Last revised 10/19/25

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

ICCV HONOLULU HAWAII OCT 19-23, 2025

ICCV 2025 Highlight

Meta NYU

Current State of Visual Representation Learning

Current State of Visual Representation Learning

Self-Supervision:

Language-Supervision:

Current State of Visual Representation Learning

Self-Supervision:

- E.g. MoCo, MAE, DINO

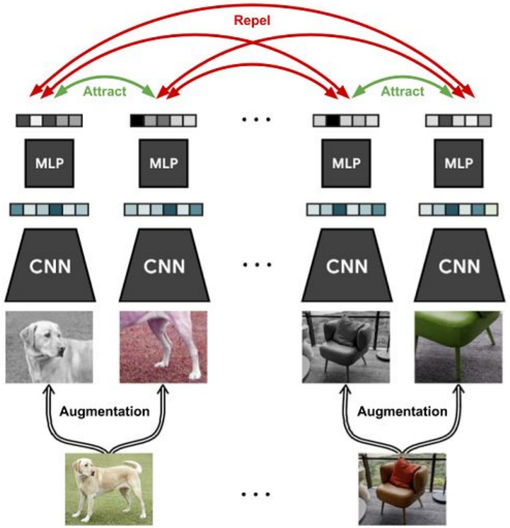
Language-Supervision:

- E.g. CLIP, SigLIP, MetaCLIP

Current State of Visual Representation Learning

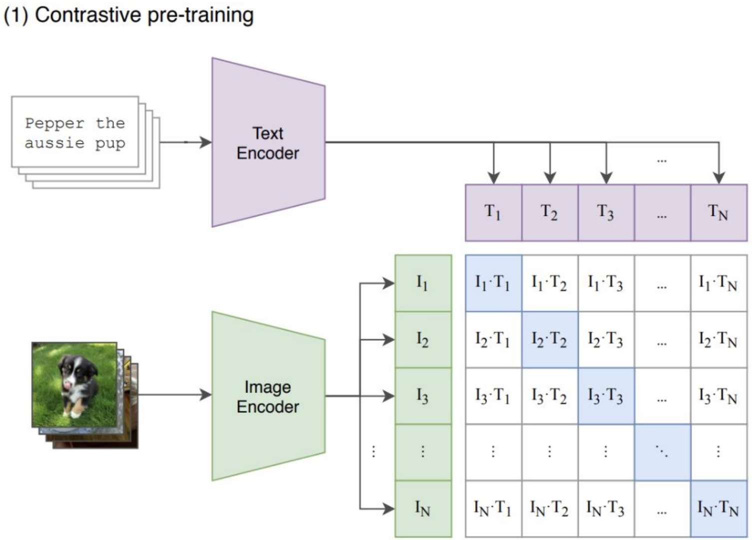
Self-Supervision:

- E.g. MoCo, MAE, DINO
- Learning from images directly (e.g. augmentation, masking)



Language-Supervision:

- E.g. CLIP, SigLIP, MetaCLIP
- Learning from language captions that describe the image



Current State of Visual Representation Learning

Self-Supervision:

- E.g. MoCo, MAE, DINO
- Learning from images directly (e.g. augmentation, masking)
- Training on ImageNet-like data (1M to >100M scale)

Language-Supervision:

- E.g. CLIP, SigLIP, MetaCLIP
- Learning from language captions that describe the image
- Training on image-text pairs from the Internet (400M to 100B scale)

Current State of Visual Representation Learning

Self-Supervision:

- E.g. MoCo, MAE, DINO
- Learning from images directly (e.g. augmentation, masking)
- Training on ImageNet-like data (1M to >100M scale)
- Good at classification, segmentation, depth estimation, etc

Language-Supervision:

- E.g. CLIP, SigLIP, MetaCLIP
- Learning from language captions that describe the image
- Training on image-text pairs from the Internet (400M to 100B scale)
- Good at classification, and widely used as backbone for **multimodal** models

The Success of CLIP as an Encoder in Multimodal Models

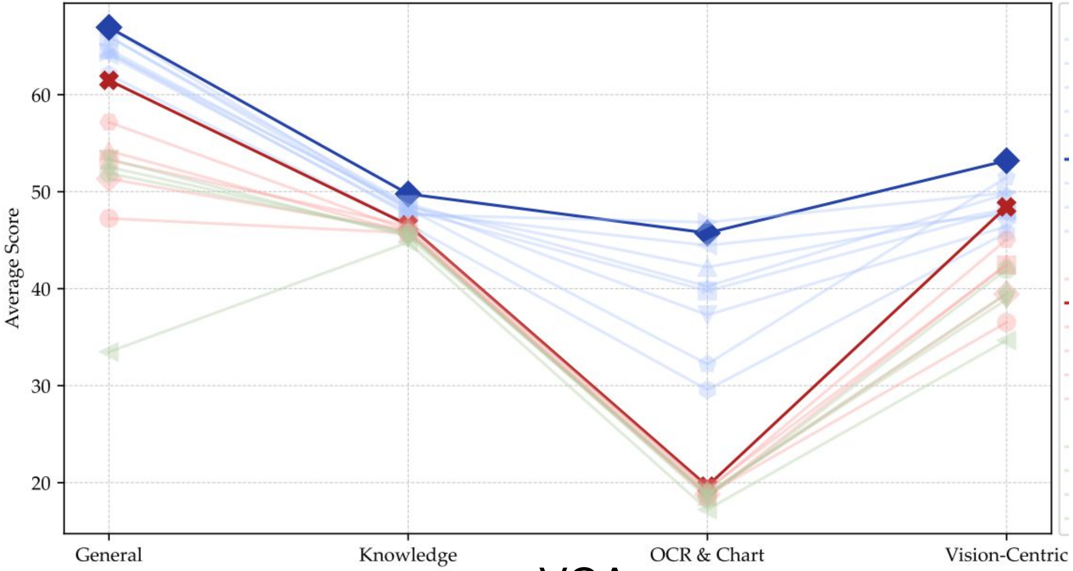
- CLIP has become the dominant visual representation learning method in multimodal models.

The Success of CLIP as an Encoder in Multimodal Models

- CLIP has become the dominant visual representation learning method in multimodal models.
  - VLM: LLaVA, Cambrian, PaliGemma, SEED-VL …
  - VLA: Pi, Otter, …
  - …

The Success of CLIP as an Encoder in Multimodal Models

- CLIP has become the dominant visual representation learning method in multimodal models.

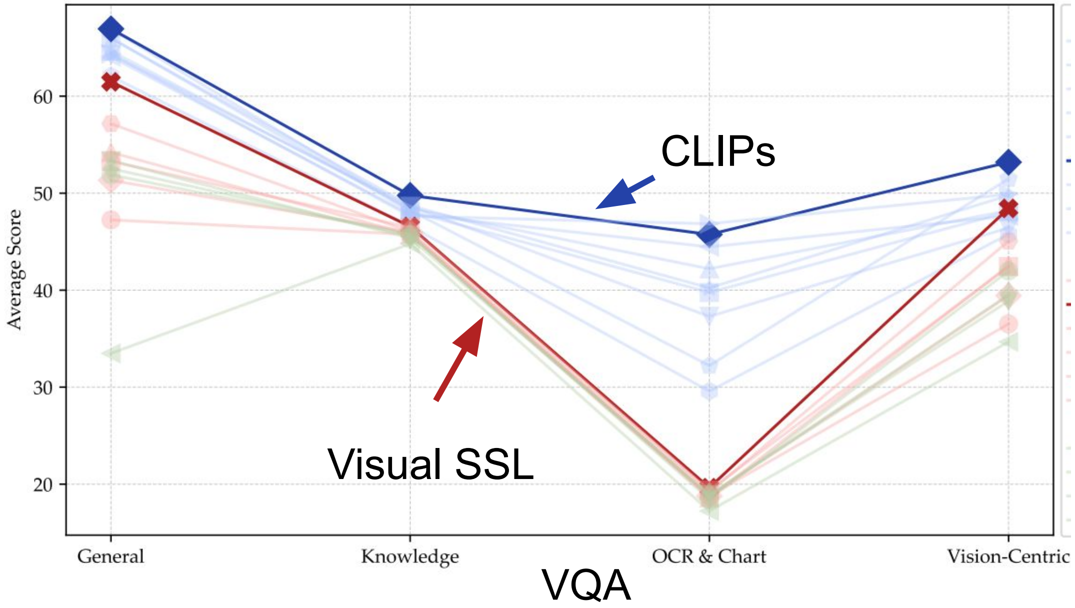


VQA

Tong, S., et al. (2024). Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs.

The Success of CLIP as an Encoder in Multimodal Models

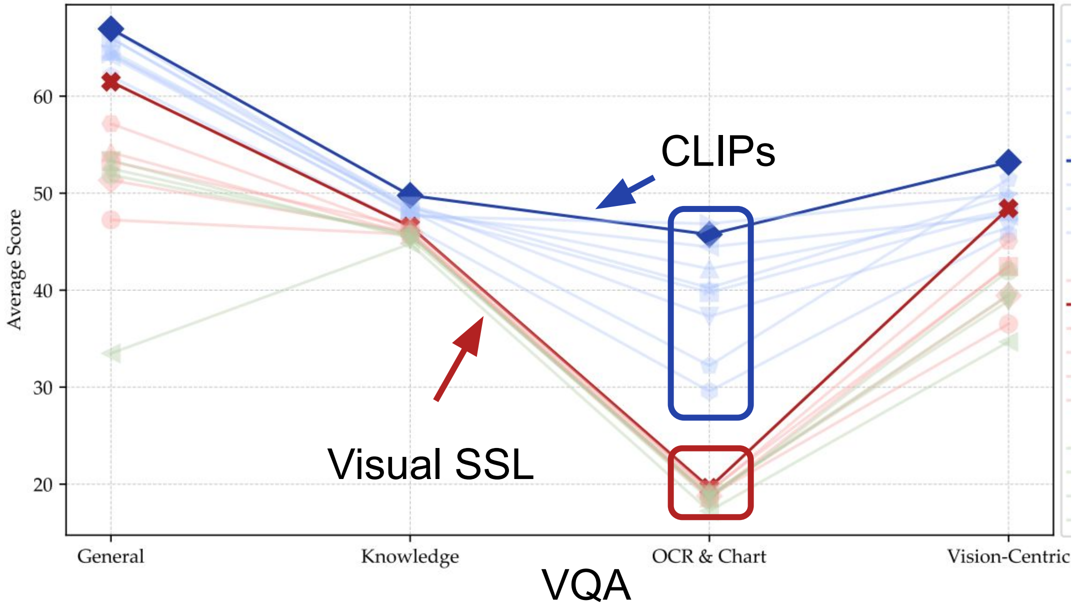
- CLIP has become the dominant visual representation learning method in multimodal models.



Tong, S., et al. (2024). Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs.

The Success of CLIP as an Encoder in Multimodal Models

- CLIP has become the dominant visual representation learning method in multimodal models.



Tong, S., et al. (2024). Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs.

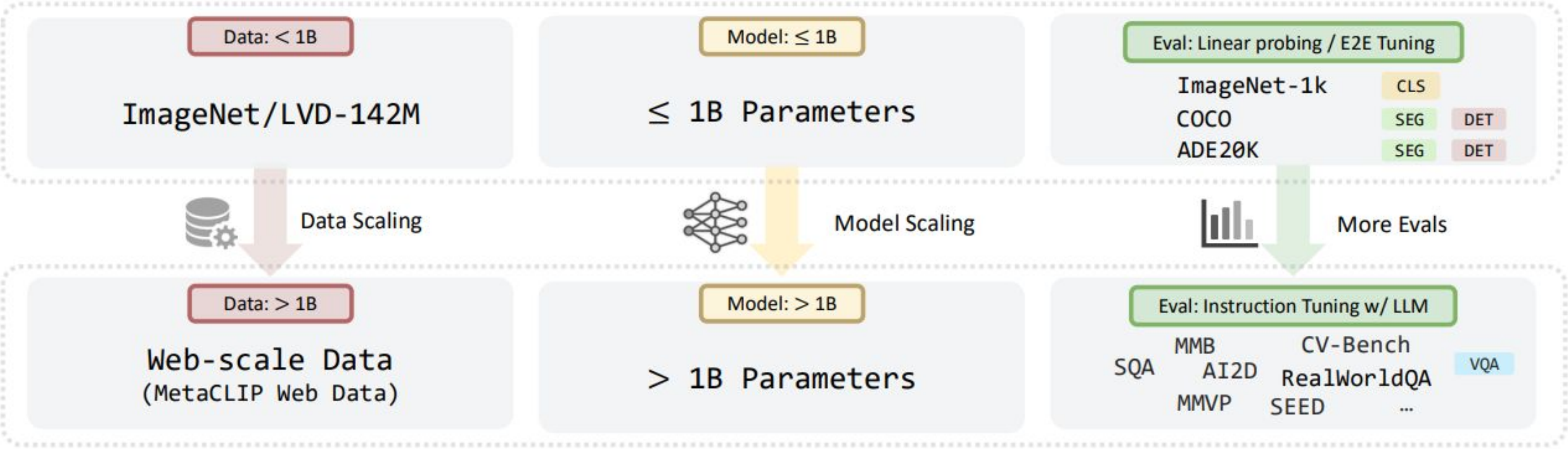
The Success of CLIP as an Encoder in Multimodal Models

- CLIP has become the dominant visual representation learning method in multimodal models.
- Is CLIP better because of **language supervision** or **data distribution**?

