Scaling Language-Free Visual Representation Learning

David Fan[1*] Shengbang Tong[1,2*] Jiachen Zhu[1,2] Koustuv Sinha[1] Zhuang Liu[1,3]
Xinlei Chen[1] Michael Rabbat[1] Nicolas Ballas[1] Yann LeCun[1,2] Amir Bar[1†] Saining Xie[2†]
[*]equal contribution [†]equal advising
[1]FAIR, Meta [2]New York University [3]Princeton University

Abstract

Visual Self-Supervised Learning (SSL) currently underperforms Contrastive Language-Image Pretraining (CLIP) in multimodal settings such as Visual Question Answering (VQA). This multimodal gap is often attributed to the semantics introduced by language supervision, even though visual SSL and CLIP models are often trained on different data. In this work, we ask the question: "Do visual self-supervised approaches lag behind CLIP due to the lack of language supervision, or differences in the training data?" We study this question by training both visual SSL and CLIP models on the same MetaCLIP data, and leveraging VQA as a diverse testbed for vision encoders. In this controlled setup, visual SSL models scale better than CLIP models in terms of data and model capacity, and visual SSL performance does not saturate even after scaling up to 7B parameters. Consequently, we observe visual SSL methods achieve CLIP-level performance on a wide range of VQA and classic vision benchmarks. These findings demonstrate that pure visual SSL can match language-supervised visual pretraining at scale, opening new opportunities for vision-centric representation learning. Code and models here.

1. Introduction

Visual representation learning has evolved along two distinct paths with different training approaches. Language-supervised methods such as Contrastive Language-Image Pretraining (CLIP) [75, 85, 108] use paired image-text data to learn representations that are enriched with linguistic semantics. Self-Supervised Learning (SSL) methods [14, 44, 55, 73, 109] learn from images alone.

Despite SSL models outperforming language-supervised models on classic vision tasks such as classification and segmentation [73], they are less commonly adopted in recent multimodal large language models (MLLMs) [1, 2, 9, 56–58, 91]. This difference in adoption is partially due to a performance gap in visual question answering (see Fig. 1), particularly for OCR & Chart interpretation tasks [78, 91].

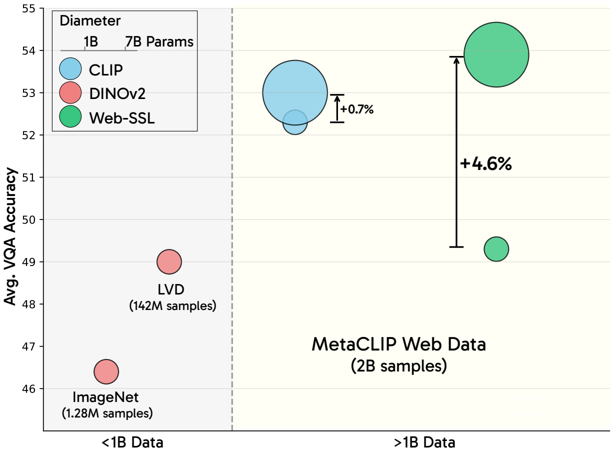

Figure 1. We compare the scaling behavior of visual SSL and CLIP on 16 VQA tasks from the Cambrian-1 suite under different data and model size regimes. Prior visual SSL methods achieved strong performance on classic vision tasks, but have underperformed as encoders for multimodal instruction-tuned VQA tasks. Our results show that with appropriate scaling of models and data, visual SSL can match the performance of language-supervised models across all evaluated domains—even OCR & Chart.

Beyond methodology differences, these approaches have also been separated by data scale and distribution (Fig. 1). CLIP models typically train on billion-scale image-text pairs from the web [18, 77, 104], while SSL methods use million-scale datasets such as ImageNet [24] or hundred-million scale data with ImageNet-like distributions [73, 76].

In this work, we investigate a fundamental question: *Is language supervision necessary to pretrain visual representations for multimodal modeling?* Rather than seeking to replace language-supervised approaches, we aim to understand the intrinsic capabilities and limitations of visual self-supervision at scale for multimodal applications. To conduct a fair comparison, we train SSL models on the same billion-scale web data used for state-of-the-art CLIP models—specifically the MetaCLIP dataset [104]. This approach controls for data distribution differences when comparing visual SSL and CLIP.

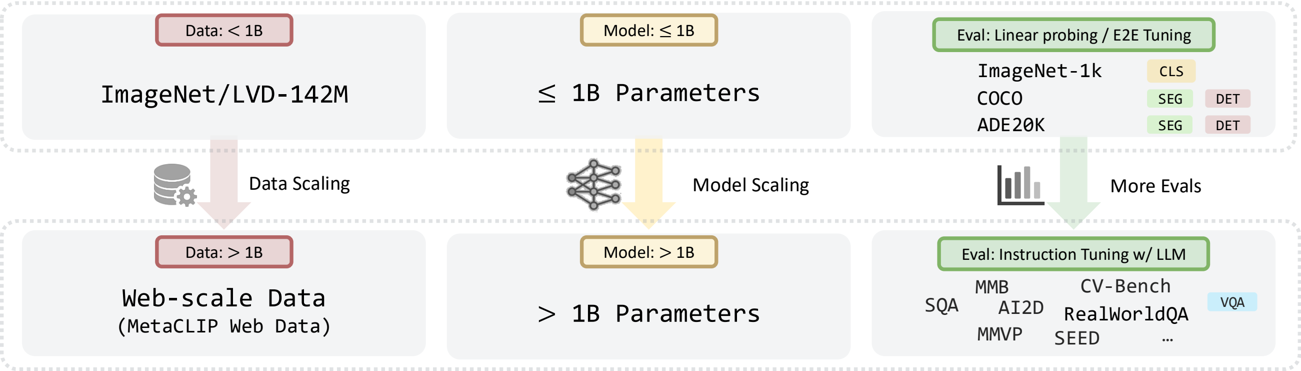

Figure 2. **Visual SSL 2.0 changes.** In this work, we adopt three improvements to the visual SSL pipeline: 1) Training on billion-scale web data, curated through the MetaCLIP [104] pipeline, to move beyond "conventional" datasets; 2) Scaling model architecture from sub-billion parameter models to models exceeding 1 billion parameters; and 3) Incorporating VQA as an evaluation protocol to comprehensively assess visual features. These changes enable us to study visual SSL at a larger scale and observe previously unseen scaling trends.

