A. Implementation Details

Training. For training Web-DINO, Web-MAE, and CLIP models, we closely follow the existing open-source codebases: the official DINOv2 and MAE repositories, and the MetaCLIP codebase which builds on top of the OpenCLIP codebase [21]. We use Fully Sharded Data Parallel (FSDP) [110] for distributed training of larger models.

For Web-DINO and CLIP pretraining, we follow the exact recipe and hyperparameters from the original paper for their largest model. For MAE pretraining, we observe that training becomes more prone to divergence as model size increases. To mitigate this, we reduce the learning rate from 2.4e-3 to 1.6e-3 and extend the warmup period to 80K iterations. Table 3 provides a summary of the pretraining hyperparameters for training on 2B samples.

Model	Batch Size	Learning Rate	Warmup
Web-DINO	3072	3.5e-4	100K
Web-MAE	4096	1.6e-3	80K
CLIP	32768	4e-4	2K

Table 3. **Hyperparameters for Web-DINO, Web-MAE and CLIP.**

VQA evaluation. For VQA evaluation, we follow Tong et al. [91, 92] and use Cambrian-Alignment data for MLP projector training and Cambrian-7M for MLP and LLM fine-tuning. We finetune on top of Llama-3 8B Instruct [2]. The vision encoder is frozen throughout finetuning. We excluded LAION [77] images from the Cambrian data to comply with safety standards. We first encode the images at the model's original input resolution using the pretrained vision encoder. Next, we extract features from the final encoder layer. Following prior approaches [91, 92], we then resize the resulting token sequence to a fixed length of 576 tokens through bilinear interpolation. This ensures consistency across evaluations despite variations in input image resolutions. We report configurations in Tab. 4.

Classic vision evaluation. We follow the evaluation procedure in DINOv2 [73] for all classic vision evaluation: linear probe on ImageNet1k [24], ADE20K [111], and NYU Depth v2 [79]. For ImageNet-1k, we evaluate models with their pretrained image resolution; For ADE20K and NYU Depth v2, we use the settings from Oquab et al. [73]. For ADE20K, we follow DINOv2 and report the **linear** and **+ms** setting. For NYU Depth v2, we report **lin. 1** and **lin. 4**. See the original paper for additional details.

Model architectures. To recap, we first borrowed the ViT-g architecture from Oquab et al. [73] and named it ViT-1B for consistent notation. We then define 2B, 3B, 5B, and 7B architectures inspired by language model scaling. Specifically, the 2 - 7B architectures are wider than the 1B variant, inspired by language model recipes. Our 7B architecture is almost identical to the Llama-2 7B design, except for the patch embedding layer which is unique to ViTs. See Table 5.

Text filtering. In Question 4, we introduced the "Light" and "Heavy" filters which retain 50.3% and 1.3% of MC-2B respectively. Specifically, we use a small MLLM, SmolVLM2 [3], to identify images containing text, using prompts such as *"Does this image contain any readable text?"*. The intention is not to achieve perfect filtering, but rather to skew the data distribution in the general desired direction. See Fig. 8 for a visualization of the filtering process and some examples. This results in two curated datasets:

(i) Light filter: Retains 50.3% of the original data, primarily consisting of images with some textual content. Prompt used: *"Does this image contain any readable text? Answer only yes or no."*

(ii) Heavy filter: Retains only 1.3% of the data, focusing mainly on charts and documents. Prompt used: *"Please think carefully before answering. Does this image contain charts, tables, or documents with readable text? Answer only yes or no."*

B. Full Results

We include full results of all experiments presented in Sec. 3 and Sec. 4.

B.1. Web-DINO

Scaling up model sizes. We show quantitative results of scaling up the model under VQA evaluation in Tab. 6 and classic vision evaluation in Tab. 7. These are the numerical results for Sec. 3.1.

