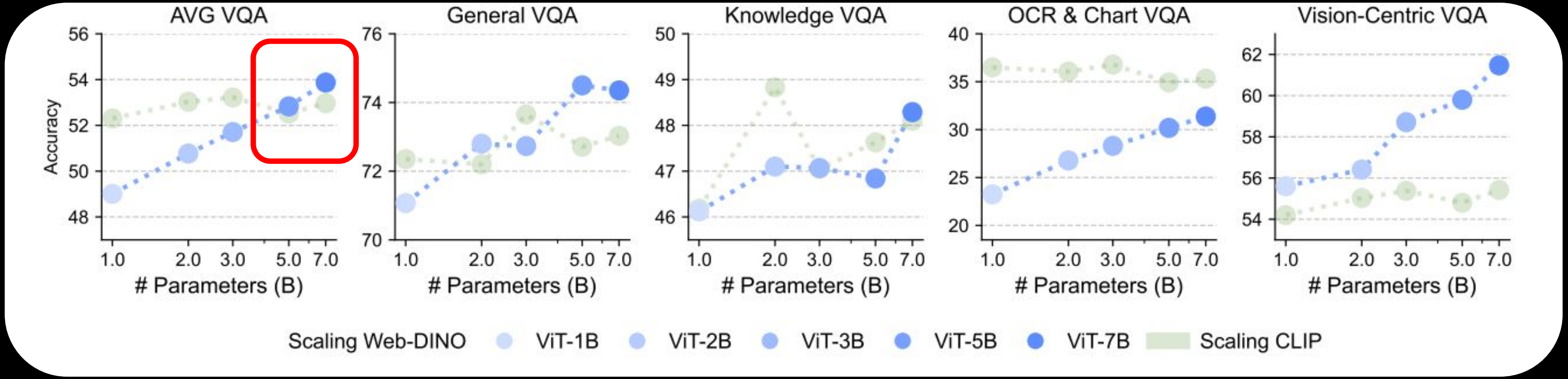
Scaling Up Model

1. Web-DINO scales log-linearly *w.r.t* to model sizes



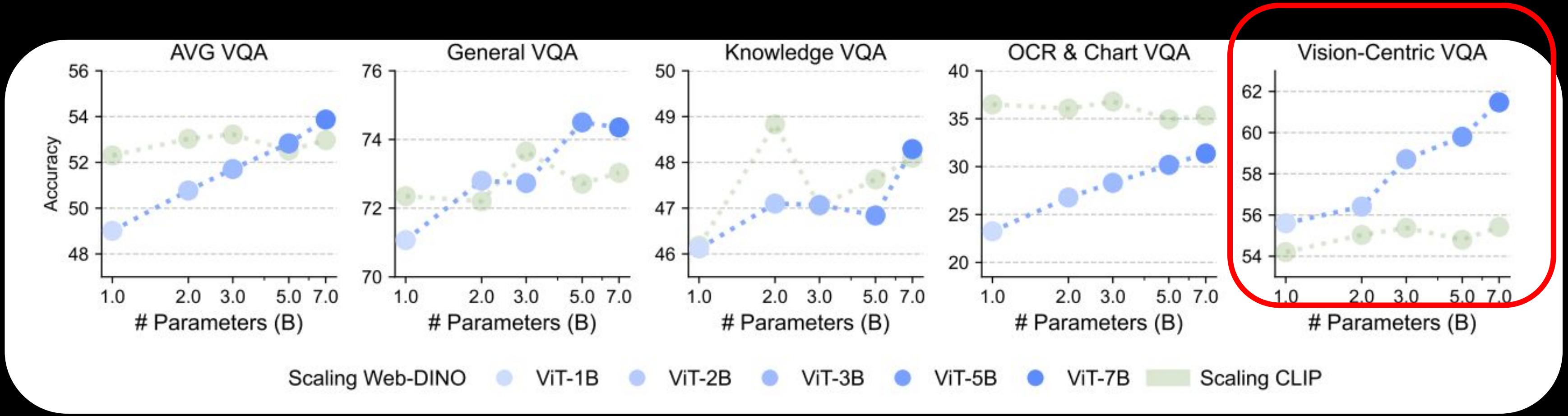
Scaling Up Model

1. Web-DINO scales log-linearly *w.r.t* to model sizes
2. Under same conditions, Web-DINO scales better than CLIP



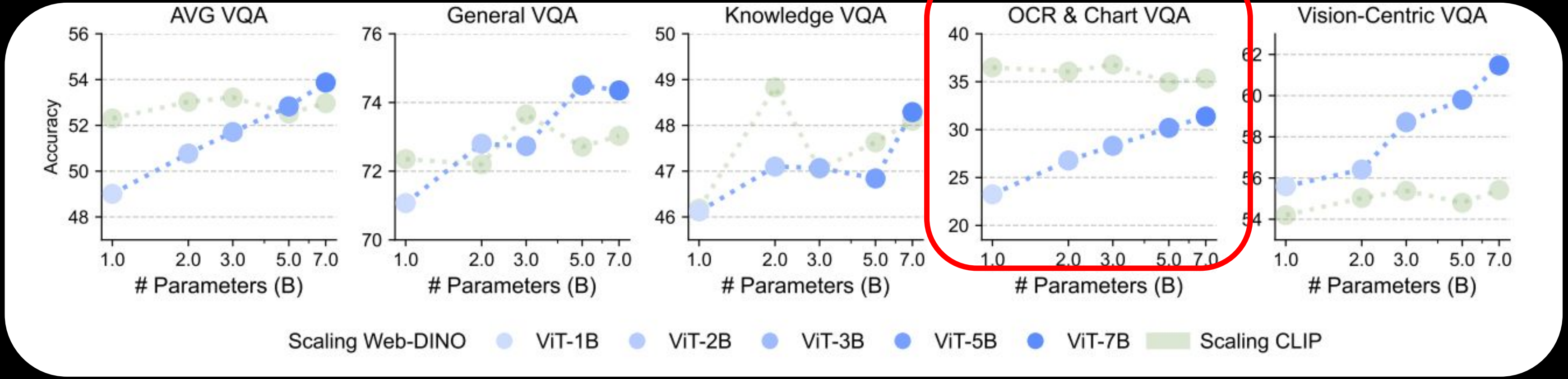
Scaling Up Model

1. Web-DINO scales log-linearly *w.r.t* to model sizes
2. Under same conditions, Web-DINO scales better than CLIP
3. Web-DINO continues to excel on Vision-Centric VQA