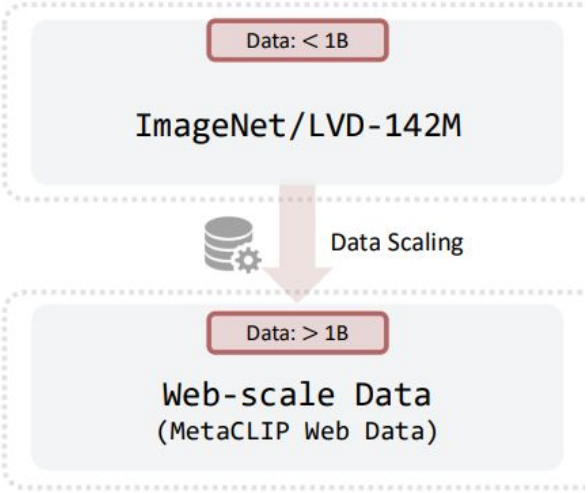
The Success of CLIP as an Encoder in Multimodal Models

- CLIP has become the dominant visual representation learning method in multimodal models.
- Is CLIP better because of **language supervision** or **data distribution**?
- To really understand this, we need controlled comparisons on the data.

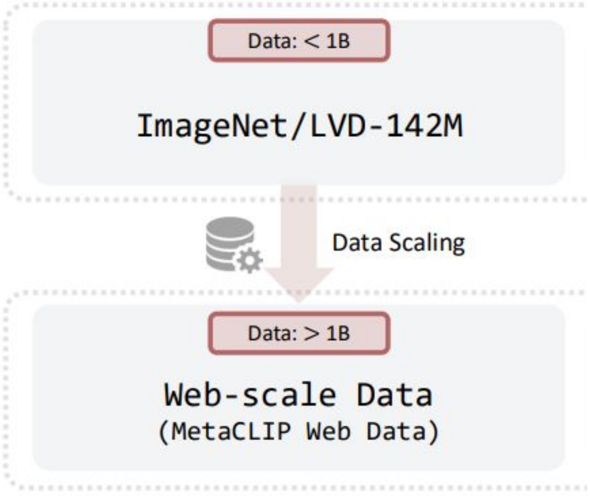
WebSSL: Towards Modernizing Visual SSL



WebSSL: Towards Modernizing Visual SSL



WebSSL: Towards Modernizing Visual SSL



ImageNet / LVD-142M[1]:

Million scale ImageNet
or
ImageNet-like distribution
of mostly natural images

→

Web-Scale Images:

Billion scale diverse "random" images from the Internet

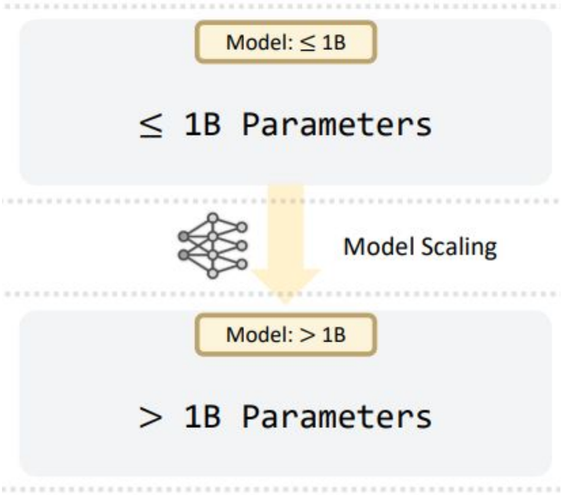
E.g. MetaCLIP[2] ("*MC-2B*")

We only use the images for SSL

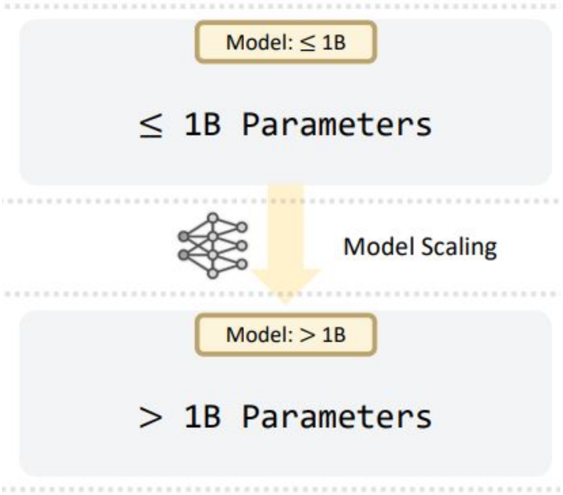
[1] Oquab, M., et al. (2023). DINOv2: Learning Robust Visual Features without Supervision

[2] Xu, H., et al. (2023). Demystifying CLIP Data.

WebSSL: Towards Modernizing Visual SSL



WebSSL: Towards Modernizing Visual SSL



Less than 1B params:

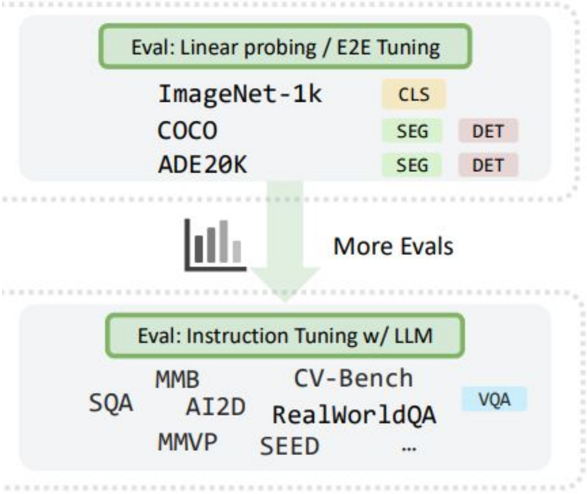
ViT-B, ViT-L, ViT-H, ViT-g

→

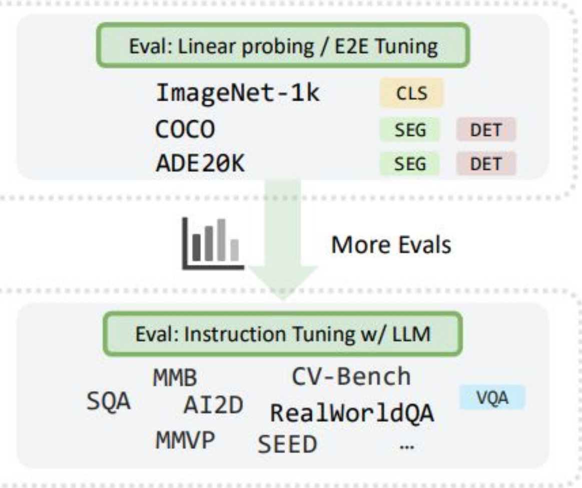
More than 1B params:

ViT-1B, …, VIT-7B and beyond

WebSSL: Towards Modernizing Visual SSL

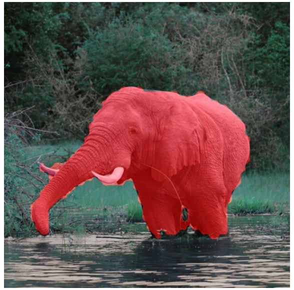


WebSSL: Towards Modernizing Visual SSL



Classic Vision Eval:

Classification, segmentation, depth estimation, etc.



Elephant

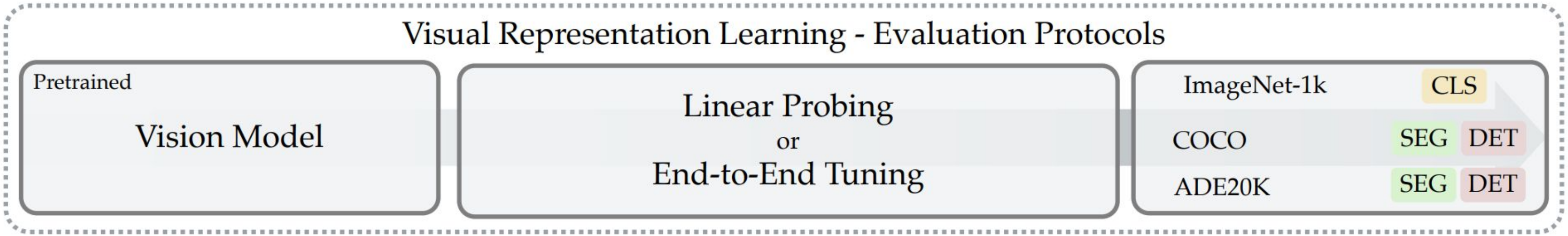
→

VQA as a Vision Eval:

Assesses wider range of capabilities and more diverse questions

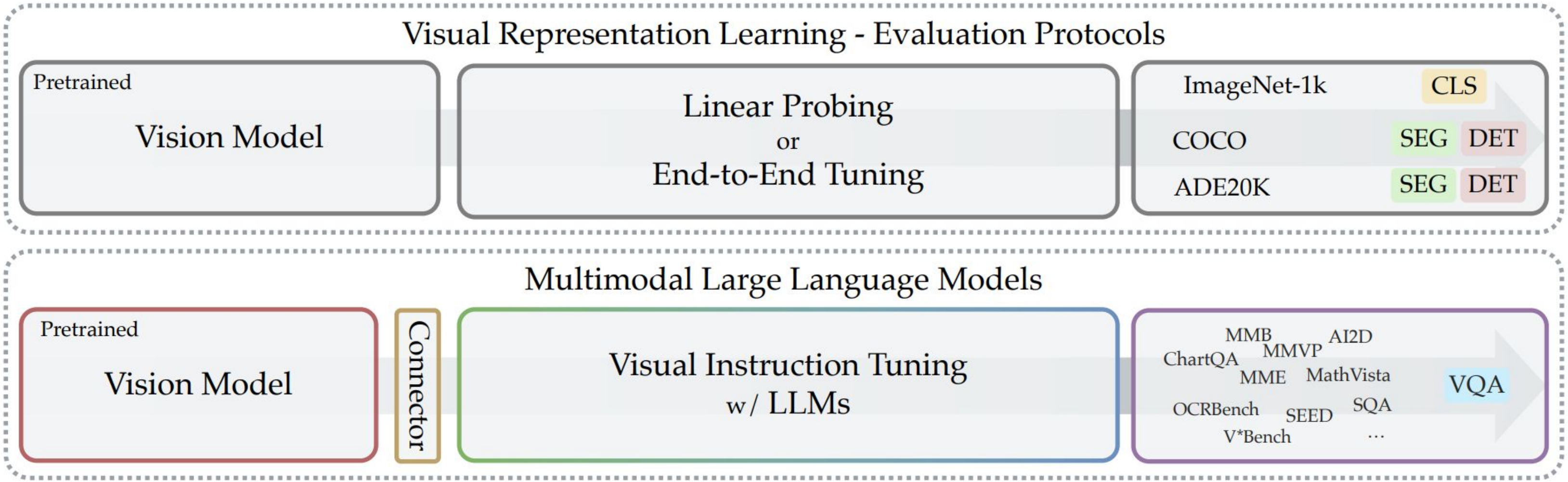
How many cars are in the image?

Evaluation Setup



Tong, S., et al. (2024). Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs.

Evaluation Setup

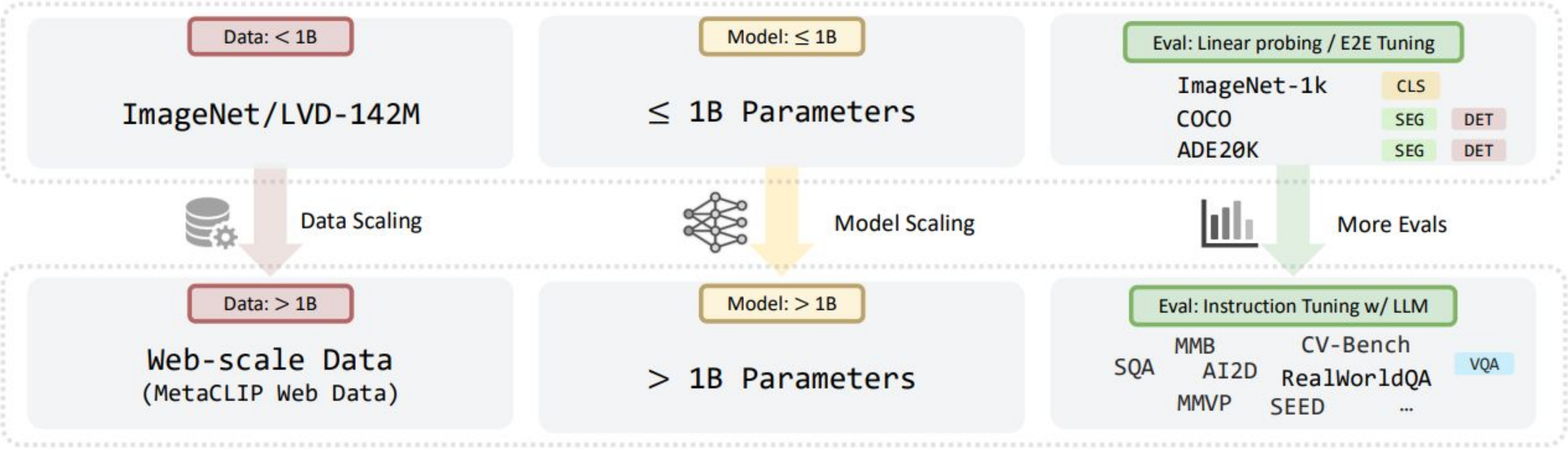


We use Cambrian with a *frozen* vision encoder (but finetuned adapter + LLM) to evaluate on VQA tasks: **General, Knowledge, OCR&Chart, Vision-Centric**

Tong, S., et al. (2024). Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs.