Scaling up data sizes. We show quantitative results of scaling up the number of data seen with Web-DINO ViT-7B on VQA evaluation in Tab. 8 and classic vision evaluation in Tab. 9. These are the numerical results for Sec. 3.2.

Scaling down training data. We show VQA evaluation results from training Web-DINO on less diverse data–ImageNet-1k, in Tab. 10. These are the full results for scaling down training data experiments in Question 2.

B.2. Web-MAE

We show VQA evaluation results from scaling up MAE trained on MC-2B, in Tab. 11. These are the full results for Question 1.

Backbone	Data		Adapter			Instruction Tuning		
LLM	Adapter	Instruction Tuning	LR	WD	BS	LR	WD	BS
Llama-3 8B Instruct	Cambrian Adapter Data	Cambrian-7M	1.00e-5	0.0	512	4.00e-5	0	512

Table 4. **Hyperparameters for all VQA experiments.** We exclude LAION [77] from Cambrian data.

Model	Width	Depth	Heads	MLP
ViT-1B	1536	40	24	6144
ViT-2B	2688	24	21	10752
ViT-3B	3072	26	24	12288
ViT-5B	3584	32	28	14336
ViT-7B	4096	32	32	16384

Table 5. **Model architecture details.** For consistency, we denote ViT-g from Oquab et al. [73] as ViT-1B.

B.3. Scaled CLIP Models

We show VQA evaluation results from scaling up MetaCLIP [104] trained on MC-2B, in Tab. 12. These are the full results for Sec. 3.1. In contrast to visual SSL methods in Tab. 7 and Tab. 11, CLIP models do not exhibit clear scaling behavior.

B.4. Text Filtered Models

We provide full results for Question 4. As shown in Tab. 13, SSL models learn features particularly well-suited for OCR & Chart tasks when trained on datasets with a higher concentration of text-rich images. This suggests that visual SSL is sensitive to the underlying training distribution and can be effectively steered toward specific downstream applications, such as OCR & Chart.

B.5. Baseline Models

In Tab. 14, we provide full VQA results for the reference off-shelf models that we evaluated in Sec. 5.

C. Additional Results

We include extra results that were not presented in the main text.

C.1. Classic Vision Performance of Scaling CLIP

In Tab. 15, we provide full classic vision results for CLIP trained on MC-2B with 2 billion images seen, for models ranging from 1B to 7B parameters. In Fig. 10, we visualize the classic vision performance of CLIP compared to Web-DINO trained on the same data, and see that CLIP consistently underperforms Web-DINO on classic vision evals across all model sizes.

C.2. Scaling CLIP Text Encoder

Throughout the paper, we used the standard 300M parameter text encoder from OpenCLIP (24 layers, 16 heads per layer, 1024 dimension). To test whether the capacity of the text encoder is a bottleneck for training larger CLIP vision encoders, we scale the text encoder to 600M and 1B parameters. The results in Tab. 16 suggest that increasing the capacity of CLIP's text encoder does not improve performance.

C.3. Comparison to Non-CLIP Language-Supervised Models

AIMv2 [31] is also a recent language-supervised vision encoder that leverages a text decoding and masked image modeling objective, unlike CLIP-family models which are contrastively pretrained. As AIMv2 is trained on different data, the comparison to Web-SSL is not quite apples-to-apples — similar to the other comparisons in Tab. 2 — but offers an additional perspective into Web-SSL's performance compared to other model families.

In Tab. 17, we compare AIMv2 to Web-SSL and observe that while Web-SSL achieves slightly lower performance on VQA, it is still competitive with AIMv2 despite not receiving any text supervision and using different data. Web-SSL does better on image classification and semantic segmentation, while doing significantly worse on depth estimation.

D. High Resolution Adaption of Web-SSL

Following Oquab et al. [73], we further fine-tune our model under higher resolution settings of 378×378 and 518×518 for 20k iterations. We use a batch size of 2048 and a correspondingly lower learning rate of 1.41e-5. All other parameters remain exactly the same as previously specified, including the learning rate warmup ratio, given the total of 10k iterations.