For evaluation, we primarily use visual question answering (VQA) as a framework to evaluate SSL models across a diverse set of capabilities at scale. VQA evaluation suites span vision-centric, visual reasoning, and OCR & Chart tasks, and have been shown to be a more diverse testbed for assessing vision encoders [31, 91, 95, 96], reflecting the broader perception challenges found in real-world distributions. We adopt the evaluation suite proposed in Cambrian-1 [91], which evaluates performance across 16 tasks spanning 4 distinct categories of VQA: General, Knowledge, OCR & Chart, and Vision-Centric.

We train Web-SSL, a family of visual SSL models ranging from 1 to 7 billion parameters, using the above setting for direct and controlled comparison to CLIP. As a result of our empirical study, we contribute several insights:

- Visual SSL can match and even surpass language-supervised methods for visual pretraining, on a wide range of VQA tasks—even on language-related tasks such as OCR & Chart understanding (Fig. 3).
- Visual SSL scales well with respect to model capacity (Fig. 3) and data (Fig. 4), indicating that SSL has significant untapped potential.
- Visual SSL can maintain competitive traditional vision performance on classification and segmentation, even while improving at VQA (Fig. 7).
- Training on a higher ratio of images containing text is especially effective for improving OCR & Chart performance (Question 4). Exploring data composition is a promising direction.

This work serves as a proof of concept that offers a compelling vision-centric alternative to the recent CLIP-dominated trend, and opens new opportunities for future research. Our Web-SSL vision models are open-sourced[1], and we hope to inspire the broader community to unlock the full potential of visual SSL in the multimodal era.

2. From Visual SSL 1.0 to 2.0

Here, we describe our experimental setup, which extends previous SSL works by (1) scaling data to billion-scale images (Sec. 2.1), (2) scaling model size beyond 1B parameters (Sec. 2.2), and (3) evaluating vision models using diverse VQA tasks (Sec. 2.3), in addition to classic vision benchmarks such as ImageNet-1k [24] and ADE20k [111].

2.1. Beyond ImageNet Pretraining

To study whether visual SSL can match the performance of CLIP, we start by adopting the same data that drove CLIP's success. We thus leverage the MetaCLIP dataset [103, 104], which has enabled the most successful open-source reproduction of CLIP to-date.[2] We use 2 billion samples from MetaCLIP, which we refer to as MC-2B. We train SSL methods on only the images, and CLIP on the image-text pairs. This controls for data distribution and size as confounding variables, and enables a fairer comparison of the pretraining methods themselves.

2.2. Scaling Up Vision Models to Billion Scale

We can also increase model size. Inspired by advancements in scaling language models [11, 52, 72], we train Vision Transformers (ViTs) with 1B, 2B, 3B, 5B, and 7B parameters, on only the images from MC-2B, to study the properties of larger-scale visual SSL models trained on web-scale data. We adapt ViT-g from Oquab et al. [73] as ViT-1B, and define new configurations for ViT-2B to 7B; see Appendix A for model details.

2.3. Multimodal LLMs as an Evaluation Protocol

In addition to conventional evaluation protocols, such as ImageNet-1k linear probe, we also evaluate our vision encoders using VQA, a flexible and robust evaluation proto-

[1] https://github.com/facebookresearch/webssl

[2] The data used to train the original CLIP is closed-source.

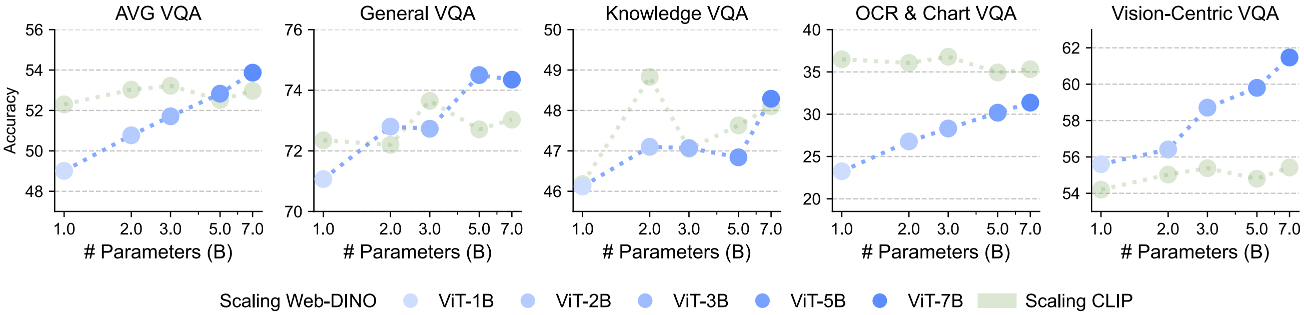

Figure 3. **Scaling behavior of Web-DINO and CLIP ViTs trained on MC-2B.** The x-axis shows model sizes from 1B to 7B parameters on a log scale. We observe novel "scaling behavior" with Web-DINO models across all categories, with particularly pronounced improvements in the OCR & Chart and Vision-Centric domains as model size increases. In contrast, CLIP models demonstrate limited scaling benefits, with performance saturating at moderate model sizes. The two model families exhibit complementary strengths: CLIP models excel at OCR & Chart VQA, and Web-DINO models are superior at Vision-Centric VQA, while remaining competitive in all other categories.

col that reflects the diversity of real-world perceptual challenges [91, 95], as shown in Fig. 2.

Here, we study all vision encoders using the same controlled setting to ensure fair comparison. Specifically, we use the same two-stage visual instruction tuning procedure and data as Cambrian-1 [91]. First, a lightweight MLP adapter is added to project the vision encoder features into the same dimensionality as the LLM, and only this MLP adapter is trained. In the second stage, both the MLP adapter and LLM are finetuned. To enable controlled comparison, the vision encoder remains frozen in both stages, and all experiments use the same training recipe as well as Llama-3 8B Instruct [94] backbone. We provide detailed training datasets and hyperparameters in Appendix A.