"Is language supervision or the data more important?"

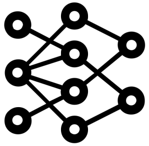
“Is language supervision or the data more important?”



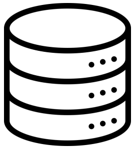
Let’s train WebSSL and find out via controlled experiments!

WebSSL

1. Scaling up model



2. Scaling up data



WebSSL: Scaling Up Model

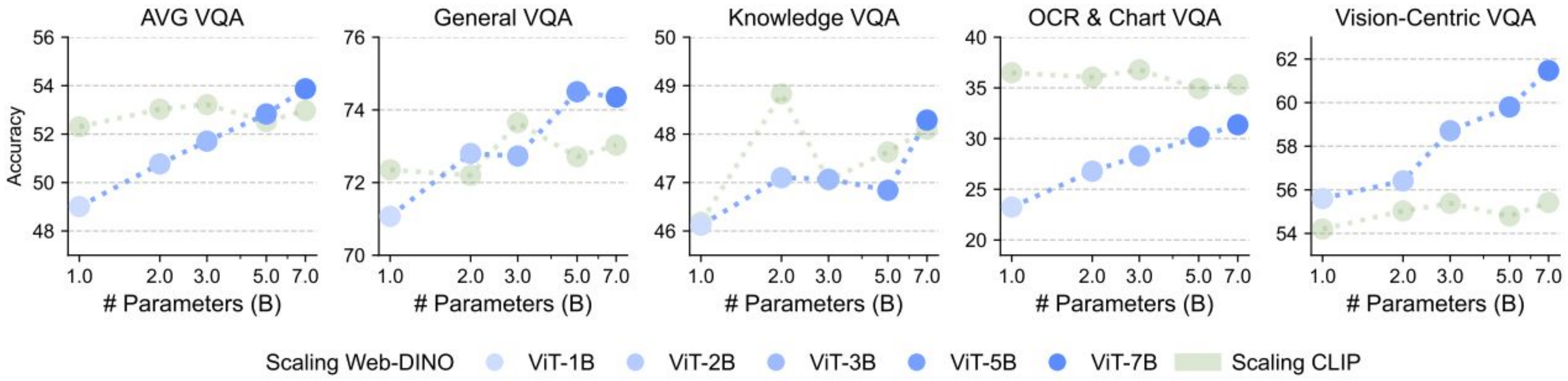
WebSSL: Scaling Up Model

- **Data**: *MC-2B*, 2 billion samples seen
- **Model**: ViT-1B, ViT-2B, ViT-3B, ViT-5B, ViT-7B
- **Method**: DINOv2 (SSL) vs. CLIP (Language-Supervised)
- **Eval**: Use VQA as evaluation and categorize Cambrian eval benchmarks:

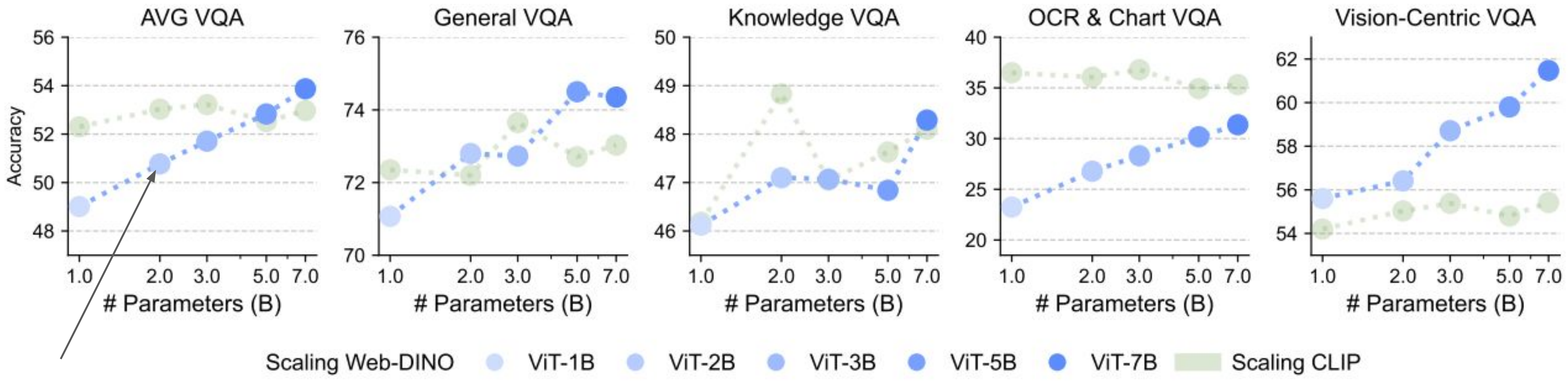
General	Knowledge	OCR & Chart	Vision-Centric
MMBench-En	AI2D	ChartQA	CV-Bench 2D
MME	MathVista	DocVQA	CV-Bench 3D
GQA	MMMU	OCRBench	MMVP
SEED	ScienceQA	TextVQA	RealWorldQA

Tong, S., et al. (2024). Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs.