We also provided detailed benchmark results of high-resolution adaptation of Web-DINO in Tab. 18.

E. Evaluation

Tab. 19 lists evaluation benchmarks used and their purposes.

F. Pretraining Dataset Cards

For reference, in Tab. 20 we include the data composition of LVD-142M, which was used to train the off-shelf DINOv2 model [73]. LVD-142M is a carefully curated data mix closely aligned with downstream classic vision evaluation tasks. In comparison, we leverage MetaCLIP data, which is less curated and collected from 15 snapshots of CommonCrawl (CC).

Vision Backbone		General				Knowledge				OCR & Chart				Vision-Centric			
Model	Average	MME^P	MMB	$SEED^I$	GQA	SQA^I	$MMMU^V$	$MathVista^M$	AI2D	ChartQA	OCRBench	TextVQA	DocVQA	MMVP	RealWorldQA	CV-$Bench^{2D}$	CV-$Bench^{3D}$
Web-DINO ViT-1B	49.01	1731.52	65.37	69.92	62.40	72.58	35.33	12.30	64.28	19.20	9.40	47.41	17.00	37.33	57.12	64.80	63.16
Web-DINO ViT-2B	50.77	1760.80	68.98	71.29	62.89	73.67	31.77	15.90	67.06	23.30	15.60	49.20	19.00	38.00	57.38	65.85	64.41
Web-DINO ViT-3B	51.71	1757.27	68.04	71.84	63.19	73.57	33.00	14.40	67.32	25.68	17.10	50.45	20.00	42.66	56.86	69.49	65.83
Web-DINO ViT-5B	52.83	1840.81	70.01	72.39	63.56	75.06	32.11	12.40	67.77	26.96	22.10	50.64	21.00	44.66	57.64	67.75	69.16
Web-DINO ViT-7B	53.87	1823.76	68.98	73.02	64.22	74.61	35.11	14.00	69.43	28.80	23.59	51.10	22.00	48.00	59.34	69.96	68.58

Table 6. **VQA Evaluation: Web-DINO trained on MC-2B with 2 billion images seen.**

Vision Backbone	IN1k lin.	ADE20K lin.	ADE20K +ms.	NYUd lin. 1 (↓)	NYUd lin. 4 (↓)
Web-DINO ViT-1B	84.70	46.60	50.97	0.364	0.345
Web-DINO ViT-2B	85.16	50.55	52.32	0.351	0.335
Web-DINO ViT-3B	85.66	50.17	53.12	0.348	0.328
Web-DINO ViT-5B	85.84	49.54	53.27	0.378	0.335
Web-DINO ViT-7B	86.00	49.08	54.65	0.380	0.339

Table 7. **Classic Vision Evaluation: Web-DINO trained on MC-2B with 2 billion images seen.**

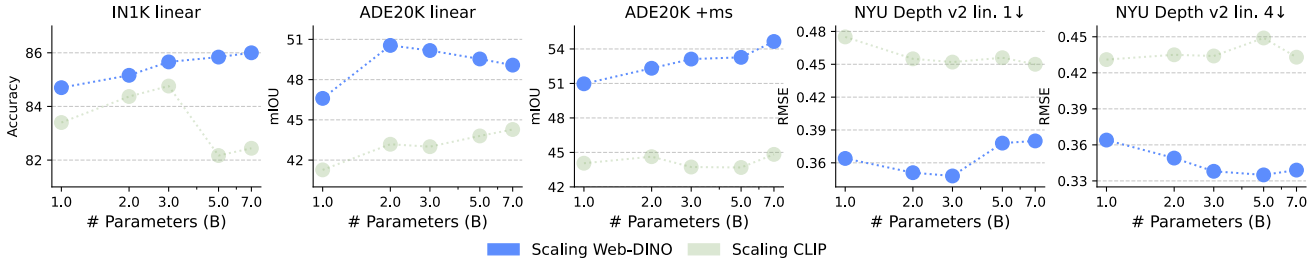

Figure 10. Classic vision evals for CLIP vs. DINO trained on 2 billion images seen from MC-2B. Lower is better for NYU Depth.