We then report results on the Cambrian-1 [91] evaluation suite, which is comprised of 16 VQA benchmarks spanning four established domains: General, Knowledge, OCR & Chart, and Vision-Centric. The average VQA performance is the average of the four subcategories. Each subcategory has 4 benchmarks and is equally weighted.

3. Scaling Visual SSL

In this section, we explore the scaling behavior of visual SSL models with respect to both model and data size, as a result of training on only images from MC-2B. We focus on DINOv2 [73] as the visual SSL method in this section, and discuss MAE [44] in Sec. 4.

In Sec. 3.1, we increase model size from 1B to 7B while keeping the training data fixed at 2 billion MC-2B images—unless otherwise denoted. We use the off-shelf training code and recipe for each method. In Sec. 3.2, we shift our focus to scaling total data seen for a fixed model size, and analyze how performance evolves as the number of images seen during training increases from 1 billion to 8 billion.

3.1. Scaling Model

The intention of scaling model size is both to find the ceiling of visual SSL under this new data regime, and to identify any unique behavior that emerges in larger models.

We thus pretrain DINOv2 ViT models, ranging from 1B to 7B parameters, using 2 billion unlabeled images at 224×224 resolution from MC-2B—*without* high-resolution adaptation [73]—to ensure fair comparison with CLIP. We refer to these models as Web-DINO throughout the paper. For a controlled comparison, we also train CLIP models of the same sizes on the same data.

We evaluate each model with VQA and present the results in Fig. 3. We will first discuss the overall performance trend and then turn to specific category performance. To the best of our knowledge, this is the first instance of a vision encoder trained purely with visual self-supervision achieving performance parity with language-supervised encoders on VQA—even in the OCR & Chart category, which is traditionally considered to be highly text-dependent.

Performance trend. We compare the performance trend as model capacity increases in Fig. 3. Web-DINO's Average, OCR & Chart, and Vision-Centric VQA performance improves nearly log-linearly with increasing model size, while General and Knowledge improve to a smaller degree. In contrast, CLIP's performance in all VQA categories largely saturates after 3B parameters. This suggests that while smaller CLIP models may be more data-efficient, this advantage largely dissipates for larger CLIP models. The continual improvement from increasing Web-DINO model capacity also suggests that visual SSL benefits from larger model capacity, and that scaling visual SSL past 7B parameters is a promising direction.

Category-specific performance. In terms of category-specific performance, DINO also increasingly outperforms CLIP on Vision-Centric VQA and largely closes the gap with CLIP on OCR & Chart and Average VQA (Fig. 3), as

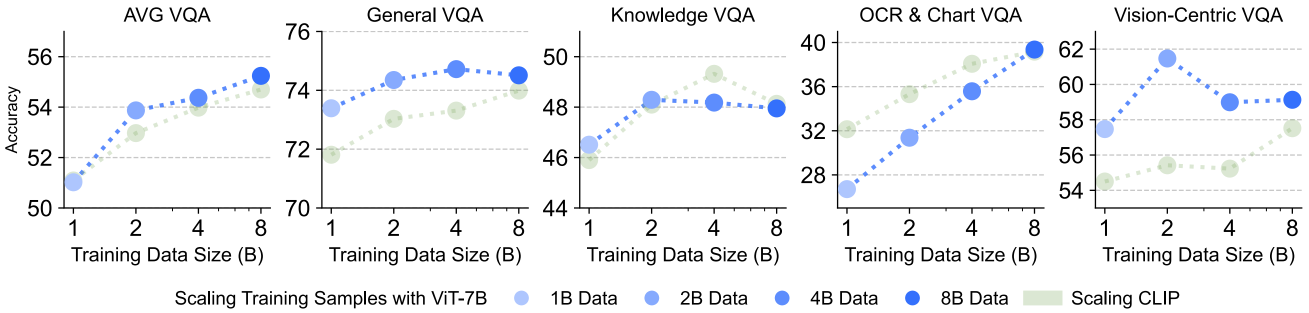

Figure 4. **Scaling up examples seen when training Web-DINO-7B.** Performance across different VQA categories as training data increases from 1B to 8B images. While General and Vision-Centric tasks show diminishing returns after 2B images, OCR & Chart tasks demonstrate continued improvement, contributing to steady gains in average performance. Further, Web-DINO consistently outperforms same-size (ViT-7B) CLIP models with different training samples seen. The x-axis plots training data size on a log-scale.

model size increases. At 5B parameters and above, DINO can exceed the Average VQA performance of CLIP. These results suggest that vision-only models, when trained on CLIP-distribution images, can develop strong visual features comparable to those of language-supervised models.

3.2. Scaling Examples Seen

Previously, we focused on single-epoch training, where each of the 2B unique images in MC-2B is seen only once. Here, we investigate the impact of increasing the number of examples seen by training Web-DINO ViT-7B on data ranging from 1 billion to 8 billion images from MC-2B.

As shown in Fig. 4, General and Knowledge VQA performance improves incrementally with more examples seen, saturating at 4B and 2B examples respectively. Vision-Centric VQA performance improves sharply from 1B to 2B examples, and saturates beyond 2B examples. In contrast, OCR & Chart is the only category that shows consistent improvement with more examples seen. This suggests that as the model sees more data, it learns a representation that is increasingly well-suited for text-related tasks, yet without marked degradation on other capabilities.

Furthermore, when compared to a CLIP model of the same size (ViT-7B), Web-DINO consistently outperforms CLIP on average VQA performance given the same number of samples seen (Fig. 4). Notably, after seeing 8B samples, Web-DINO closes the performance gap with the CLIP model on OCR & Chart VQA tasks. This provides further evidence suggesting that visual SSL models have the potential to scale better than language-supervised models.

Collectively, the results in Fig. 3 and 4 indicate that as model size and examples seen increase, visual SSL learns features that are increasingly effective for VQA in general, but especially on OCR & Chart. Our results suggest that CLIP-based models do not hold an absolute advantage compared to visual SSL. In Sec. 4, we delve deeper into the underlying mechanisms driving this trend.