Scaling Up Model

1. Web-DINO scales log-linearly *w.r.t* to model sizes
2. Under same conditions, Web-DINO scales better than CLIP
3. Web-DINO continues to excel on Vision-Centric VQA
4. The gap on OCR & Chart is closing!



Scaling Up Data

Scaling Up Data

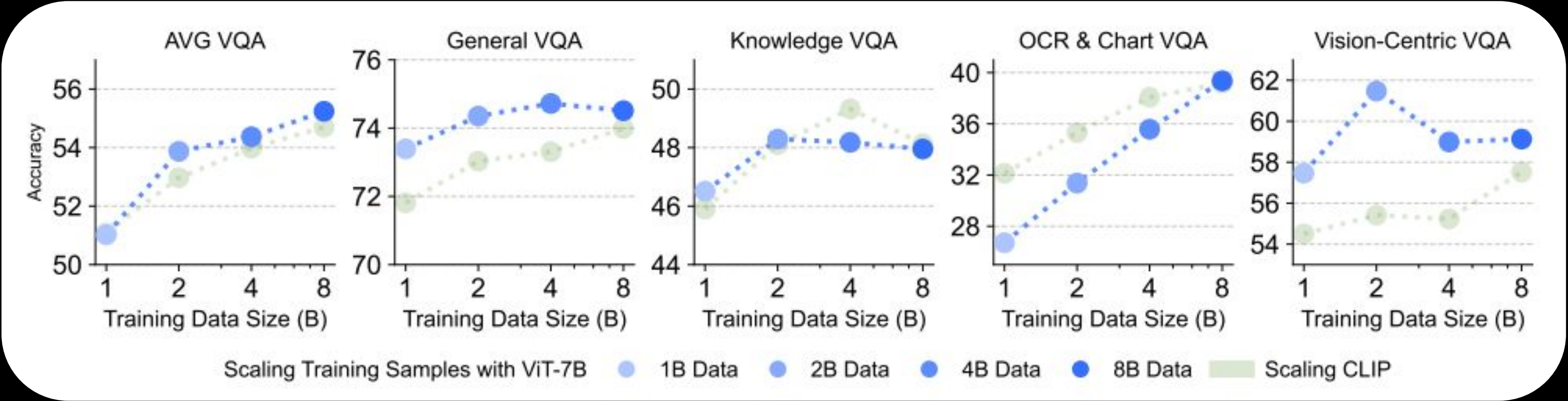
Data: *MC-2B*:
- 1 billion samples seen
- 2 billion samples seen
- 4 billion samples seen
- 8 billion samples seen

Model: ViT-7B

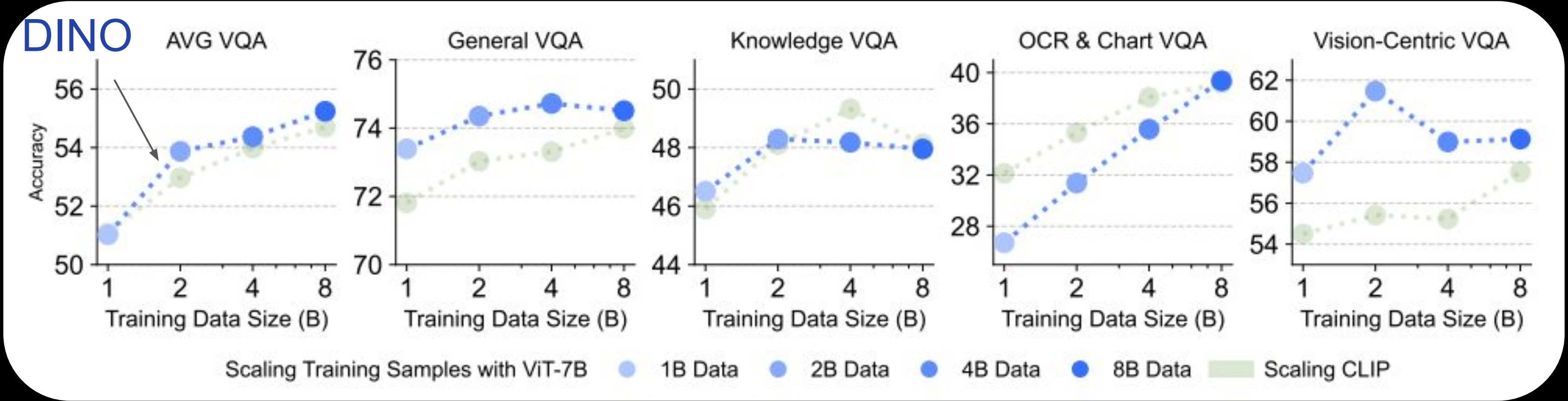
Method: DINOv2

Eval: Use VQA as evaluation.