WebSSL: Scaling Up Model

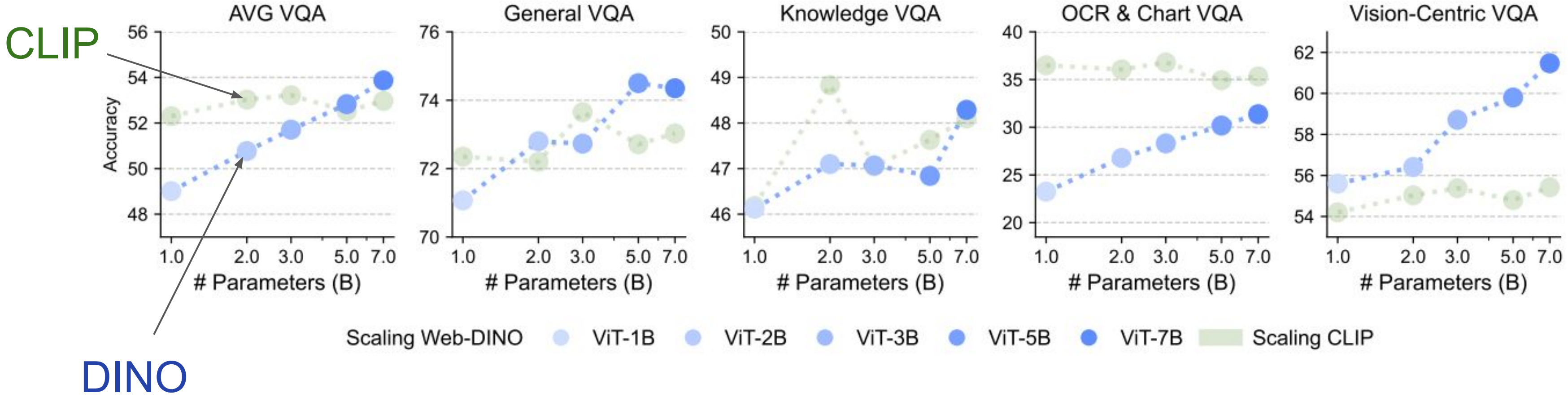


WebSSL: Scaling Up Model



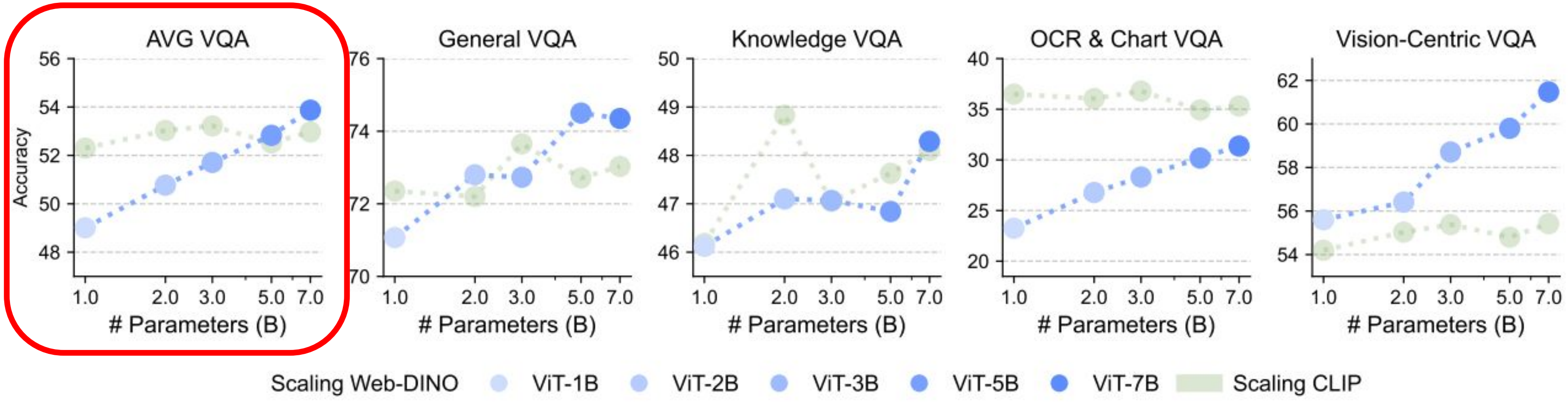
DINO

WebSSL: Scaling Up Model



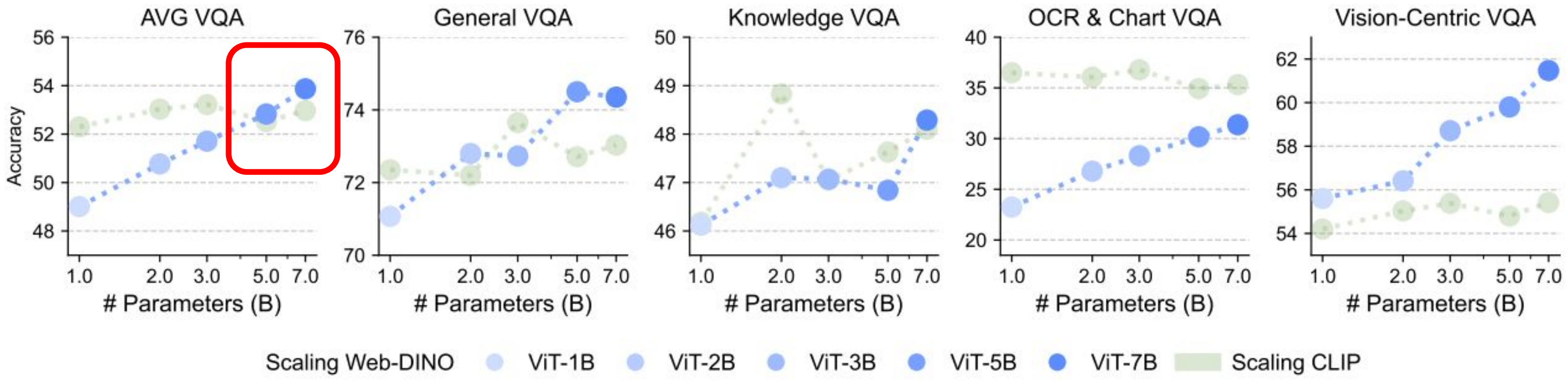
WebSSL: Scaling Up Model

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



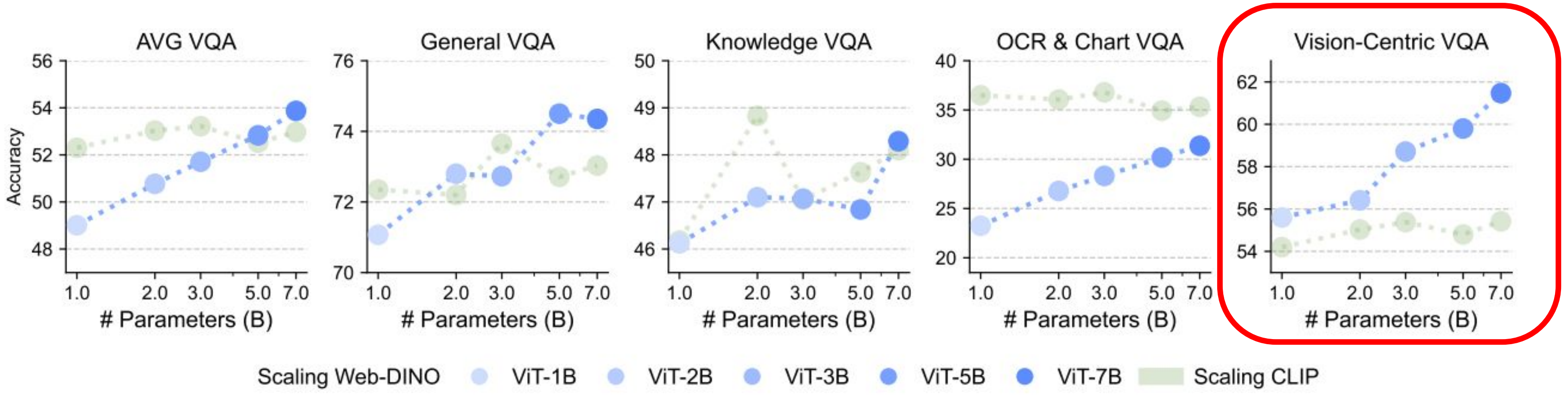
WebSSL: Scaling Up Model

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



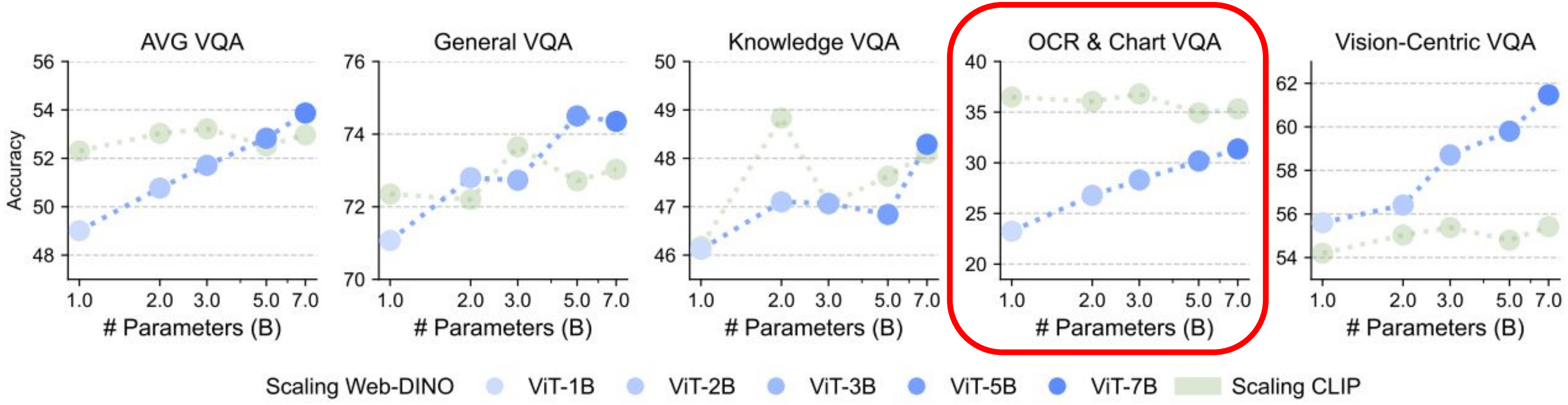
WebSSL: Scaling Up Model

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



WebSSL: 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!

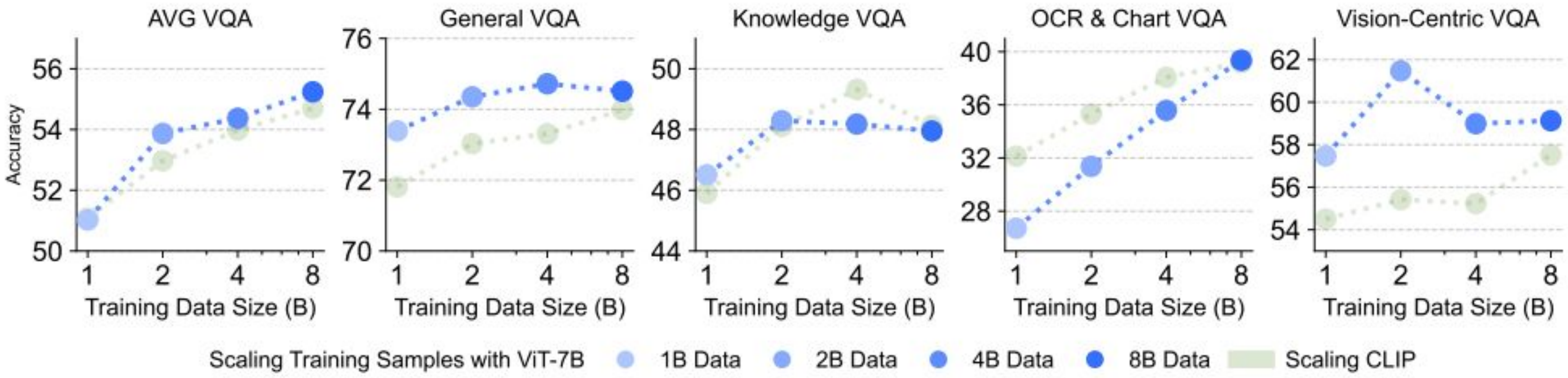


WebSSL: Scaling Up Data

WebSSL: Scaling Up Data

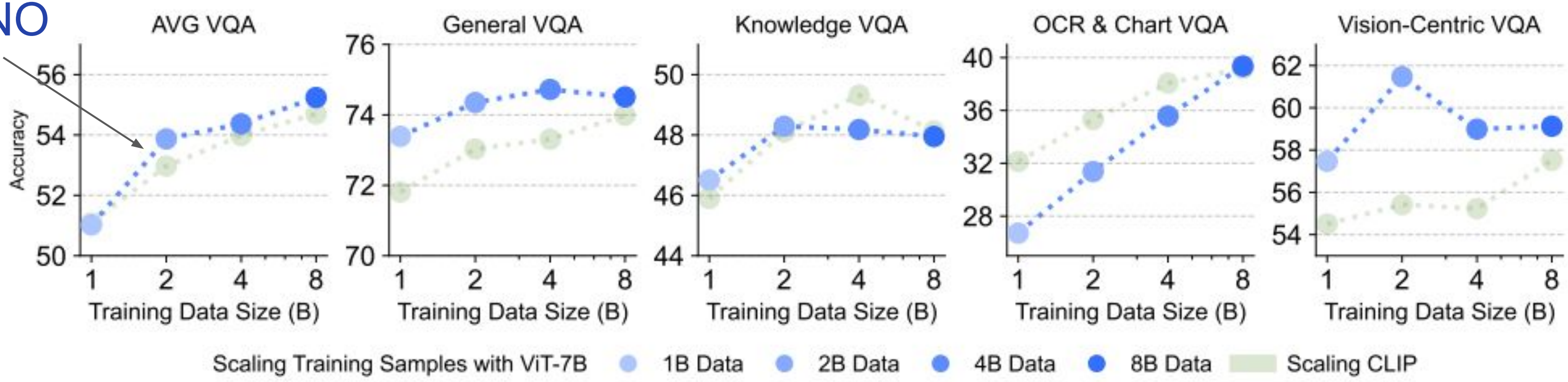
- **Data**: *MC-2B*:
  - 1 billion samples seen
  - 2 billion samples seen
  - 4 billion samples seen
  - 8 billion samples seen
- **Model**: ViT-7B
- **Method**: DINOv2 (SSL) vs. CLIP (Language-Supervised)
- **Eval**: Use VQA as evaluation.