4. Scaling Analysis and Findings

In Sec. 3, we demonstrated that visual SSL models scale well with model size and training set size. These observations raise further questions about the generality and implications of these phenomena. To deepen our understanding, we investigate five key aspects, including whether scaling behavior extends to other vision-only models (Question 1), if SSL models also exhibit scaling behavior on smaller and more conventional data (Question 2), and whether SSL can retain competitive performance on classic vision tasks (Question 3). Additionally, we explore why scaling particularly enhances OCR & Chart performance (Question 4), and highlight emergent properties that arise via scaling visual SSL (Question 5). Next, we provide detailed analysis.

Question 1

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

In previous sections, we derived our findings from DINOv2, a joint embedding visual SSL method. Here, we extend our analysis to a masked modeling based visual SSL method—Masked Autoencoder (MAE) [44]. We train MAE on MC-2B (denoted as Web-MAE) using ViT models ranging from 1B to 5B parameters and compare the results with Web-DINO models in Fig. 5.

Web-MAE models exhibit similar scaling behavior to Web-DINO models, with average VQA performance improving consistently as model size increases. Compared to joint embedding methods, Web-MAE models learn features that are particularly well-suited for OCR & Chart tasks but underperform in other domains. These results suggest that the "scaling behavior" observed in VQA tasks generalizes across different visual SSL methods. We also note that different visual SSL approaches learn distinct representations even when trained under the same conditions, as demonstrated by Web-MAE's OCR performance.

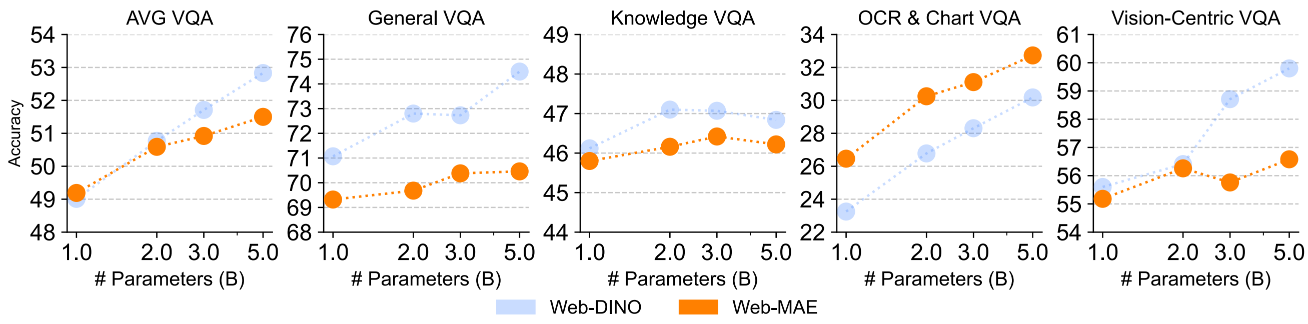

Figure 5. **Web-MAE trained on MC-2B.** Web-MAE also exhibits consistent scaling behavior as model size increases. Notably, Web-MAE demonstrates better performance in OCR & Chart tasks, achieving higher accuracy than Web-DINO across all model sizes.

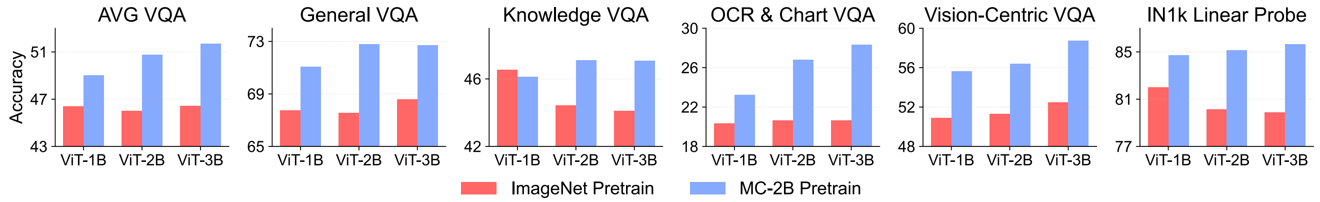

Figure 6. **Comparison of ImageNet-1k and MC-2B Pretraining.** Increasing the diversity and scale of pretraining data improves model performance on VQA accuracy and ImageNet linear probing. Unlike MC-2B pretraining, ImageNet pretraining exhibits no clear trend.

Question 2

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

We pretrain Web-DINO 1B, 2B, and 3B models for 300 epochs on ImageNet-1k, a conventional SSL pretraining dataset, following the recipe from [73]. We compare these variants to those trained on MC-2B. We evaluate their downstream VQA performance and ImageNet-1k linear probing results. As shown in Fig. 6, models pretrained on ImageNet-1k exhibit consistently inferior performance across all metrics. Moreover, unlike models trained on MC-2B, those trained on ImageNet-1k do not improve with increasing model sizes. This highlights the importance of training visual SSL on more diverse and larger datasets. This echoes recent findings that increasing data size and diversity drives LLM scaling [22, 48, 52], and that pretraining data distribution is critical to downstream performance [62].

Question 3

How do scaled models perform on classic vision tasks?

We evaluate Web-DINO models, ranging from 1B to 7B parameters, on classic vision benchmarks including linear probing on ImageNet-1k [24], semantic segmentation on ADE20K [111], and depth estimation on NYUv2 [79]. Following the evaluation protocol of DINOv2 [73], we freeze the vision encoder; see Appendix A for details. As shown in Fig. 7, Web-DINO's performance improves modestly with increasing model size. Web-DINO achieves strong performance across all benchmarks, outperforming MetaCLIP by a significant margin and remaining competitive with off-shelf DINOv2, even outperforming it on ADE20K +ms. Note that the comparison with off-shelf DINOv2 is not exactly apples-to-apples, as we do not use high-resolution adaptation [73], in order to maintain the same input resolution as CLIP. Additionally, the DINOv2 training data has a higher correlation with these classic vision benchmarks, detailed further in Appendix F. These differences suggest that there remains considerable room for further improvement in our model's classic vision performance.