G. Limitations

In this work, we focus on training visual SSL models without using language. The main limitation of vision-only models, compared to language-supervised models, is that they do not support zero-shot image classification out of the box. However, by integrating visual SSL models into MLLM frameworks through instruction tuning, we show they can achieve impressive downstream performance across classification and other tasks. Another way to achieve zero-shot image classification is to use LiT-style adaptation [51, 107], but this is out-of-scope for our work as we do not use language supervision. To focus on comparing the vision encoder, we fixed the base LLM for visual instruction tuning to Llama-3 8B Instruct [2]. We hypothesize that the findings using other LLM backbones would be similar, however this is not in scope for our work. Additionally, while we demonstrate that visual SSL scales well on MetaCLIP data, we leave the exploration of even larger and/or uncurated datasets to future work.

Vision Backbone		General				Knowledge				OCR & Chart				Vision-Centric			
Model	Average	MME^P	MMB	$SEED^I$	GQA	SQA^I	$MMMU^V$	$MathVista^M$	AI2D	ChartQA	OCRBench	TextVQA	DocVQA	MMVP	RealWorldQA	$CV\text{-}Bench^{2D}$	$CV\text{-}Bench^{3D}$
Web-DINO ViT-7B (1B Data)	51.02	1785.97	68.12	72.54	63.60	73.87	32.88	12.70	66.58	23.60	15.20	49.04	19.00	43.33	57.12	68.35	61.08
Web-DINO ViT-7B (2B Data)	53.87	1823.76	68.98	73.02	64.22	74.61	35.11	14.00	69.43	28.80	23.59	51.10	22.00	48.00	59.34	69.96	68.58
Web-DINO ViT-7B (4B Data)	54.37	1827.12	71.39	72.61	63.53	72.73	34.00	18.90	67.09	35.12	30.00	53.19	24.00	45.33	55.94	69.68	65.00
Web-DINO ViT-7B (8B Data)	55.24	1811.05	71.30	72.14	64.04	72.43	35.66	15.20	68.52	35.52	36.40	56.53	29.00	46.00	57.90	70.53	62.08

Table 8. **VQA Evaluation: Web-DINO ViT-7B trained on MC-2B with increased number of images seen.**

Vision Backbone	IN1k lin.	ADE20K lin.	ADE20K +ms.	NYUd lin. 1 (↓)	NYUd lin. 4 (↓)
Web-DINO ViT-7B (2B Data)	86.00	49.08	54.65	0.380	0.339
Web-DINO ViT-7B (4B Data)	86.33	47.41	54.66	0.416	0.363
Web-DINO ViT-7B (8B Data)	86.52	42.14	52.55	0.491	0.376

Table 9. **Classic Vision Evaluation: Web-DINO ViT-7B trained on MC-2B with increased number of images seen.**

Vision Backbone		General				Knowledge				OCR & Chart				Vision-Centric			
Model	Average	MME^P	MMB	$SEED^I$	GQA	SQA^I	$MMMU^V$	$MathVista^M$	AI2D	ChartQA	OCRBench	TextVQA	DocVQA	MMVP	RealWorldQA	$CV\text{-}Bench^{2D}$	$CV\text{-}Bench^{3D}$
Web-DINO ViT-1B	46.39	1704.30	59.27	66.43	60.12	71.29	32.77	18.70	63.40	17.56	4.90	44.93	14.00	32.00	52.41	62.81	56.41
Web-DINO ViT-2B	45.99	1666.01	60.13	66.64	60.19	68.71	34.88	12.10	62.07	18.60	4.39	45.55	14.00	32.66	52.67	62.07	57.83
Web-DINO ViT-3B	46.43	1729.40	60.56	66.99	60.24	70.50	31.88	11.70	62.30	17.52	4.80	45.18	15.00	31.33	53.20	62.77	62.50
Web-DINO ViT-5B	46.28	1661.25	59.27	67.24	61.10	69.41	31.55	10.90	61.46	18.72	4.60	45.53	15.00	34.00	53.07	64.57	61.08