WebSSL: Scaling Up Data

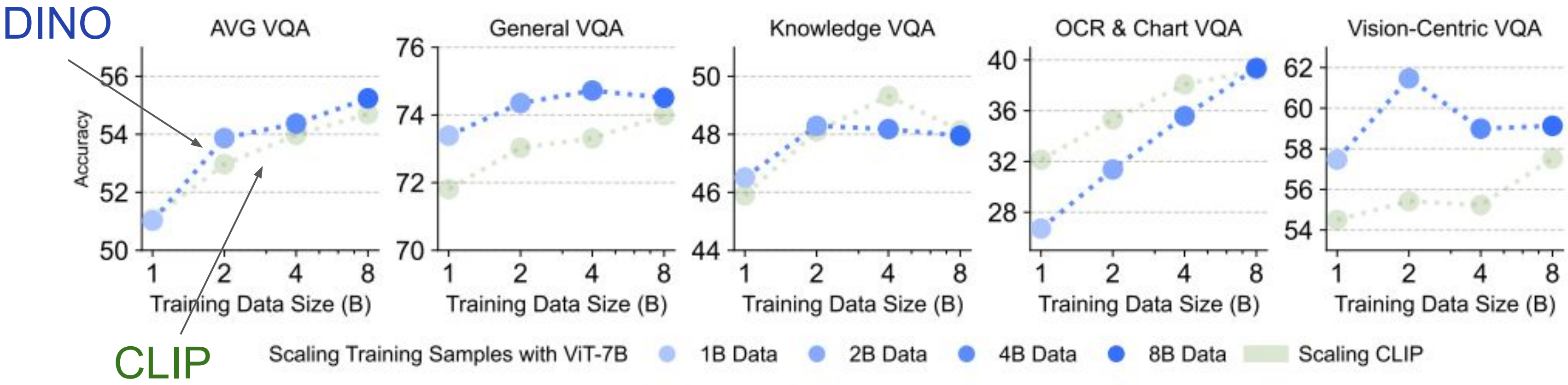


WebSSL: Scaling Up Data

DINO

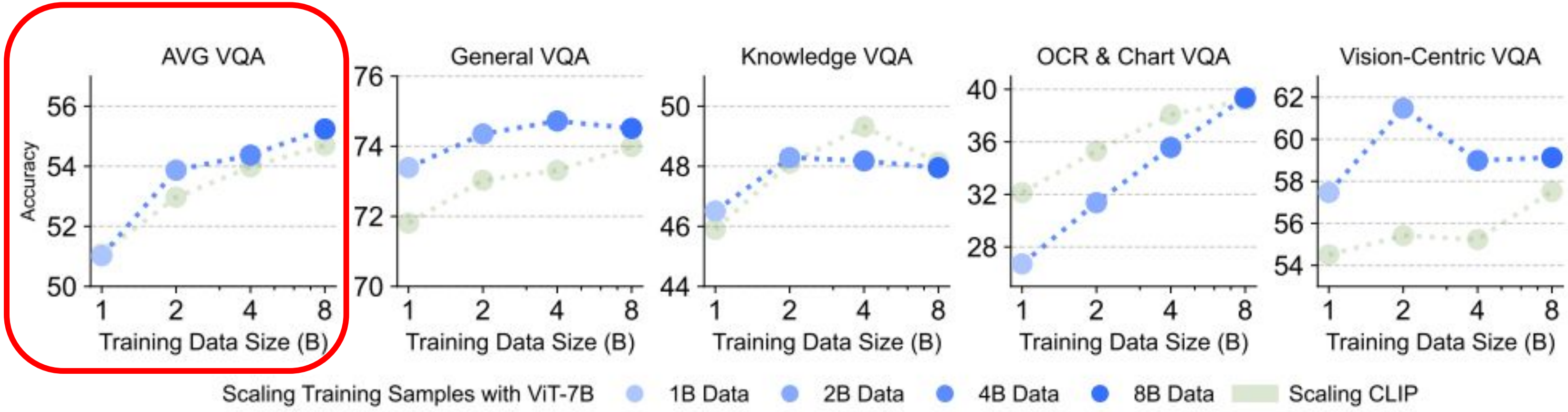


WebSSL: Scaling Up Data



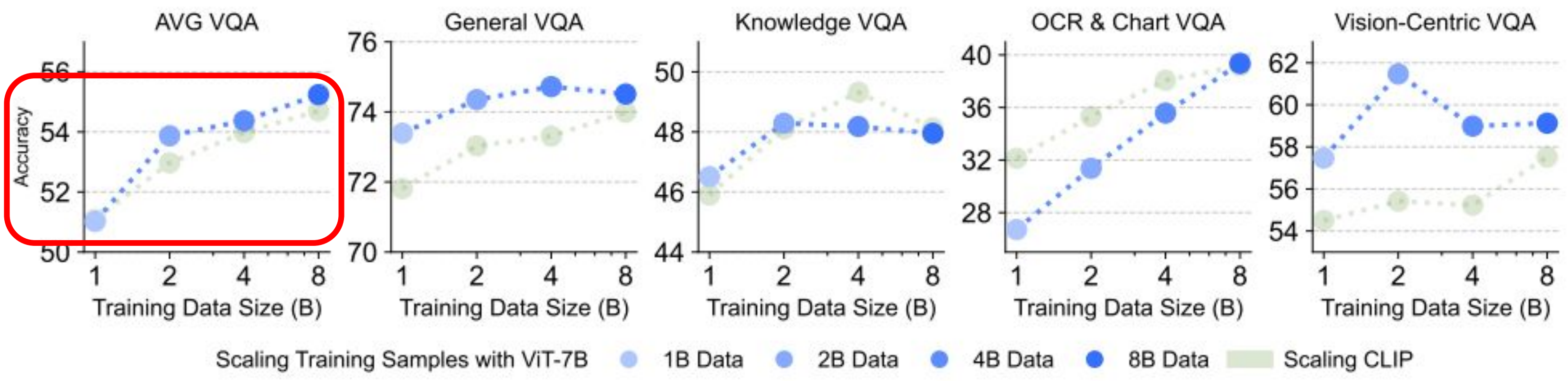
WebSSL: Scaling Up Data

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



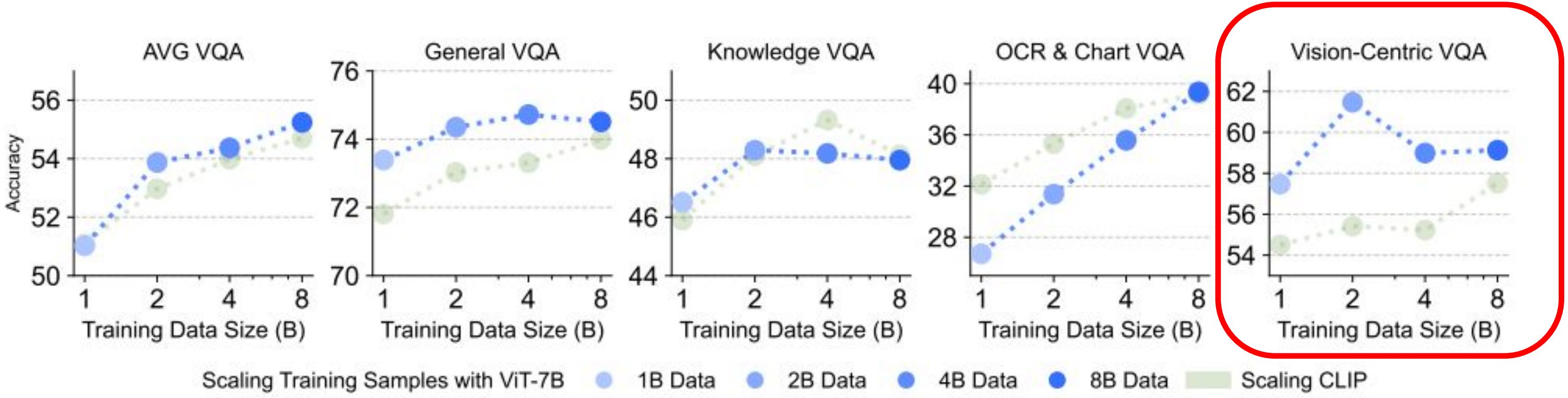
WebSSL: Scaling Up Data

1. Model improves *w.r.t* to more data seen
2. SSL models consistently outperform CLIP models at all data sizes



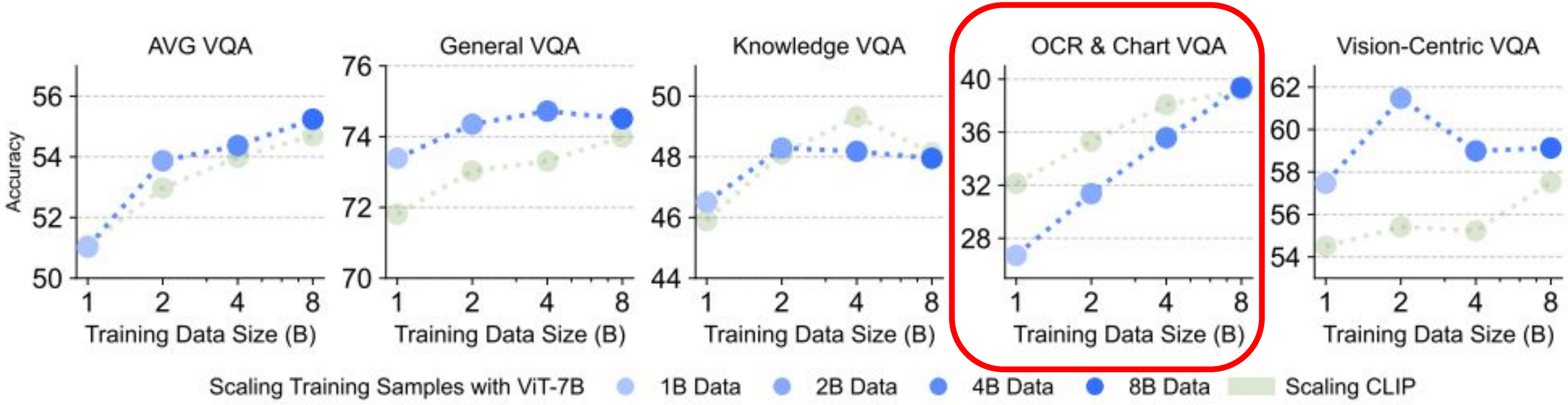
WebSSL: Scaling Up Data

1. Model improves *w.r.t* to more data seen
2. SSL models consistently outperform CLIP models at all data sizes
3. SSL models are better "visual" models



WebSSL: Scaling Up Data

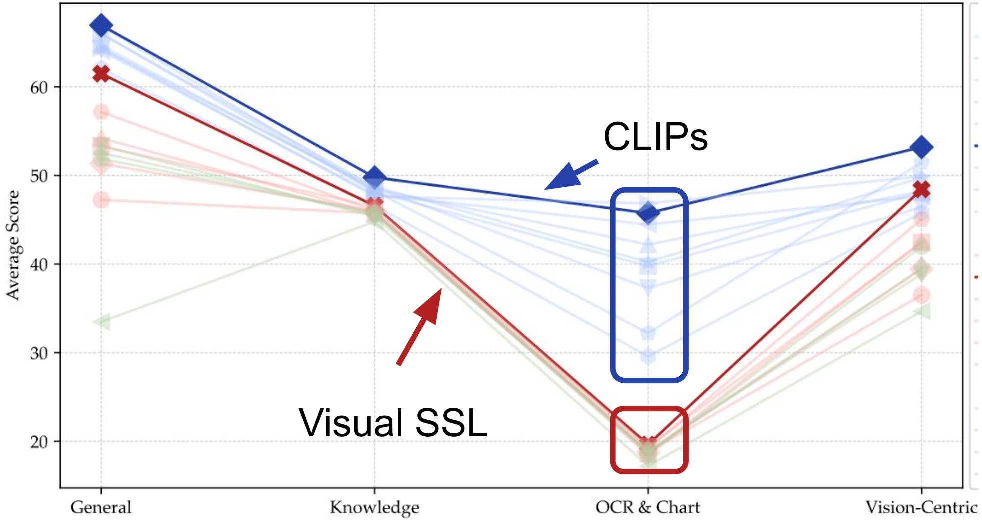
1. Model improves *w.r.t* to more data seen
2. SSL models consistently outperform CLIP models at all data sizes
3. SSL models are better “visual” models
4. Gap closes on OCR & Chart!



WebSSL: Scaling Up Data

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

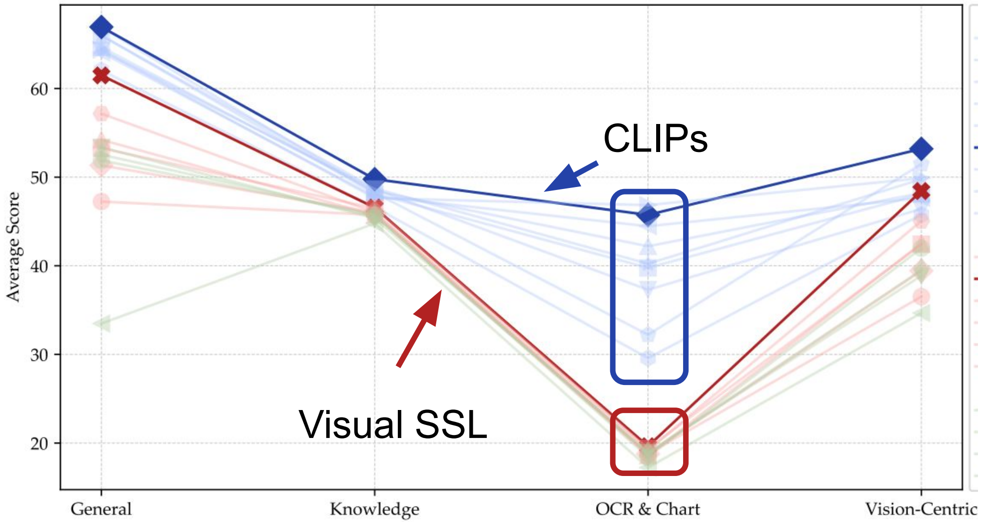
Before



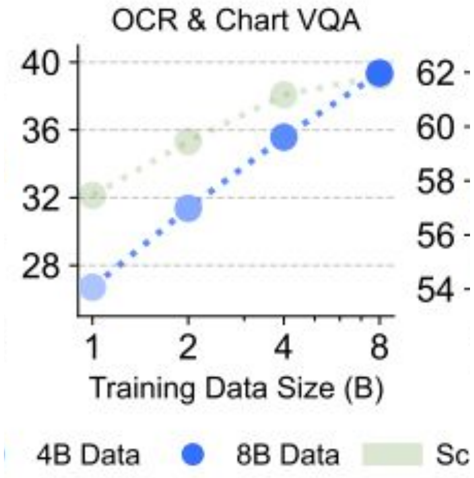
WebSSL: Scaling Up Data

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

Before



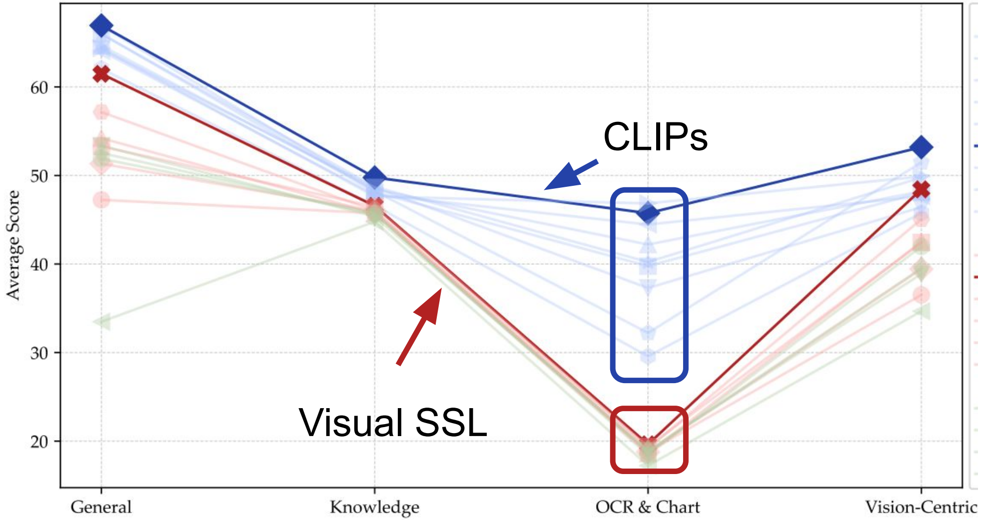
Now



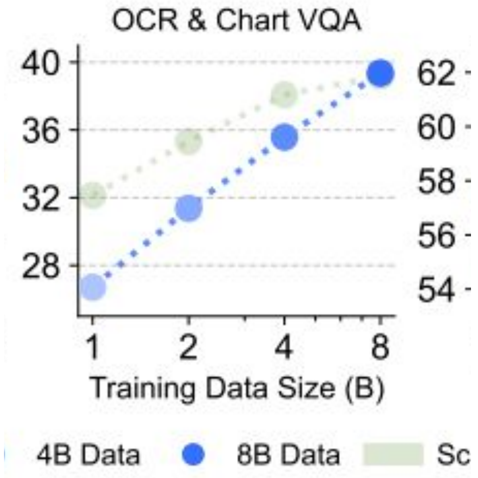
WebSSL: Scaling Up Data

VQA capability is **not unique** to language-supervised vision encoders!
SSL vision encoders can do just as well at scale :)

Before



Now



Takeaways from Scaling Up WebSSL

SSL performance improves with …

1. Larger model size
2. More data seen

SSL scales better than CLIP and is competitive with CLIP when controlling for the data.

Takeaways from Scaling Up WebSSL

SSL performance improves with …

1. Larger model size
2. More data seen

SSL scales better than CLIP and is competitive with CLIP when controlling for the data.

So it’s more about the **data**, not language supervision!

Deep Dive and Analysis

Deep Dive and Analysis

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

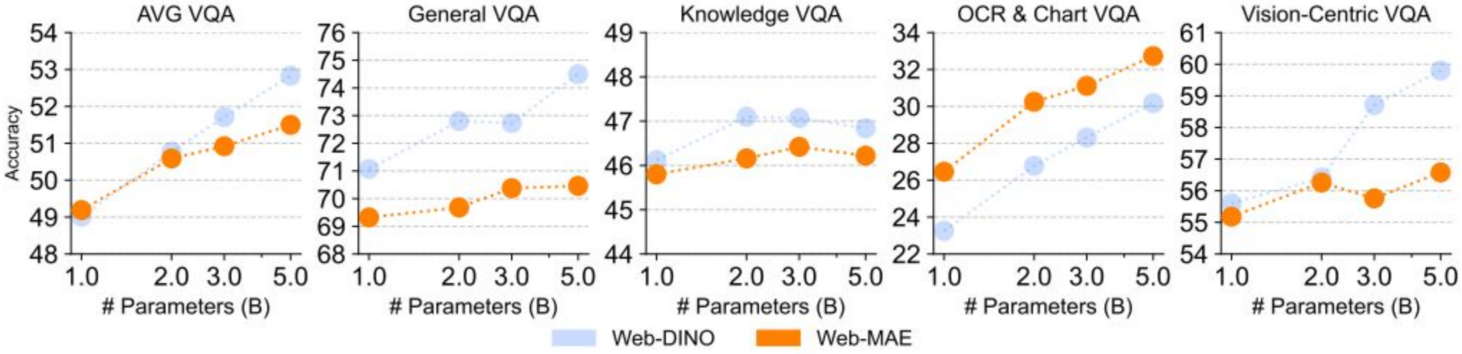
Q1. Does the observed scaling behavior generalize to other visual SSL methods?