However, we observe that the scaling behavior in classic vision tasks is less pronounced compared to VQA. This finding, along with insights from previous work [31, 70, 91], reinforces the value of VQA as a comprehensive vision model evaluation framework. While classic benchmarks remain important, VQA provides a complementary view into model performance via offering a diverse set of tasks that are grounded in real-world perceptual challenges.

Question 4

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

In Sec. 3, we observed that increasing model size and examples seen leads to unprecedented improvements in OCR & Chart performance for visual SSL models. This is surprising since current off-the-shelf visual SSL methods are notably poor at OCR & Chart understanding compared to language-supervised models [78, 91].

One possible explanation is that web-scale image datasets already contain a degree of textual information.

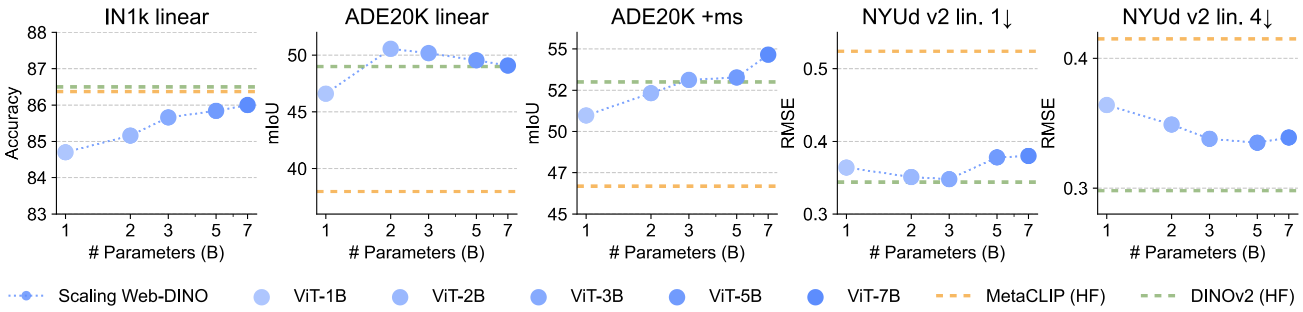

Figure 7. **Performance of Web-DINO models on classic vision tasks.** All models achieve strong performance across ImageNet-1k classification, ADE20K segmentation, and NYU Depth estimation. Web-DINO outperforms MetaCLIP (HF) and is competitive with DINOv2 (HF). (HF) denotes the largest official Hugging Face released version. ↓ means lower is better; by default higher is better.

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

Table 1. **Impact of data filtering on SSL model performance.** We compare Web-DINO ViT-2B models trained on MC-2B with different levels of text filtering (full, 50.3%, and 1.3%) against CLIP ViT-2B trained on full MC-2B. OCR & Chart performance improves with progressively aggressive filtering. Despite receiving zero language supervision, SSL models can surpass CLIP in text-centric tasks.

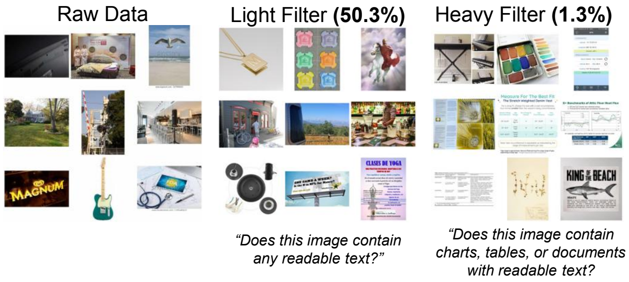
Raw Data
Light Filter (50.3%)
Heavy Filter (1.3%)
MAGNUM
KING & BEACH
"Does this image contain any readable text?"
"Does this image contain charts, tables, or documents with readable text?"

Figure 8. **Examples of filtered MC-2B images.** The Light filter (*Middle*) identifies images containing text, retaining 50.3% of the images. The Heavy filter (*Right*) identifies images explicitly containing charts and documents, retaining only 1.3% of MC-2B.

Unlike object-centric datasets such as ImageNet, images from the web often contain text (e.g. labels, signs, diagrams, etc.). Larger capacity and more data might aid visual SSL models to extract and leverage this textual information.

To test this hypothesis, we apply an off-the-shelf MLLM—SmolVLM2 [3]—to identify images containing text. See Fig. 8 for qualitative examples and Appendix A for details. This results in two curated datasets: (i) Light filter: retains 50.3% of Web-DINO and contains images with any textual content. (ii) Heavy filter: retains 1.3% of MC-2B and contains images with charts, tables, or documents.

We train Web-DINO ViT-2B models on these filtered datasets, with each experiment using 2 billion seen examples (meaning filtered datasets undergo multiple epochs). As shown in Tab. 1, the model trained on lightly filtered data outperforms the full data variant by +6.4% on OCR & Chart, while maintaining strong performance in other categories. The model trained on heavily filtered data performs better and outperforms even the language-supervised CLIP ViT-2B trained on full data by +4.3% on OCR & Chart. Likewise, heavy filtering also improves Average VQA performance, outperforming the full data Web-DINO ViT-2B by +2.6% and even the full data CLIP ViT-2B by +0.7%. This means that is it possible for visual SSL models to outperform CLIP models of the same size, with only a fraction of the total data (in this case 1.3% of MC-2B).

The improvement in OCR & Chart from training on heavily filtered data is particularly pronounced for ChartQA (+24.2%), OCRBench (+13.8%), and DocVQA (+13.0%), while performance remains competitive in all other categories. These results demonstrate that self-supervised visual models, when trained on images containing more text in them, can develop high-quality text understanding capabilities. It suggests that data composition—rather than purely scale or language supervision—is crucial for developing strong OCR & Chart understanding abilities.

Although it is not surprising that skewing the data in favor of OCR & Chart would improve OCR & Chart capabilities, it is surprising that simple data filtering can outperform language supervision on the full data. This simple proof of concept suggests that similar techniques may be used to help visual SSL bridge future gaps in other capabilities.