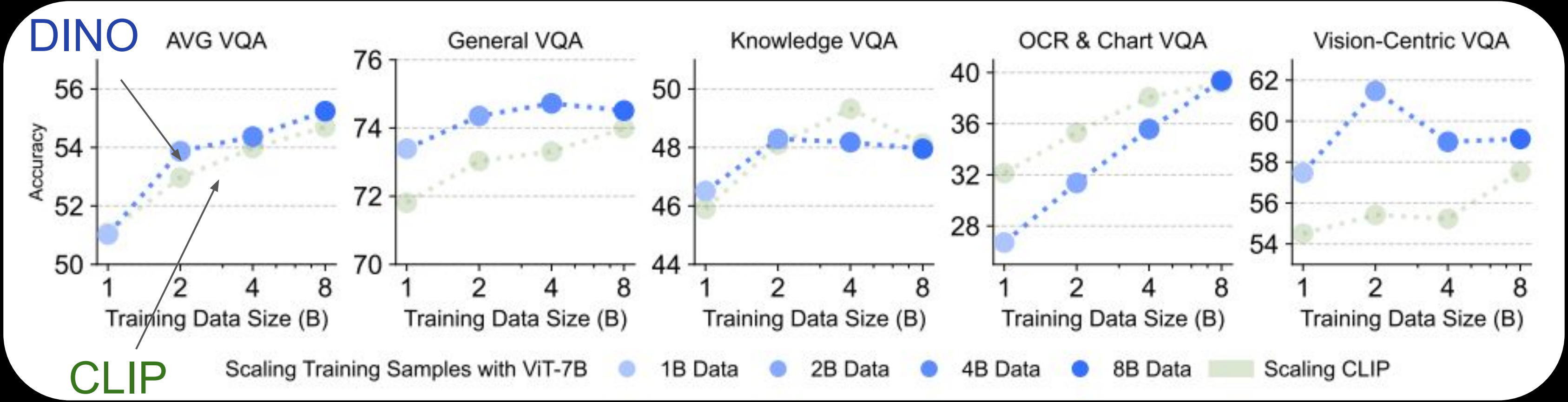
Scaling Up Data



Scaling Up Data

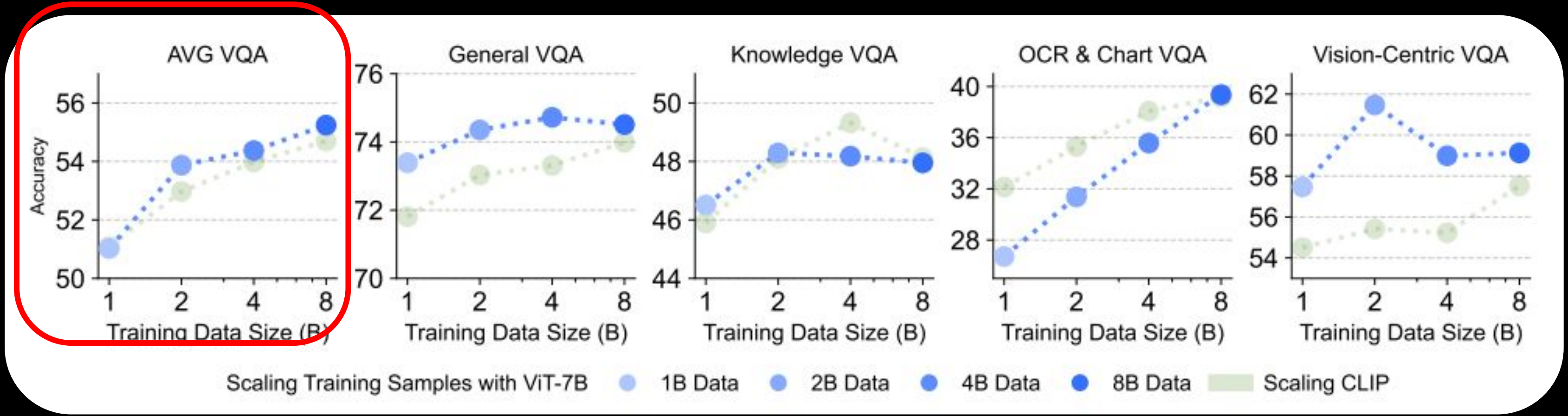


Scaling Up Data



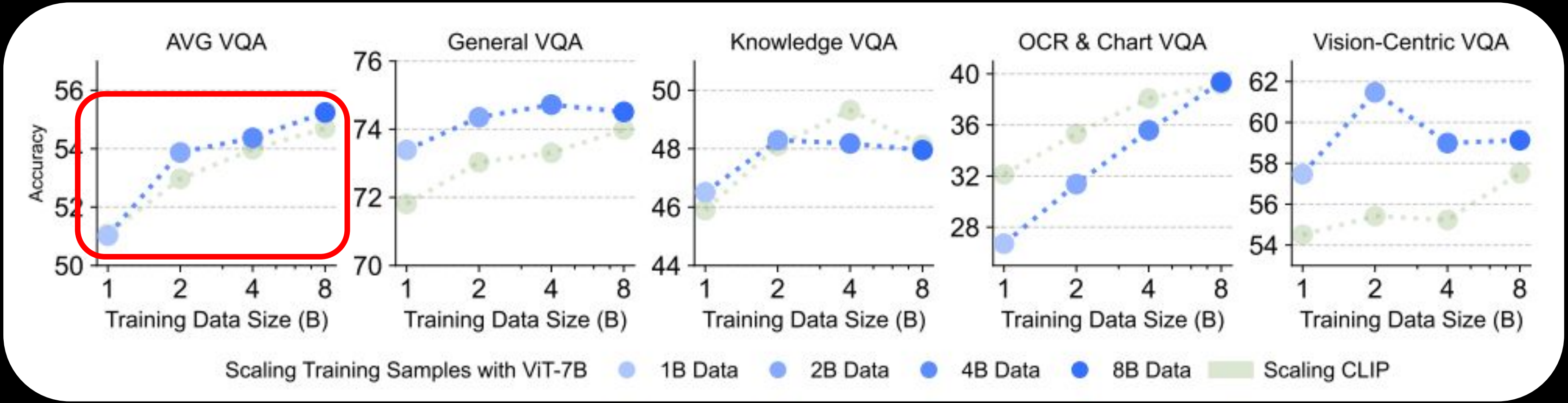
Scaling Up Data

1. Model improves *w.r.t* to more data seen



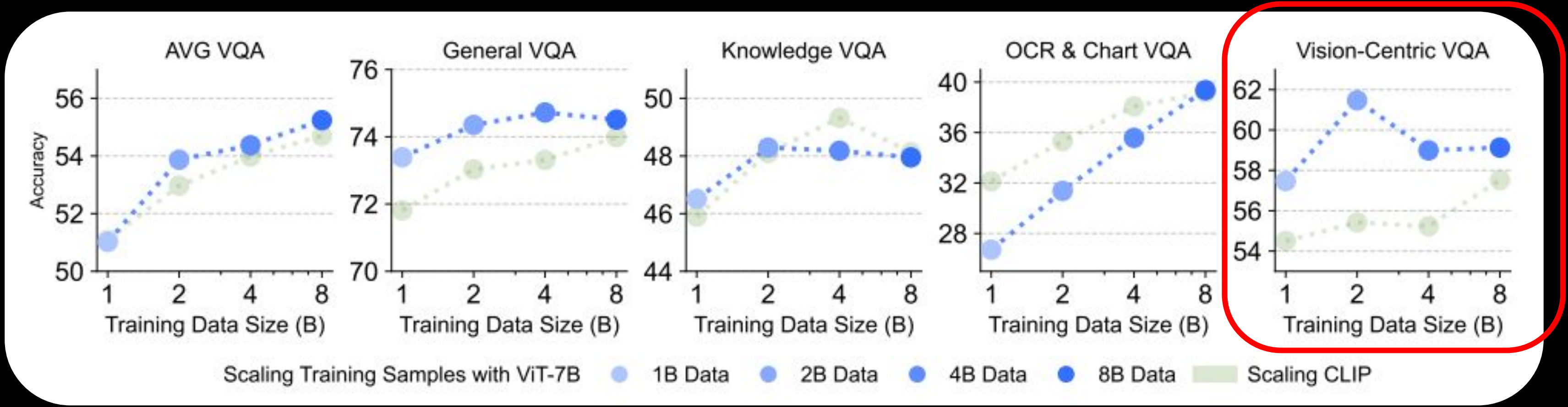
Scaling Up Data

1. Model improves *w.r.t* to more data seen
2. SSL models consistently outperforms CLIP models