Table 10. **VQA Evaluation: Web-DINO trained on ImageNet-1k.**

Vision Backbone		General				Knowledge				OCR & Chart				Vision-Centric			
Model	Average	MME^P	MMB	$SEED^I$	GQA	SQA^I	$MMMU^V$	$MathVista^M$	AI2D	ChartQA	OCRBench	TextVQA	DocVQA	MMVP	RealWorldQA	$CV\text{-}Bench^{2D}$	$CV\text{-}Bench^{3D}$
Web-MAE ViT-1B	49.19	1736.22	62.02	68.38	60.05	73.27	33.11	12.90	63.92	23.60	16.40	47.84	18.00	36.66	52.81	70.42	60.83
Web-MAE ViT-2B	50.59	1700.16	63.57	69.21	60.93	72.48	32.22	15.50	64.44	29.00	23.20	48.78	20.00	38.00	55.16	67.98	63.91
Web-MAE ViT-3B	50.92	1723.85	64.69	69.71	60.94	72.13	34.33	13.50	65.70	30.92	24.60	48.92	20.00	37.33	54.64	64.15	66.91
Web-MAE ViT-5B	51.50	1710.13	65.12	70.13	61.10	72.63	32.66	13.90	65.67	33.80	26.50	49.60	21.00	38.00	53.72	66.69	67.91

Table 11. **VQA Evaluation: Web-MAE trained on MC-2B.**

Vision Backbone		General				Knowledge				OCR & Chart				Vision-Centric			
Model	Average	MME^P	MMB	$SEED^I$	GQA	SQA^I	$MMMU^V$	$MathVista^M$	AI2D	ChartQA	OCRBench	TextVQA	DocVQA	MMVP	RealWorldQA	$CV\text{-}Bench^{2D}$	$CV\text{-}Bench^{3D}$
MetaCLIP ViT-1B	52.30	1813.70	68.90	69.45	60.35	74.07	33.55	12.70	64.41	33.20	34.59	52.15	26.00	37.33	52.15	65.47	61.83
MetaCLIP ViT-2B	53.03	1787.39	68.81	69.54	61.08	75.16	34.66	20.10	65.38	32.80	32.90	52.55	26.00	37.33	52.94	65.19	64.67
MetaCLIP ViT-3B	53.22	1873.67	68.72	70.33	61.85	77.29	32.77	11.80	66.35	32.16	34.40	54.58	26.00	35.33	55.55	65.57	65.08
MetaCLIP ViT-5B	52.52	1779.03	70.10	70.26	61.53	72.43	33.44	17.90	66.74	30.04	32.20	52.49	25.00	39.33	54.50	64.22	61.16
MetaCLIP ViT-7B	52.97	1827.80	69.93	69.47	61.33	74.91	35.55	16.80	65.15	32.12	32.10	52.07	25.00	39.33	54.11	65.08	63.16

Table 12. **VQA Evaluation: MetaCLIP trained on MC-2B with 2 billion images seen.**