Question 5

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

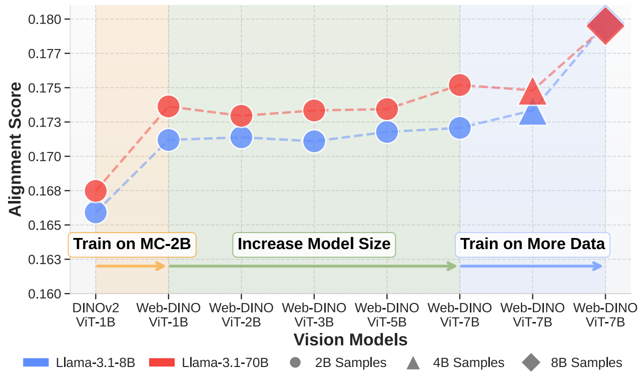

Figure 9. **Alignment score between Web-DINO and LLMs.** Moving from DINOv2 to Web-DINO improves the alignment between the image and the corresponding text representations obtained by LLMs. Increasing model size from 1B to 7B parameters shows gradual improvement, while training on larger data quantities (4B/8B samples) yields the most significant alignment gains.

Thus far, we have seen that visual SSL models can not only become competitive with CLIP models, but also that they can excel at tasks previously thought to require language. This raises an important question: why do vision-only models learn features that work well for multimodal models, even in the absence of language supervision?

We hypothesize that SSL models learn features increasingly aligned with language as model size and examples seen increases. Following Huh et al. [50], we evaluate intrinsic representational alignment by computing a matching metric between the vision encoder and language model, using image-text pairs from the Wikipedia Captions dataset [84]. We use off-the-shelf DINOv2 [73] and Web-DINO as vision encoders, and off-the-shelf Llama-3.1 8B and 70B [94] as the language models, *without* any visual instruction tuning nor alignment procedure.

As shown in Fig. 9, we observe three key trends: (1) training on more diverse data (MC-2B) improves alignment with LLMs (DINOv2 ViT-1B → Web-DINO ViT-1B); (2) increasing the vision model size leads to slightly higher alignment (Web-DINO ViT-1B → ViT-7B); and (3) seeing more training samples further enhances alignment (Web-DINO ViT-7B trained on 2B samples → 8B samples).

These findings suggest that as model size and, in particular, training samples scale, vision models naturally develop text-sensitive features and achieve strong alignment with LLMs, without explicit language supervision.

5. The Web-SSL Model Family

Next, we analyze the overall best performing vision encoders using both VQA and classic vision benchmarks. In Table 2, we show the best results of our vision encoders against recent off-the-shelf vision encoders, in terms of VQA and classic vision tasks.

For VQA, all vision encoders—including off-the-shelf models—are evaluated using the same visual instruction tuning setup detailed in Sec. 2.3, and mainly 224×224 input resolution for the purpose of fair comparison. Because the goal is not to produce a state-of-the-art MLLM, we did not employ techniques such as unfreezing the vision encoder, resolution tiling [59], and spatial visual aggregator [91].

For classic vision, we follow the evaluation procedure from Oquab et al. [73] and evaluate linear probe performance on ImageNet-1k [24], ADE20K [111], and NYU Depth v2 [79]. The input resolution differs between classic vision tasks, but each model tested uses the same exact settings from Oquab et al. [73]. We emphasize that the primary motivation is still to provide controlled insights, even though other off-shelf models are trained on different data.

Performance at 224px. Web-DINO can outperform off-the-shelf MetaCLIP in both VQA and classic vision tasks. Web-DINO is even able to match the performance of SigLIP and SigLIP2 on VQA despite seeing 5× less data and receiving no language supervision. In general, Web-DINO outperforms all off-shelf language-supervised CLIP models at traditional vision benchmarks. Although our best Web-DINO model is 7B parameters, the results from Sec. 3.1 and Sec. 3.2 suggest that CLIP models saturate beyond moderate model and data sizes, while visual SSL improves progressively with increasing model and data size. Web-DINO also outperforms off-the-shelf visual SSL methods, including DINOv2 [73], in all VQA categories. Web-DINO is also competitive in traditional vision benchmarks.

Performance beyond 224px. Next, we discuss the performance of higher resolution models. Following Oquab et al. [73], we additionally fine-tune Web-DINO for 20k steps. We do this for resolutions of 378 and 518, to compare against the higher-resolution off-shelf versions of SigLIP as well as DINOv2. See Appendix D for training details. From 224 to 378 to 518 resolution, Web-DINO improves steadily at average VQA, with notable gains in OCR & Chart performance. Classic vision performance improves modestly with higher resolution. At 384 resolution, Web-DINO trails behind SigLIP. At 518 resolution, Web-DINO is largely able to bridge the gap. The results suggest that Web-DINO may benefit from further increasing high-resolution adaptation.

6. Related Work

Visual self-supervised learning methods. Early visual SSL methods explored various pretext tasks for pretraining [6, 25, 37, 71, 74, 97, 109]. More recently, research has converged on two primary approaches: joint embedding methods and masked image modeling. Joint embedding methods learn invariant features by aligning representations of different augmented views [12, 14–17, 19, 34, 40, 43, 55, 69], while masked modeling [4, 5, 7, 13, 28, 44, 99, 100, 112] learns by predicting masked visual inputs. Our

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

Table 2. **Comparison with other vision models.** Web-DINO ViT-7B achieves competitive performance with CLIP models on VQA without language supervision and surpasses them on traditional vision tasks. Compared to other self-supervised models like DINOv2, Web-DINO significantly narrows the performance gap with CLIP on VQA tasks, particularly excelling in OCR & Chart understanding. These results demonstrate that SSL can effectively produce strong visual representations for both multimodal and classic vision tasks.

work complements SSL research focused on pretraining algorithms, by taking off-the-shelf training code and training visual SSL at scale with a controlled experimental setup.

Data used to train vision models. Both supervised [26, 41, 63, 102] and SSL vision models have traditionally relied on standard datasets such as MNIST [54], CIFAR-10 [53], and ImageNet [24, 76]. More recently, self-supervised methods have scaled to larger unlabeled datasets, such as YFCC [90], LVD-142M [73], and IG-3B [81]; however, these methods still exhibit a significant performance gap compared to language-supervised models on VQA.