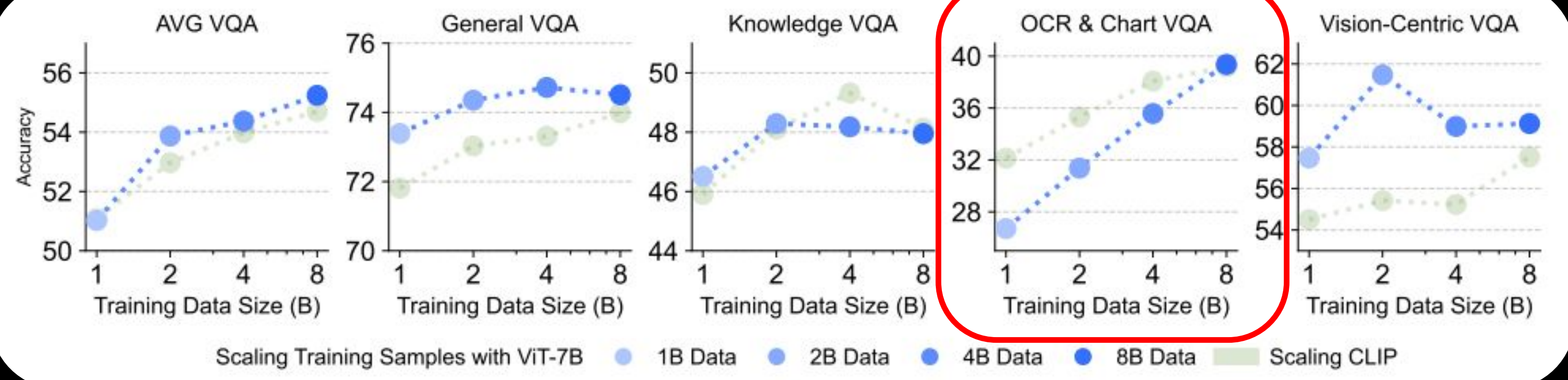
Scaling Up Data

1. Model improves *w.r.t* to more data seen
2. SSL models consistently outperforms CLIP models
3. SSL models are better "visual" model



Scaling Up Data

1. Model improves *w.r.t* to more data seen
2. SSL models consistently outperforms CLIP models
3. SSL models are better “visual” model
4. Gap closes on OCR & Chart.