Vision Backbone		General				Knowledge				OCR & Chart				Vision-Centric			
Model	Average	MME^{P}	MMB	$SEED^{I}$	GQA	SQA^{I}	$MMMU^{V}$	$MathVista^{M}$	AI2D	ChartQA	OCRBench	TextVQA	DocVQA	MMVP	RealWorldQA	$CV\text{-}Bench^{2D}$	$CV\text{-}Bench^{3D}$
Web-DINO ViT-1B (No Filter)	49.01	1731.52	65.37	69.92	62.40	72.58	35.33	12.30	64.28	19.20	9.40	47.41	17.00	37.33	57.12	64.80	63.16
Web-DINO ViT-1B (Light Filter)	50.73	1690.89	65.54	70.68	62.63	70.99	33.89	17.80	63.69	26.12	21.80	50.56	20.00	36.00	56.86	64.84	65.75
Web-DINO ViT-1B (Heavy Filter)	49.44	1593.79	61.40	65.34	59.53	71.19	31.33	14.90	64.83	36.92	24.09	50.09	27.00	21.33	53.20	66.53	63.66
Web-DINO ViT-2B (No Filter)	50.77	1760.80	68.98	71.29	62.89	73.67	31.77	15.90	67.06	23.30	15.60	49.20	19.00	38.00	57.38	65.85	64.41
Web-DINO ViT-2B (Light Filter)	53.38	1768.67	68.38	71.80	63.24	74.16	33.88	31.40	67.38	31.40	27.30	51.26	23.00	39.33	56.47	61.13	65.50
Web-DINO ViT-2B (Heavy Filter)	53.65	1743.56	65.29	69.28	61.19	74.86	32.22	14.50	67.42	47.48	29.40	52.80	32.00	40.00	54.50	65.85	64.50

Table 13. **VQA Evaluation: Web-DINO trained on text filtered MC-2B.**

Vision Backbone		General				Knowledge				OCR & Chart				Vision-Centric			
Model	Average	MME^{P}	MMB	$SEED^{I}$	GQA	SQA^{I}	$MMMU^{V}$	$MathVista^{M}$	AI2D	ChartQA	OCRBench	TextVQA	DocVQA	MMVP	RealWorldQA	$CV\text{-}Bench^{2D}$	$CV\text{-}Bench^{3D}$
CLIP Models																	
MetaCLIP ViT-H_{224px}	54.91	1860.58	72.93	70.96	62.22	77.88	36.88	15.00	67.32	35.60	33.40	55.10	29.00	41.33	53.46	68.53	65.91
SigLIP ViT-$SO400M_{224px}$	55.36	1807.30	72.76	71.83	62.68	76.74	35.44	14.00	68.65	33.08	40.20	56.61	28.00	47.33	56.99	66.42	64.66
SigLIP ViT-$SO400M_{384px}$	59.97	1892.16	73.71	73.00	63.80	77.83	33.88	20.00	69.78	54.24	46.40	63.53	50.00	46.00	58.43	67.37	66.91
SigLIP2 ViT-$SO400M_{224px}$	56.32	1789.26	73.36	72.20	62.60	74.96	35.55	22.40	69.85	35.76	42.00	59.68	31.00	44.00	54.24	69.88	64.16
SigLIP2 ViT-$SO400M_{384px}$	61.98	1895.70	74.57	72.24	64.81	79.27	36.33	19.90	72.24	59.68	52.90	67.15	54.00	49.33	54.77	70.73	69.00
SSL Models																	
DINOv2 ViT-g_{224px}	49.25	1785.25	64.86	70.89	62.89	72.03	32.11	12.40	62.37	17.96	5.50	47.06	15.00	47.33	56.33	65.92	66.08
DINOv2 ViT-g_{378px}	47.94	1734.38	64.26	71.50	62.21	71.04	33.11	9.60	63.08	17.76	5.00	45.59	15.00	41.33	56.47	63.79	60.58
DINOv2 ViT-g_{518px}	47.91	1694.08	62.45	70.64	62.87	71.29	33.55	11.80	63.37	18.32	5.10	46.27	15.00	37.33	56.60	65.36	61.83
I-JEPA ViT-H $_{224px}$	44.78	1598.15	60.01	64.04	57.66	68.91	34.55	10.20	62.07	16.72	4.00	42.99	14.00	29.33	49.93	57.39	57.16
MAE ViT-H_{224px}	45.21	1697.06	56.87	56.41	60.51	70.74	32.11	11.50	61.30	17.40	5.50	45.38	14.00	27.33	53.46	61.19	64.75