In contrast, language-supervised models [20, 31, 75, 85, 86, 88, 104, 108] leverage significantly larger image-text datasets, from WIT-400M [75] to billion-scale web data [30, 33, 77, 104], with some using up to 100B image-text pairs [98], and even synthetic captions [66]. Studies suggest that pretraining data distribution is more critical for downstream performance than specific training methodologies [29, 62].

Our work bridges these paradigms by pretraining SSL models on web-scale data. Through controlled experiments (Sec. 3 and 4), we show that (1) visual SSL models are sensitive to the training distribution, and (2) increasing data diversity and quantity significantly improves performance on a diverse range of VQA tasks.

Evaluating vision models. Classic works have primarily used image classification [10, 24, 45, 46, 53, 54] to evaluate learned representations. More recent SSL research has expanded evaluation to include image segmentation [23, 27, 42, 111], depth estimation [36, 79, 82], and video classification [8, 38, 83]. Language-supervised models [75, 108], due to their two-tower encoder structure, commonly use zero-shot image classification to assess the quality of learned image and text features.

Our work follows recent proposals [31, 70, 91] to evaluate vision encoders on a broader range of VQA tasks [32, 39, 61, 89, 101, 105, 106] using MLLMs. These VQA tasks complement traditional vision benchmarks by assessing visual features on a more diverse range of real-world perceptual challenges. As shown in Sec. 3 and Sec. 4, we find that visual SSL trained on web-scale data learns representations that continue to improve on VQA benchmarks, and—to a lesser degree—also on traditional vision benchmarks.

7. Discussion

We show that large-scale visual encoders that are trained with self-supervised language-free objectives can produce high quality visual features for multimodal models. Our results echo the "bitter lesson" [87] and suggest that imposing less supervision—including language—remains a promising direction for advancing the field of computer vision. We hope our work will inspire further exploration of vision-only approaches, which will enable the construction of next generation vision models that excel at both traditional vision and modern multimodal capabilities.

8. Acknowledgements

We thank Ellis Brown, John Nguyen, Junlin Han, Shengyi Qian, Tyler Zhu, Yuexiang Zhai, Druv Pai, Shusheng Yang, Jihan Yang, Muzi Tao, Boyang Zheng, and Anjali Gupta for reviewing this manuscript. We thank Hu Xu and the MetaCLIP paper authors for creating the MetaCLIP dataset. We thank Mido Assran, Mikael Henaff, Daniel Bolya, Hu Xu, Mark Ibrahim, Russ Howes, and Matthew Muckley for their insightful feedback. We thank Michaël Ramamonjisoa and Marc Szafraniec for their help with image segmentation and depth estimation evaluations. Lastly, we thank Ananya Saxena, Cody Olsen, Mack Ward, Maxwell Taylor, Kalyan Saladi, Dev Satpathy, Dinesh Kannappan, Xiaodong Ma, Jacob Kahn, Gabriel Synnaeve, and Shubho Sengupta for infrastructure support. ST is supported by the Meta AI Mentorship (AIM) Program. SX also acknowledges support from the OpenPath AI Foundation, IITP grant funded by the Korean Government (MSIT) (No. RS-2024-00457882, National AI Research Lab Project), and NSF Award IIS-2443404.

References

[1] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Devendra Chaplot, Jessica Chudnovsky, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024. 1

[2] AI@Meta. Llama 3 model card. 2024. 1, 14, 16

[3] Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, et al. Smollm2: When smol goes big–data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025. 6, 14

[4] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, 2023. 7

[5] Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan L Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. In *CVPR*, 2024. 7

[6] Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, et al. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*, 2023. 7

[7] Amir Bar, Florian Bordes, Assaf Shocher, Mido Assran, Pascal Vincent, Nicolas Ballas, Trevor Darrell, Amir Globerson, and Yann LeCun. Stochastic positional embeddings improve masked image modeling. In *ICML*, 2024. 7

[8] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *NeurIPS*, 2021. 8

[9] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024. 1

[10] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, 2014. 8

[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020. 2

[12] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 7

[13] João Carreira, Dilara Gokay, Michael King, Chuhan Zhang, Ignacio Rocco, Aravindh Mahendran, Thomas Albert Keck, Joseph Heyward, Skanda Koppula, Etienne Pot, et al. Scaling 4d representations. *arXiv preprint arXiv:2412.15212*, 2024. 7

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 1, 7

[15] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021.

[16] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[17] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021. 7

[18] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. In *ICLR*, 2023. 1

[19] Yubei Chen, Adrien Bardes, Zengyi Li, and Yann LeCun. Bag of image patch embedding behind the success of self-supervised learning. *arXiv preprint arXiv:2206.08954*, 2022. 7

[20] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024. 8

[21] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, 2023. 14

[22] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul

Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. arxiv 2022. *arXiv preprint arXiv:2204.02311*, 10:1, 2022. 5

[23] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 8

[24] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1, 2, 5, 7, 8, 14, 20

[25] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015. 7

[26] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 8

[27] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 8

[28] David Fan, Jue Wang, Shuai Liao, Yi Zhu, Vimal Bhat, Hector Santos-Villalobos, Rohith MV, and Xinyu Li. Motion-guided masking for spatiotemporal representation learning. In *ICCV*, 2023. 7

[29] Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data determines distributional robustness in contrastive language image pre-training (clip). In *ICML*, 2022. 8

[30] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. In *ICLR*, 2024. 8

[31] Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor Guilherme Turrisi da Costa, Louis Béthune, Zhe Gan, et al. Multimodal autoregressive pre-training of large vision encoders. *arXiv preprint arXiv:2411.14402*, 2024. 2, 5, 8, 15

[32] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: a comprehensive evaluation benchmark for multimodal large language models. corr abs/2306.13394 (2023), 2023. 8, 20

[33] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. In *NeurIPS*, 2024. 8

[34] Quentin Garrido, Yubei Chen, Adrien Bardes, Laurent Najman, and Yann Lecun. On the duality between contrastive and non-contrastive self-supervised learning. In *ICLR*, 2023. 7

[35] Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. Planting a seed of vision in large language model. *arXiv preprint arXiv:2307.08041*, 2023. 20