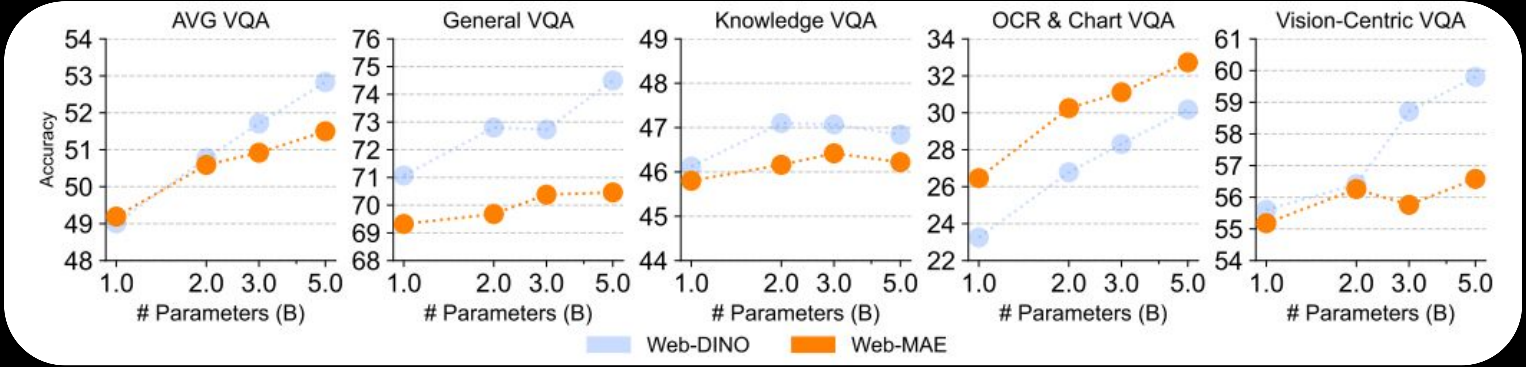


Q1. Does the observed scaling behavior generalize to other visual SSL methods?

Q1. Does the observed scaling behavior generalize to other visual SSL methods?

We conduct similar experiments on MAE

And YES!

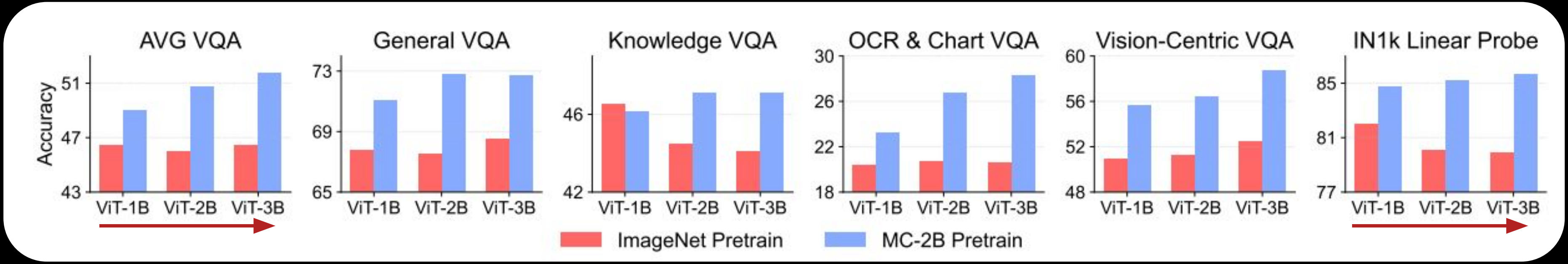


Q2.Does visual SSL exhibit similar scaling behavior on smaller scale conventional data such as ImageNet?

Q2.Does visual SSL exhibit similar scaling behavior on smaller scale conventional data such as ImageNet?

We conduct similar experiments training on ImageNet-1k

No obvious scaling trend on both VQA and ImageNet-1k



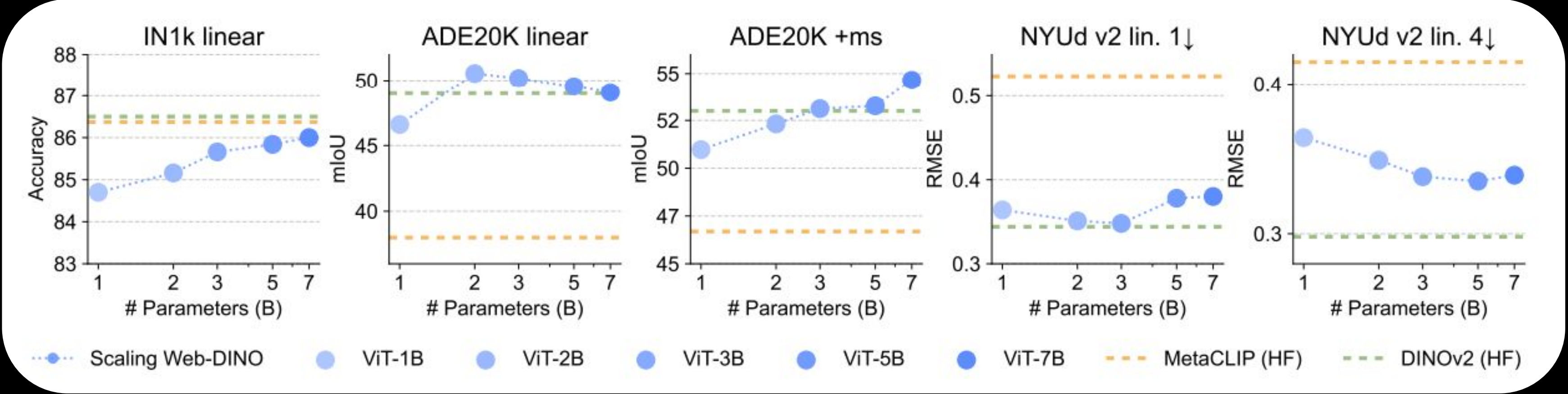
Q3.How do scaled models perform on classic vision tasks?