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

Answer: we conduct similar experiments on MAE (another SSL method) to see if the behavior is unique to DINO or not

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

Answer: we conduct similar experiments on MAE (another SSL method)

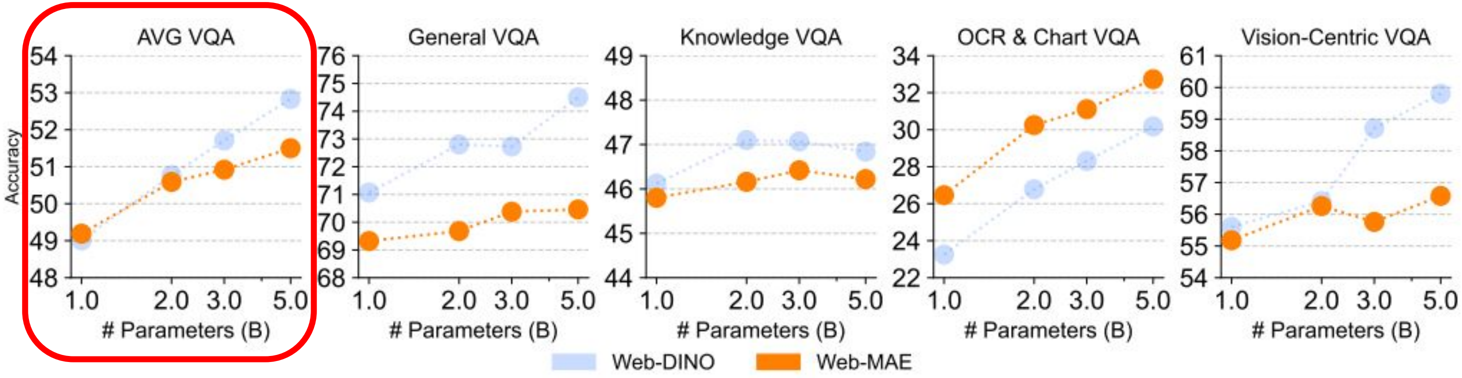


He, K., et al. (2021). Masked Autoencoders Are Scalable Vision Learners.

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

Answer: we conduct similar experiments on MAE (another SSL method)

1. MAE improves as well when trained on web-scale images!

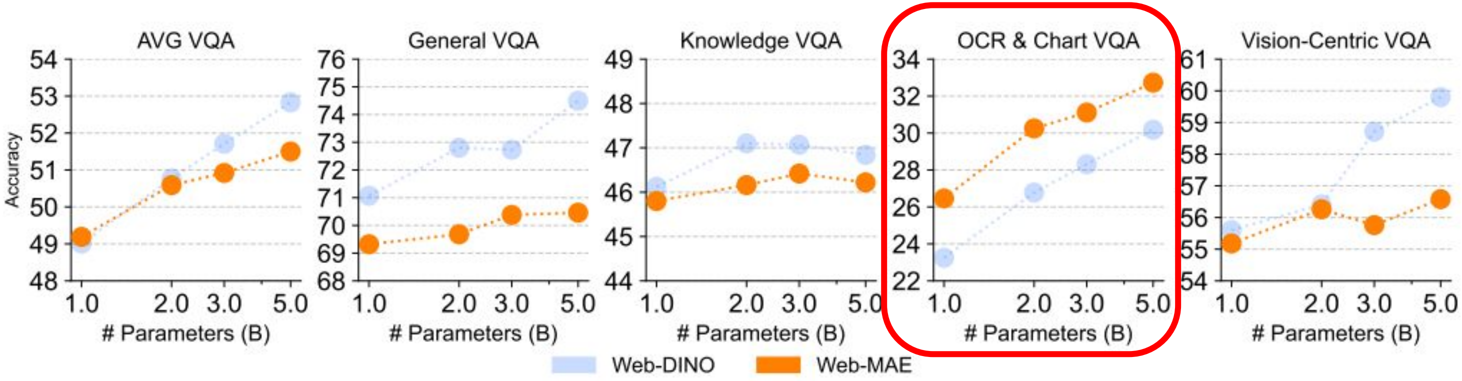


He, K., et al. (2021). Masked Autoencoders Are Scalable Vision Learners.

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

Answer: we conduct similar experiments on MAE (another SSL method)

1. MAE improves as well when trained on web-scale images!
2. Yet different SSL methods still learn different features
    a. MAE is consistently better than DINO at OCR & Chart



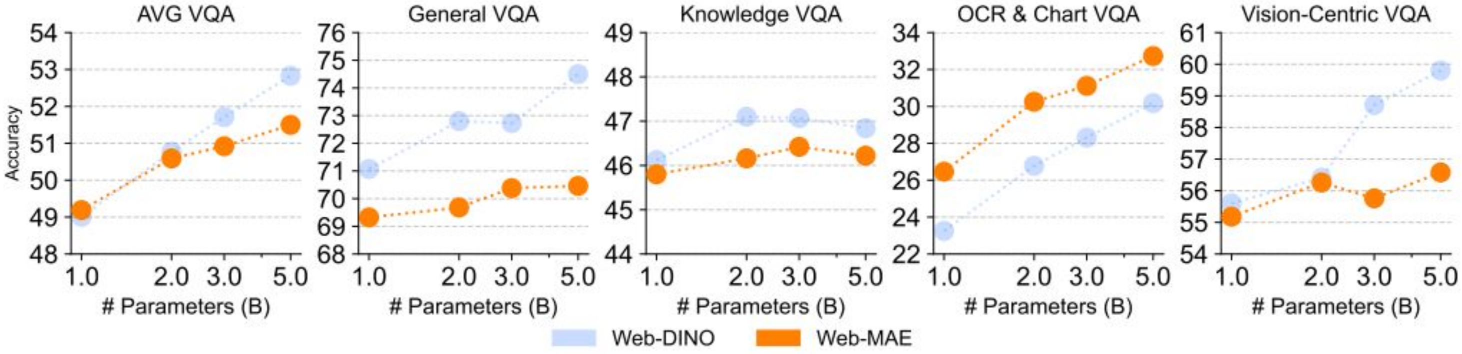
He, K., et al. (2021). Masked Autoencoders Are Scalable Vision Learners.

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

Answer: we conduct similar experiments on MAE (another SSL method)

1. MAE improves as well when trained on web-scale images!
2. Yet different SSL methods still learn different features
   a. MAE is consistently better than DINO at OCR & Chart

Yes, the observed behavior generalizes to other SSL methods!



He, K., et al. (2021). Masked Autoencoders Are Scalable Vision Learners.

Deep Dive and Analysis

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

    A: Yes, it does!

Deep Dive and Analysis

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

   A:Yes, it does!

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

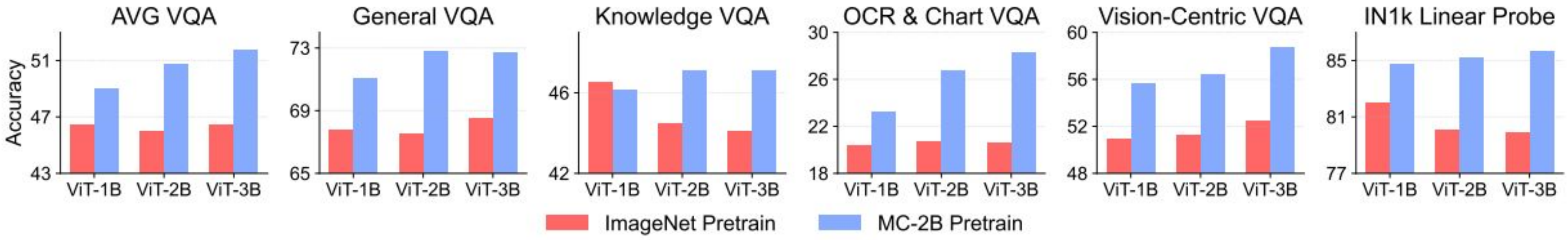
Q2.Does visual SSL exhibit similar scaling behavior on smaller scale conventional data such as ImageNet?

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

Answer: we conduct similar experiments training on ImageNet-1k

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

Answer: we conduct similar experiments training on ImageNet-1k

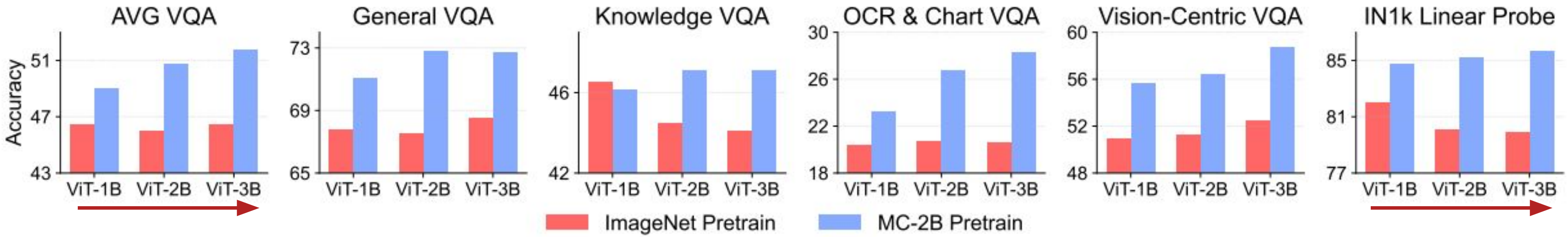


Q2.Does visual SSL exhibit similar scaling behavior on smaller scale conventional data such as ImageNet?

Answer: we conduct similar experiments training on ImageNet-1k

No obvious scaling on both VQA and ImageNet-1k evaluation.

We need large and diverse data in order to scale SSL.



Deep Dive and Analysis

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

    A: Yes, it does!

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

    A: No, it doesn’t. We need large and diverse data.

Deep Dive and Analysis

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

   A: Yes, it does!

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

   A: No, it doesn’t. We need large data.

3. How do WebSSL models perform on classic vision tasks?

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

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

Answer: Evaluate our trained Web-DINO on classic vision benchmarks with linear probes.

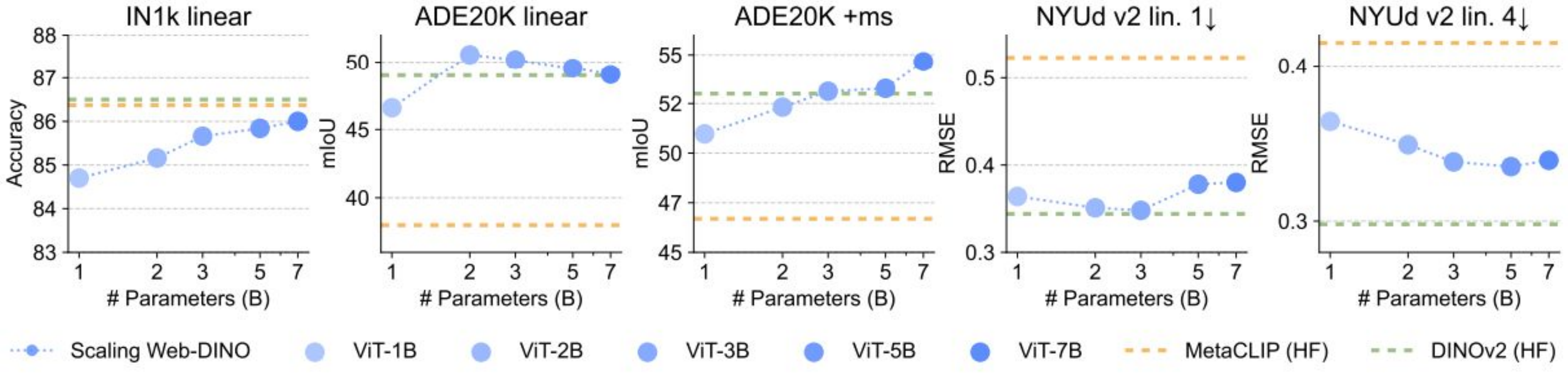
Q3. How do WebSSL models perform on classic vision tasks?

Answer: Evaluate our trained Web-DINO on classic vision benchmarks with linear probes.

- Classification:
  - ImageNet-1k
- Segmentation:
  - ADE20k (last layer)
  - ADE20k (multi-scale)
- Depth Estimation:
  - NYUd v2 (last layer)
  - NYUd v2 (four layers)

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

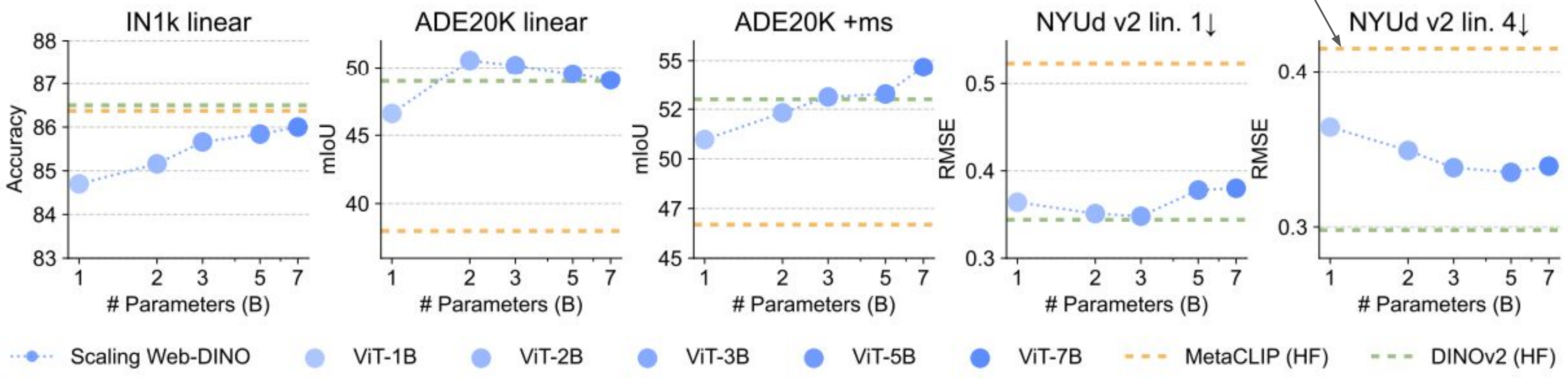
Answer: Evaluate our trained Web-DINO on classic vision benchmarks with linear probes.