[36] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 8

[37] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018. 7

[38] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017. 8

[39] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. 8

[40] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *NeurIPS*, 2020. 7

[41] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 8

[42] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 8

[43] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. arxiv e-prints, art. In *CVPR*, 2019. 7

[44] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 1, 3, 4, 7

[45] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Xiaodong Song. Natural adversarial examples. 2021 ieee. In *CVPR*, 2019. 8

[46] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. 2021 ieee. In *ICCV*, 2020. 8

[47] Tuomo Hiippala, Malihe Alikhani, Jonas Haverinen, Timo Kalliokoski, Evanfiya Logacheva, Serafina Orekhova, Aino Tuomainen, Matthew Stone, and John A Bateman. Ai2d-rst: A multimodal corpus of 1000 primary school science diagrams. *Language Resources and Evaluation*, 55:661–688, 2021. 20

[48] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. In *NeurIPS*, 2023. 5

[49] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 20

[50] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. In *ICML*, 2024. 7

[51] Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, et al. Dinov2 meets text: A unified framework for image-and pixel-level vision-language alignment. *arXiv preprint arXiv:2412.16334*, 2024. 16

[52] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 2, 5

[53] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 8

[54] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998. 8

[55] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1): 1–62, 2022. 1, 7

[56] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1

[57] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[58] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 1

[59] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 7

[60] Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023. 20

[61] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multimodal model an all-around player? In *ECCV*, 2024. 8, 20

[62] Zhuang Liu and Kaiming He. A decade's battle on dataset bias: Are we there yet? In *ICLR*, 2025. 5, 8

[63] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022. 8

[64] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, 2022. 20

[65] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2023. 20

[66] Kevis-Kokitsi Maninis, Kaifeng Chen, Soham Ghosh, Arjun Karpur, Koert Chen, Ye Xia, Bingyi Cao, Daniel Salz, Guangxing Han, Jan Dlabal, et al. Tips: Text-image pretraining with spatial awareness. *ICLR*, 2025. 8

[67] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *ACL*, 2022. 20

[68] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, 2021. 20

[69] Ishan Misra and Laurens Van Der Maaten. Self-supervised learning of pretext-invariant representations. In *CVPR*, 2020. 7

[70] Muhammad Ferjad Naeem, Yongqin Xian, Xiaohua Zhai, Lukas Hoyer, Luc Van Gool, and Federico Tombari. Silc: Improving vision language pretraining with self-distillation. In *ECCV*, 2024. 5, 8

[71] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016. 7

[72] OpenAI. Chatgpt, 2022. 2

[73] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. In *TMLR*, 2023. 1, 2, 3, 5, 7, 8, 14, 15, 21

[74] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 7

[75] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 8

[76] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 1, 8

[77] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022. 1, 8, 14, 15

[78] Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, et al. Eagle: Exploring the design space for multimodal llms with mixture of encoders. *arXiv preprint arXiv:2408.15998*, 2024. 1, 5

[79] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 5, 7, 8, 14, 20

[80] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, 2019. 20

[81] Mannat Singh, Quentin Duval, Kalyan Vasudev Alwala, Haoqi Fan, Vaibhav Aggarwal, Aaron Adcock, Armand Joulin, Piotr Dollár, Christoph Feichtenhofer, Ross Girshick, et al. The effectiveness of mae pre-pretraining for billion-scale pretraining. In *ICCV*, 2023. 8

[82] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, 2015. 8

[83] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 8

[84] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2443–2449, 2021. 7

[85] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 1, 8

[86] Quan Sun, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. Eva-clip-18b: Scaling clip to 18 billion parameters. *arXiv preprint arXiv:2402.04252*, 2024. 8

[87] Richard Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 2019. 8

[88] Zineng Tang, Long Lian, Seun Eisape, XuDong Wang, Roei Herzig, Adam Yala, Alane Suhr, Trevor Darrell, and David M. Chan. Tulip: Towards unified language-image pretraining, 2025. Preprint. 8

[89] Muzi Tao and Saining Xie. What does a visual formal analysis of the world's 500 most famous paintings tell us about multimodal LLMs? In *The Second Tiny Papers Track at ICLR 2024*, 2024. 8

[90] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 8

[91] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *NeurIPS*, 2024. 1, 2, 3, 5, 7, 8, 14, 20

[92] Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164*, 2024. 14

[93] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, 2024. 20

[94] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3, 7

[95] Michael Tschannen, Manoj Kumar, Andreas Steiner, Xiaohua Zhai, Neil Houlsby, and Lucas Beyer. Image captioners are scalable vision learners too. In *NeurIPS*, 2024. 2, 3

[96] Bo Wan, Michael Tschannen, Yongqin Xian, Filip Pavetic, Ibrahim Alabdulmohsin, Xiao Wang, André Susano Pinto, Andreas Steiner, Lucas Beyer, and Xiaohua Zhai. Locca: Visual pretraining with location-aware captioners. *arXiv preprint arXiv:2403.19596*, 2024. 2

[97] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *ICCV*, 2015. 7

[98] Xiao Wang, Ibrahim Alabdulmohsin, Daniel Salz, Zhe Li, Keran Rong, and Xiaohua Zhai. Scaling pre-training to one hundred billion data for vision language models. *arXiv preprint arXiv:2502.07617*, 2025. 8

[99] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, 2022. 7

[100] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *CVPR*, 2023. 7

[101] xAI. grok, 2024. 8, 20

[102] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *arXiv preprint arXiv:1611.05431*, 2016. 8

[103] Hu Xu, Po-Yao Huang, Xiaoqing Ellen Tan, Ching-Feng Yeh, Jacob Kahn, Christine Jou, Gargi Ghosh, Omer Levy, Luke Zettlemoyer, Wen-tau Yih, et al. Altogether: Image captioning via re-aligning alt-text. *arXiv preprint arXiv:2410.17251*, 2024. 2

[104] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. In *ICLR*, 2024. 1, 2, 8, 15, 21

[105] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024. 8, 20

[106] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Ming Yin, Botao Yu, Ge Zhang, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024. 8

[107] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18123–18133, 2022. 16

[108] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 1, 8

[109] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016. 1, 7

[110] Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023. 14

[111] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 2019. 2, 5, 7, 8, 14, 20

[112] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 7