Q3.How do scaled models perform on classic vision tasks?

Evaluate our trained Web-DINO on classic vision benchmarks

Q3.How do scaled models perform on classic vision tasks?

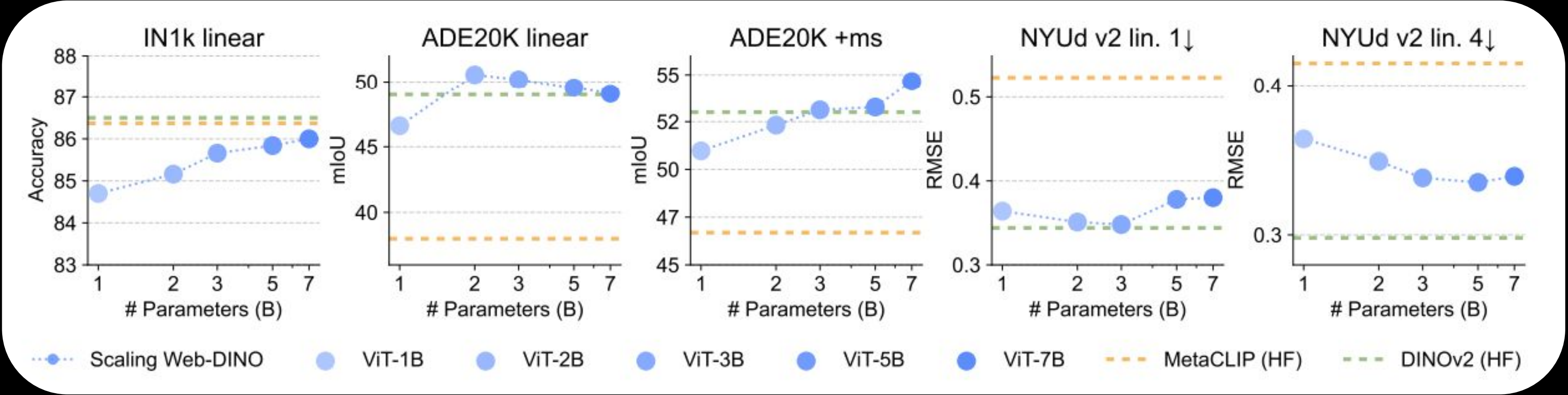
Evaluate our trained Web-DINO on classic vision benchmarks



Q3.How do scaled models perform on classic vision tasks?

Evaluate our trained Web-DINO on classic vision benchmarks

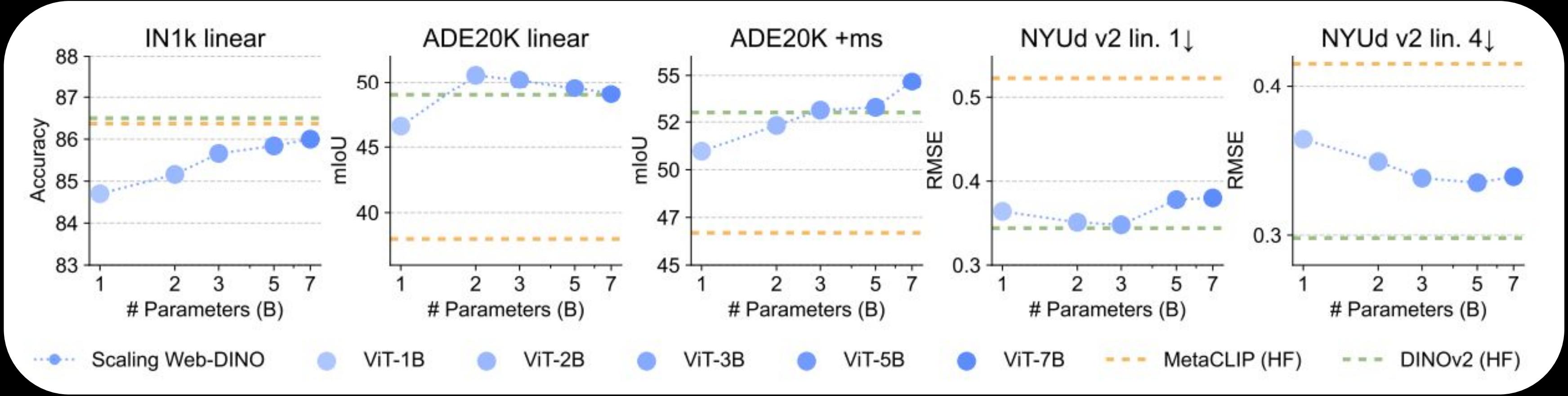
1. Web-DINO is mostly better than MetaCLIP



Q3.How do scaled models perform on classic vision tasks?

Evaluate our trained Web-DINO on classic vision benchmarks

1. Web-DINO is mostly better than MetaCLIP
2. Web-DINO remains competitive with DINOv2
   a. Challenging! Since LVD142M is retrieved from classic vision tasks train set.



Q4.Why does web-scale data improve OCR & Chart performance?