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

Answer: Evaluate our trained Web-DINO on classic vision benchmarks

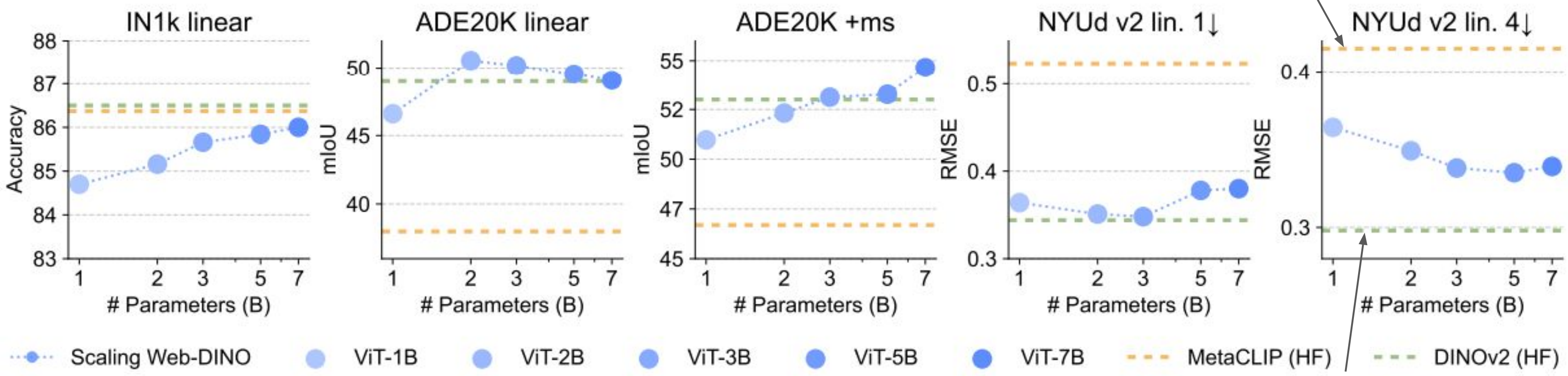
Best off-shelf MetaCLIP



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

Answer: Evaluate our trained Web-DINO on classic vision benchmarks

Best off-shelf
MetaCLIP

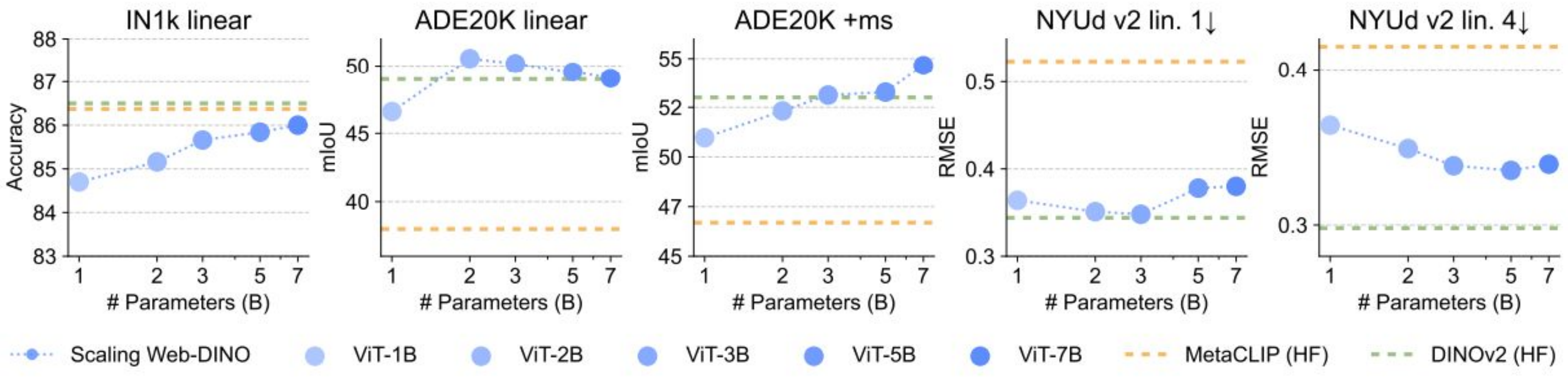


Best off-shelf
DINOv2

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

Answer: Evaluate our trained Web-DINO on classic vision benchmarks

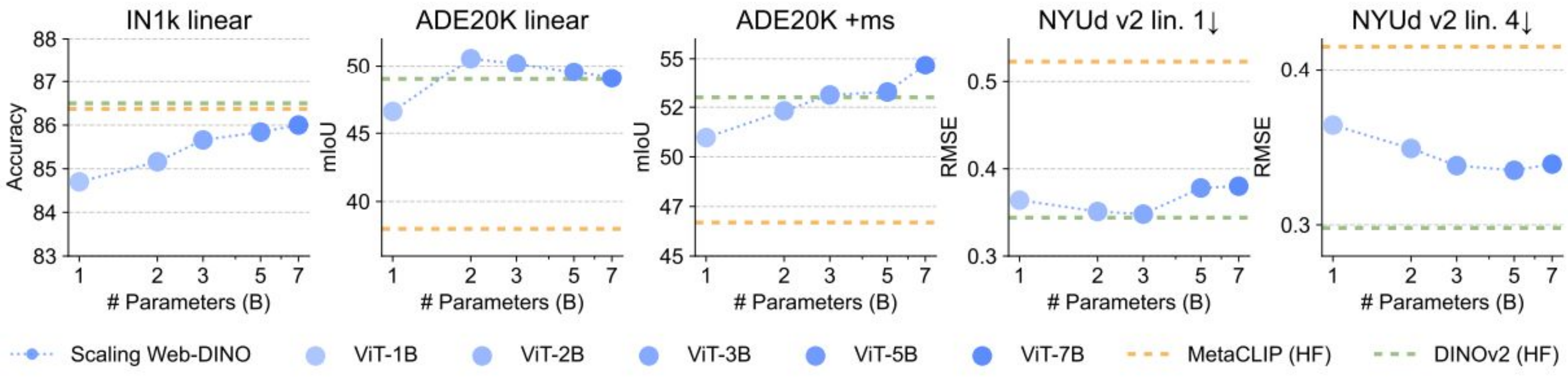
1. Web-DINO is mostly better than MetaCLIP



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

Answer: Evaluate our trained Web-DINO on classic vision benchmarks

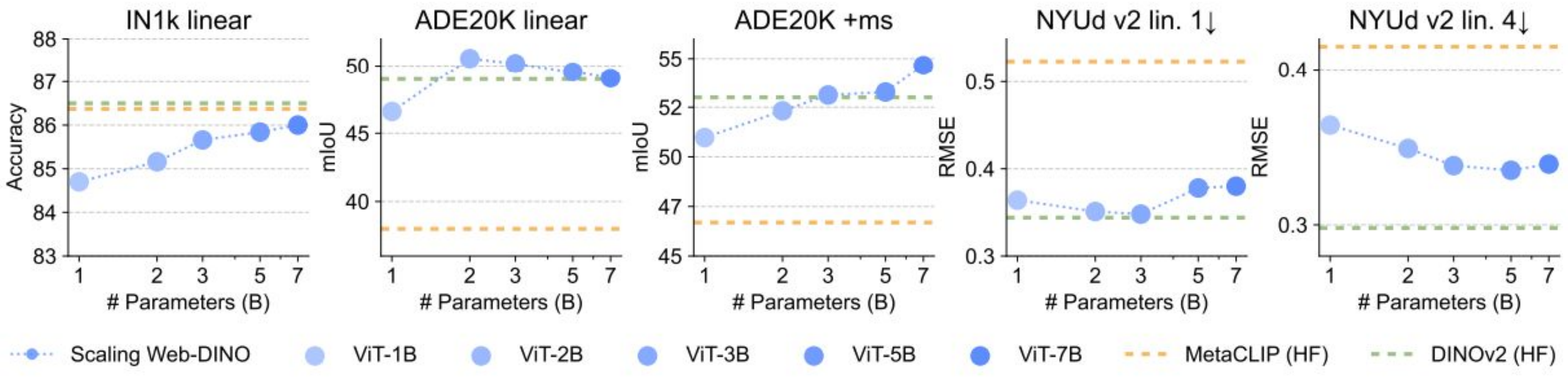
1. Web-DINO is mostly better than MetaCLIP
2. Web-DINO remains competitive with DINOv2



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

Answer: 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 (DINOv2 train data) is retrieved from classic vision tasks.



Deep Dive and Analysis

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

    A: Yes, it does!

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

    A: No, it doesn’t. We need large data.

3. How do WebSSL models perform on classic vision tasks?

    A: Better than CLIP models and competitive with DINOv2.

Deep Dive and Analysis

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

    A: Yes, it does!

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

    A: No, it doesn't. We need large data.

3. How do WebSSL models perform on classic vision tasks?

    A: Better than CLIP models and competitive with DINOv2.

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

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

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	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)

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)

Trained on images **containing *any* text**

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	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)

Trained on images **containing charts, documents, *heavy* text …**

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

1. Huge boost on OCR & Chart

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)

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

1. Huge boost on OCR & Chart
2. Other categories does not change much (no loss of generality)

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)

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

1. Huge boost on OCR & Chart
2. Other categories does not change much (no loss of generality)
3. Beats same-size clip CLIP models, even on OCR & Chart.

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)

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

1. Huge boost on OCR & Chart
2. Other categories does not change much (no loss of generality)
3. Beats same-size clip CLIP models, even on OCR & Chart.

The "text" in images contributes to improved OCR & Chart ability, and SSL methods can implicitly learn this from the data.

		VQA Evaluator					Breakdown of OCR & Chart Tasks			
Method	% of MC-2B	AVG	General	Knowledge	Vision Centric	OCR Chart	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)

Deep Dive and Analysis

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

    A: Yes, it does!

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

    A: No, it doesn't. We need large data.

3. How do WebSSL models perform on classic vision tasks?

    A: It is better than CLIP models and competitive with DINOv2.

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

    A: Because SSL models learn from text information embed in images.

Deep Dive and Analysis

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

   A: Yes, it does!

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

   A: No, it doesn't. We need large data.

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

   A: It is better than CLIP models and competitive with DINOv2.

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

   A: Because SSL models learn from text information embed in images.

5. 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?

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”

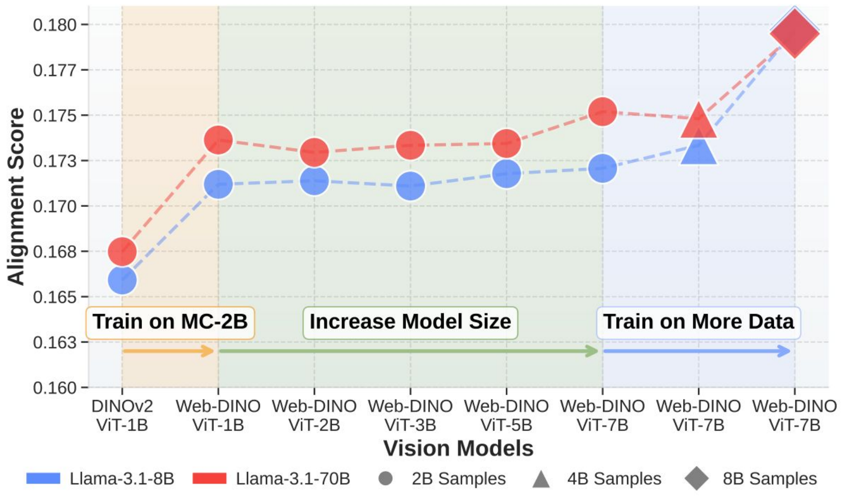
Huh, M., et al. (2024). The Platonic Representation Hypothesis.

Platonic Representation Measurements

- Frozen visual encoder + off-shelf LLM (no post-training / alignment)
- Uses 1024 Samples from WiT-1024 (A image-text dataset based on Wikipedia)
- Compute the representation from Vision Model ([cls]) and Language Model ([avg])
- For each [Image, Text], compute k=10 nearest neighbors each, measure how many overlap.
  - If 2 neighbors overlap, alignment score = 2/10 = 0.2
- Alignment Score is the average alignment score across all samples

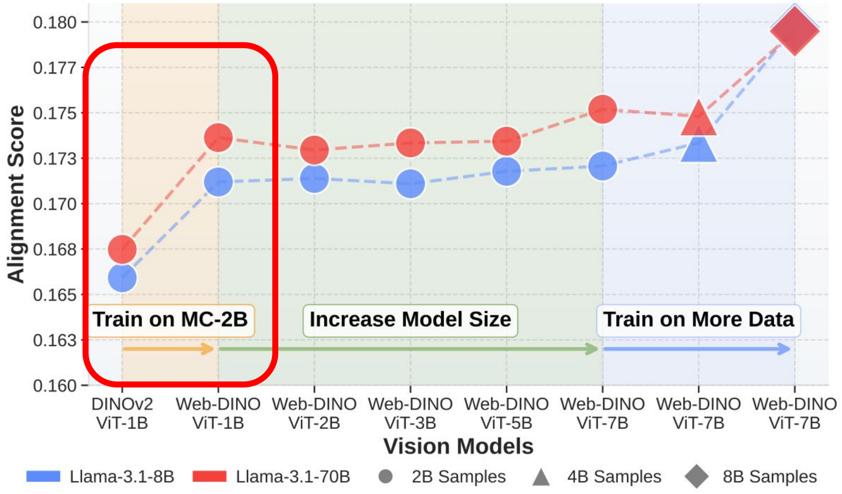
Huh, M., et al. (2024). The Platonic Representation Hypothesis.