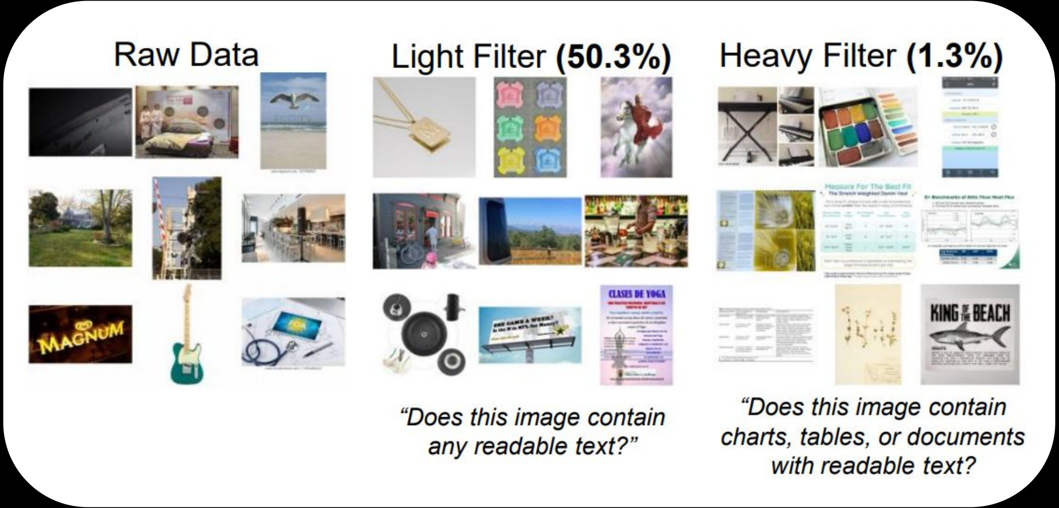
Q4.Why does web-scale data improve OCR & Chart performance?

Hypothesis: Maybe web-scale data contains very rich text information in images, and SSL models can learn from them

Q4.Why does web-scale data improve OCR & Chart performance?

Hypothesis: Maybe web-scale data contains very rich text information in images, and SSL models can learn from them

Filter images that contain text/chart/documents…



Q4. Why does web-scale data improve OCR & Chart performance?

Hypothesis: Maybe web-scale data contains very rich text information in images, and SSL models can learn from them

Method	% of MC-2B	AVG	General	Knowledge	Vision Centric	OCR Chart	ChartQA	OCRBench	TextVQA	DocVQA
		VQA Evaluator					Breakdown of OCR & Chart Tasks			
CLIP 2B	100%	53.0	72.2	48.8	55.0	36.1	32.8	32.9	52.6	26.0
Web-DINO 2B	100%	50.8	72.8	47.1	56.4	26.8	23.3	15.6	49.2	19.0
Web-DINO 2B	50.3%	53.4 (+2.6)	73.0 (+0.2)	51.7 (+4.6)	55.6 (-0.8)	33.2 (+6.4)	31.4 (+8.1)	27.3 (+11.7)	51.3 (+2.1)	23.0 (+4.0)
Web-DINO 2B	1.3%	53.7 (+2.9)	70.7 (-2.1)	47.3 (+0.2)	56.2 (-0.2)	40.4 (+13.6)	47.5 (+24.2)	29.4 (+13.8)	52.8 (+3.6)	32.0 (+13.0)

Q4.Why does web-scale data improve OCR & Chart performance?

Hypothesis: Maybe web-scale data contains very rich text information in images, and SSL models can learn from them

The "text" in images contribute to the OCR & Chart ability and SSL method can learn from it

Method	% of MC-2B	VQA Evaluator AVG	General	Knowledge	Vision Centric	OCR Chart	Breakdown of OCR & Chart Tasks ChartQA	OCRBench	TextVQA	DocVQA
CLIP 2B	100%	53.0	72.2	48.8	55.0	36.1	32.8	32.9	52.6	26.0
Web-DINO 2B	100%	50.8	72.8	47.1	56.4	26.8	23.3	15.6	49.2	19.0
Web-DINO 2B	50.3%	53.4 (+2.6)	73.0 (+0.2)	51.7 (+4.6)	55.6 (-0.8)	33.2 (+6.4)	31.4 (+8.1)	27.3 (+11.7)	51.3 (+2.1)	23.0 (+4.0)
Web-DINO 2B	1.3%	53.7 (+2.9)	70.7 (-2.1)	47.3 (+0.2)	56.2 (-0.2)	40.4 (+13.6)	47.5 (+24.2)	29.4 (+13.8)	52.8 (+3.6)	32.0 (+13.0)

Q5. Why can SSL learn strong visual representations for multimodal modeling, without language supervision?

Q5. Why can SSL learn strong visual representations for multimodal modeling, without language supervision?

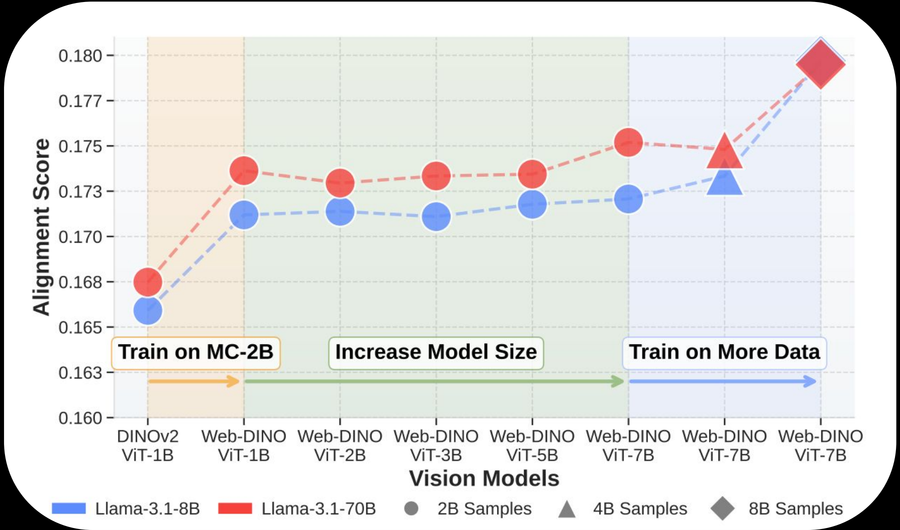
Hypothesis: SSL models learn features increasingly aligned with language as model size and examples seen increases.

Q5. Why can SSL learn strong visual representations for multimodal modeling, without language supervision?

Hypothesis: SSL models learn features increasingly aligned with language as model size and examples seen increases.

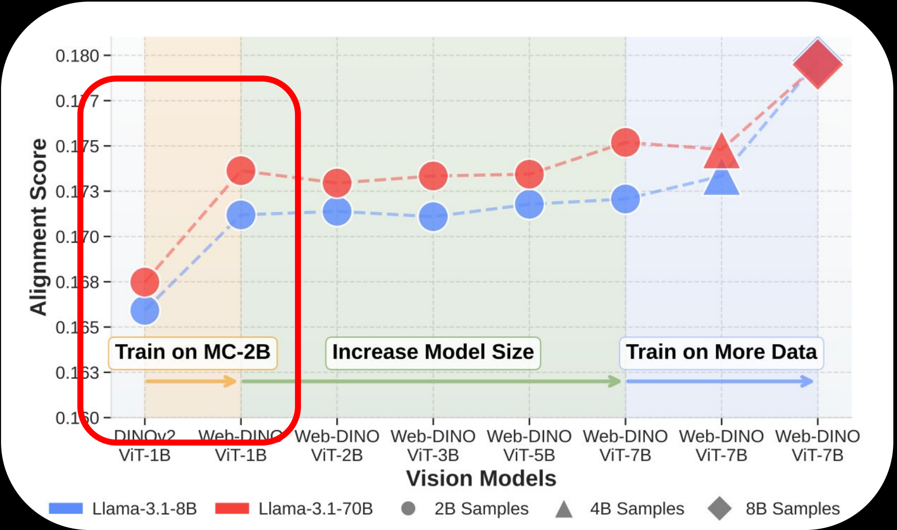
Measure its alignment with LLM via “Platonic Hypothesis”

Q5. Why can SSL learn strong visual representations for multimodal modeling, without language supervision?



Q5. Why can SSL learn strong visual representations for multimodal modeling, without language supervision?

1. Training on more diverse data (MC-2B) lead to better alignment



Q5. Why can SSL learn strong visual representations for multimodal modeling, without language supervision?

1. Training on more diverse data (MC-2B) lead to better alignment
2. Increase model size gradually lead to better alignment

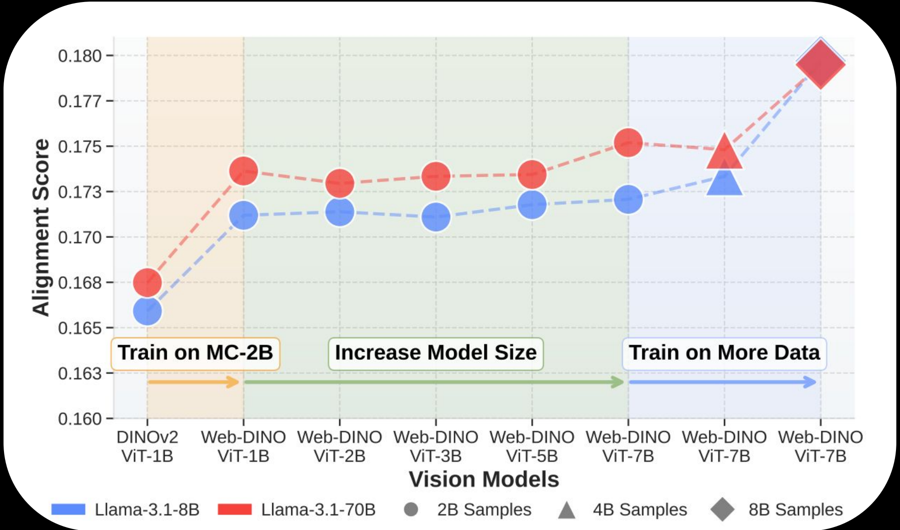
Q5. Why can SSL learn strong visual representations for multimodal modeling, without language supervision?

1. Training on more diverse data (MC-2B) lead to better alignment
2. Increase model size gradually lead to better alignment
3. Training on more data lead to better alignment

Q5. Why can SSL learn strong visual representations for multimodal modeling, without language supervision?

1. Training on more diverse data (MC-2B) lead to better alignment
2. Increase model size gradually lead to better alignment
3. Training on more data lead to better alignment

As the model scales larger or train longer, it naturally aligns more with LLM



Takeaways

- CLIP models might be the bottleneck in understanding the “visual” world and scaling up does NOT resolve the problem.

Takeaways

- CLIP models might be the bottleneck in understanding the “visual” world and scaling up does NOT resolve the problem.

- We need to develop better visual representation

Takeaways

- CLIP models might be the bottleneck in understanding the “visual” world and scaling up does NOT resolve the problem.
- We need to develop better visual representation
- Visual SSL are scalable learner: improves *w.r.t* to model and data sizes when we use VQA as evaluation

Takeaways

- CLIP models might be the bottleneck in understanding the “visual” world and scaling up does NOT resolve the problem.
- We need to develop better visual representation
- Visual SSL are scalable learner: improves *w.r.t* to model and data sizes when we use VQA as evaluation
- Visual SSL is compatible with CLIP models on VQA, even on OCR & Chart. And Visual SSL models are very good at “Vision”

Thank you!