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

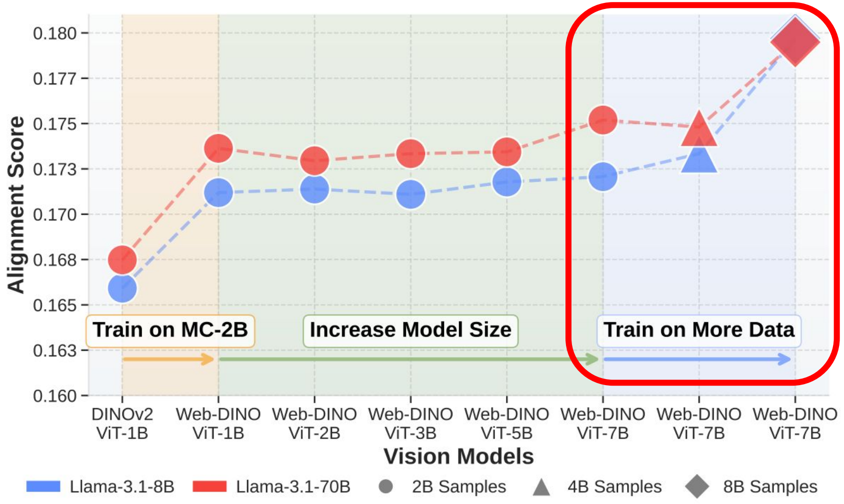


Q5. Why can SSL learn strong visual representations for multimodal modeling, without language supervision?

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 SSL scales to larger models or more data, its representation naturally aligns more with off-shelf LLMs

… without any explicit alignment!

Deep Dive and Analysis

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

   A: Yes, it does!

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

   A: No, it doesn't. We need large data.

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

   A: It is better than CLIP models and competitive with DINOv2.

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

   A: Because SSL models learn from text information embed in images.

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

   A: As SSL scales larger or train longer, the representation intrinsically aligns more with off-shelf LLMs, without any explicit alignment.

How Does WebSSL Compare with *SOTA*?

(Now the system-level comparisons are no longer apples-to-apples)

How Does WebSSL Compare with *SOTA*?

Model				MLLM Evaluator					Classic Vision Tasks				
Method	Pretrain Data	Pretrain Samples Seen	Res	AVG	General	Knowledge	OCR & Chart	Vision-Centric	IN1k lin.	ADE20K lin.	ADE20K ms.	NYUd lin. 1 (↓)	NYUd lin. 4 (↓)
Language-Supervised Models													
SigLIP ViT-SO400M	WebLI	45.0B	224	55.4	74.4	48.7	39.5	58.9	86.5	36.5	38.0	0.607	0.525
			384	60.0	76.3	50.4	53.5	59.7	87.3	39.5	47.2	0.582	0.438
SigLIP2 ViT-SO400M	WebLI	45.0B	224	56.3	74.4	50.7	42.1	58.1	87.5	41.1	44.2	0.562	0.539
			384	62.0	76.6	51.9	58.4	61.0	88.1	43.5	50.2	0.524	0.469
MetaCLIP ViT-G	MetaCLIP	12.8B	224	54.8	75.5	48.2	37.3	58.4	86.4	38.0	46.7	0.524	0.415
Visual Self-Supervised Models													
MAE ViT-H	ImageNet-1k	2.0B	224	45.2	64.6	43.9	20.6	51.7	76.6	33.3	30.7	0.517	0.483
I-JEPA ViT-H	ImageNet-22k	0.9B	224	44.7	65.4	43.9	21.2	48.4	68.8	31.6	34.6	0.548	0.520
DINOv2 ViT-g	LVD-142M	1.9B	518	47.9	70.2	45.0	21.2	55.3	86.0	49.0	53.0	0.344	0.298
Web-DINO ViT-7B	MC-2B	8.0B	224	55.2	74.5	48.0	39.4	59.1	86.5	42.1	52.6	0.491	0.376
			378	57.4	73.9	47.7	50.4	57.7	86.3	42.3	53.1	0.498	0.366
			518	59.9	75.5	48.2	55.1	60.8	86.4	42.6	52.8	0.490	0.362

How Does WebSSL Compare with *SOTA*?

1. WebSSL is competitive with CLIP models on VQA, even when using less data.

Model				MLLM Evaluator					Classic Vision Tasks				
Method	Pretrain Data	Pretrain Samples Seen	Res	AVG	General	Knowledge	OCR & Chart	Vision-Centric	IN1k lin.	ADE20K lin.	ADE20K ms.	NYUd lin. 1 (↓)	NYUd lin. 4 (↓)
Language-Supervised Models													
SigLIP ViT-SO400M	WebLI	45.0B	224	55.4	74.4	48.7	39.5	58.9	86.5	36.5	38.0	0.607	0.525
			384	60.0	76.3	50.4	53.5	59.7	87.3	39.5	47.2	0.582	0.438
SigLIP2 ViT-SO400M	WebLI	45.0B	224	56.3	74.4	50.7	42.1	58.1	87.5	41.1	44.2	0.562	0.539
			384	62.0	76.6	51.9	58.4	61.0	88.1	43.5	50.2	0.524	0.469
MetaCLIP ViT-G	MetaCLIP	12.8B	224	54.8	75.5	48.2	37.3	58.4	86.4	38.0	46.7	0.524	0.415
Visual Self-Supervised Models													
MAE ViT-H	ImageNet-1k	2.0B	224	45.2	64.6	43.9	20.6	51.7	76.6	33.3	30.7	0.517	0.483
I-JEPA ViT-H	ImageNet-22k	0.9B	224	44.7	65.4	43.9	21.2	48.4	68.8	31.6	34.6	0.548	0.520
DINOv2 ViT-g	LVD-142M	1.9B	518	47.9	70.2	45.0	21.2	55.3	86.0	49.0	53.0	0.344	0.298
Web-DINO ViT-7B	MC-2B	8.0B	224	55.2	74.5	48.0	39.4	59.1	86.5	42.1	52.6	0.491	0.376
			378	57.4	73.9	47.7	50.4	57.7	86.3	42.3	53.1	0.498	0.366
			518	59.9	75.5	48.2	55.1	60.8	86.4	42.6	52.8	0.490	0.362

How Does WebSSL Compare with *SOTA*?

1. WebSSL is competitive with CLIP models on VQA, even when using less data.
2. And better than CLIP models on classic vision.

Method	Pretrain Data	Pretrain Samples Seen	Res	AVG	General	Knowledge	OCR & Chart	Vision-Centric	IN1k lin.	ADE20K lin.	ADE20K ms.	NYUd lin. 1 (↓)	NYUd lin. 4 (↓)
				MLLM Evaluator					**Classic Vision Tasks**				
Language-Supervised Models													
SigLIP ViT-SO400M	WebLI	45.0B	224	55.4	74.4	48.7	39.5	58.9	86.5	36.5	38.0	0.607	0.525
			384	60.0	76.3	50.4	53.5	59.7	87.3	39.5	47.2	0.582	0.438
SigLIP2 ViT-SO400M	WebLI	45.0B	224	56.3	74.4	50.7	42.1	58.1	87.5	41.1	44.2	0.562	0.539
			384	62.0	76.6	51.9	58.4	61.0	88.1	43.5	50.2	0.524	0.469
MetaCLIP ViT-G	MetaCLIP	12.8B	224	54.8	75.5	48.2	37.3	58.4	86.4	38.0	46.7	0.524	0.415
Visual Self-Supervised Models													
MAE ViT-H	ImageNet-1k	2.0B	224	45.2	64.6	43.9	20.6	51.7	76.6	33.3	30.7	0.517	0.483
I-JEPA ViT-H	ImageNet-22k	0.9B	224	44.7	65.4	43.9	21.2	48.4	68.8	31.6	34.6	0.548	0.520
DINOv2 ViT-g	LVD-142M	1.9B	518	47.9	70.2	45.0	21.2	55.3	86.0	49.0	53.0	0.344	0.298
Web-DINO ViT-7B	MC-2B	8.0B	224	55.2	74.5	48.0	39.4	59.1	86.5	42.1	52.6	0.491	0.376
			378	57.4	73.9	47.7	50.4	57.7	86.3	42.3	53.1	0.498	0.366
			518	59.9	75.5	48.2	55.1	60.8	86.4	42.6	52.8	0.490	0.362

How Does WebSSL Compare with *SOTA*?

WebSSL also improves with higher resolution (more room for improvement!)

Method	Pretrain Data	Pretrain Samples Seen	Res	AVG	General	Knowledge	OCR & Chart	Vision-Centric	IN1k lin.	ADE20K lin.	ADE20K ms.	NYUd lin. 1 (↓)	NYUd lin. 4 (↓)
Language-Supervised Models													
SigLIP ViT-SO400M	WebLI	45.0B	224	55.4	74.4	48.7	39.5	58.9	86.5	36.5	38.0	0.607	0.525
			384	60.0	76.3	50.4	53.5	59.7	87.3	39.5	47.2	0.582	0.438
SigLIP2 ViT-SO400M	WebLI	45.0B	224	56.3	74.4	50.7	42.1	58.1	87.5	41.1	44.2	0.562	0.539
			384	62.0	76.6	51.9	58.4	61.0	88.1	43.5	50.2	0.524	0.469
MetaCLIP ViT-G	MetaCLIP	12.8B	224	54.8	75.5	48.2	37.3	58.4	86.4	38.0	46.7	0.524	0.415
Visual Self-Supervised Models													
MAE ViT-H	ImageNet-1k	2.0B	224	45.2	64.6	43.9	20.6	51.7	76.6	33.3	30.7	0.517	0.483
I-JEPA ViT-H	ImageNet-22k	0.9B	224	44.7	65.4	43.9	21.2	48.4	68.8	31.6	34.6	0.548	0.520
DINOv2 ViT-g	LVD-142M	1.9B	518	47.9	70.2	45.0	21.2	55.3	86.0	49.0	53.0	0.344	0.298
Web-DINO ViT-7B	MC-2B	8.0B	224	55.2	74.5	48.0	39.4	59.1	86.5	42.1	52.6	0.491	0.376
			378	57.4	73.9	47.7	50.4	57.7	86.3	42.3	53.1	0.498	0.366
			518	59.9	75.5	48.2	55.1	60.8	86.4	42.6	52.8	0.490	0.362

Takeaways

- Visual SSL improves *w.r.t* to model and data sizes when we use VQA as evaluation

Takeaways

- Visual SSL improves *w.r.t* to model and data sizes when we use VQA as evaluation
- The gap between SSL and CLIP models partly (largely) comes from **data,** not language supervision

Takeaways

- Visual SSL improves *w.r.t* to model and data sizes when we use VQA as evaluation
- The gap between SSL and CLIP models partly (largely) come from **data,** not language supervision
- Visual SSL is competitive with CLIP models on VQA, even on OCR & Chart

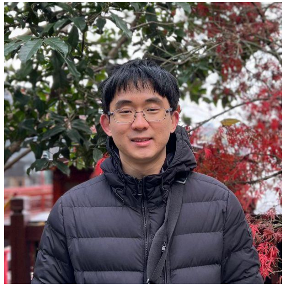
Takeaways

- Visual SSL improves *w.r.t* to model and data sizes when we use VQA as evaluation
- The gap between SSL and CLIP models partly (largely) come from data not language supervision
- Visual SSL is competitive with CLIP models on VQA, even on OCR & Chart
- Visual SSL has its unique benefits
  - Vision-centric VQA
  - Classic vision benchmarks
  - Easy to train on raw images (no need for text curation)

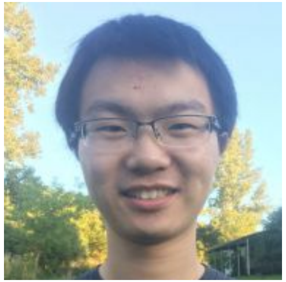
Takeaways

- Visual SSL improves *w.r.t* to model and data sizes when we use VQA as evaluation
- The gap between SSL and CLIP models partly (largely) come from data not language supervision
- Visual SSL is competitive with CLIP models on VQA, even on OCR & Chart
- Visual SSL has its unique benefits
  - Vision-centric VQA
  - Classic vision benchmarks
  - Easy to train on raw images (no need for text curation)
- We can continue to train better SSL models! (Better / More Data, Larger Model, …)

Thanks to Our Amazing Team!!!








Thank you!

Please visit us at Poster #25

(Tuesday 11:45 AM - 1:45 PM)

Open-sourced at:

https://davidfan.io/webssl/

https://github.com/facebookresearch/webssl