Table 14. **VQA Evaluation: Off-shelf CLIP and SSL models.**

Vision Backbone	IN1k lin.	ADE20K lin.	ADE20K +ms.	NYUd lin. 1 (↓)	NYUd lin. 4 (↓)
CLIP ViT-1B	83.40	41.27	44.06	0.475	0.431
CLIP ViT-2B	84.37	43.18	44.63	0.455	0.435
CLIP ViT-3B	84.76	43.00	43.73	0.452	0.434
CLIP ViT-5B	82.17	43.80	43.69	0.456	0.449
CLIP ViT-7B	82.44	44.83	44.28	0.450	0.433

Table 15. **Classic Vision Evaluation: CLIP trained on MC-2B with 2 billion images seen.**

Vision Enc.	Text Enc. Params	Avg	General	Knowledge	OCR/Chart	Vision
ViT-2B	300M	53.0	72.2	48.8	36.1	55.0
ViT-2B	600M	52.6	73.3	49.0	36.5	51.7
ViT-2B	1B	52.5	73.7	46.9	34.7	54.5

Table 16. **Scaling the CLIP text encoder:** VQA results for scaling the CLIP text encoder with ViT-2B.

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													
AIMv2 3B	DFN-2B + COYO + HQITP	12.0B	336	59.7	75.3	50.1	53.1	60.1	86.5	37.1	45.9	0.340	0.322
			448	63.8	77.0	51.1	62.5	64.5	83.2	37.0	46.9	0.153	0.125
Visual Self-Supervised Models													
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 17. **Comparison with AIMv2.** Web-DINO ViT-7B achieves competitive performance with AIMv2 on VQA without language supervision. On classic vision, DINO performs better at image classification and segmentation, while AIMv2 performs significantly better at depth estimation.

Vision Backbone		General				Knowledge				OCR & Chart				Vision-Centric			
Model	Average	MME^{P}	MMB	$SEED^{I}$	GQA	SQA^{I}	$MMMU^{V}$	$MathVista^{M}$	AI2D	ChartQA	OCRBench	TextVQA	DocVQA	MMVP	RealWorldQA	$CV\text{-}Bench^{2D}$	$CV\text{-}Bench^{3D}$
$Web\text{-}DINO_{224px}$	55.24	1811.05	71.30	72.14	64.04	72.43	35.66	15.20	68.52	35.52	36.40	56.53	29.00	46.00	57.90	70.53	62.08
$Web\text{-}DINO_{378px}$	57.43	1757.06	70.61	72.59	64.50	72.53	35.11	16.10	67.09	52.04	42.19	61.51	46.00	38.00	59.08	66.55	67.16
$Web\text{-}DINO_{518px}$	59.91	1807.08	73.79	72.92	64.78	74.36	34.66	14.50	69.43	57.28	45.70	64.48	53.00	43.33	60.52	70.08	69.41

Table 18. **VQA Evaluation: Web-DINO ViT-7B adapted to different resolution**

Benchmark	Eval	Citation
GQA	General VQA	Hudson and Manning [49]
SEED	General VQA	Ge et al. [35]
MME	General VQA	Fu et al. [32]
MMBench	General VQA	Liu et al. [61]
AI2D	Knowledge VQA	Hiippala et al. [47]
ScienceQA	Knowledge VQA	Lu et al. [64]
MathVista	Knowledge VQA	Lu et al. [65]
MMMU	Knowledge VQA	Yue et al. [105]
TextVQA	OCR & Chart VQA	Singh et al. [80]
DocVQA	OCR & Chart VQA	Mathew et al. [68]
ChartQA	OCR & Chart VQA	Masry et al. [67]
OCRBench	OCR & Chart VQA	Liu et al. [60]
MMVP	Vision-Centric VQA	Tong et al. [93]
RealWorldQA	Vision-Centric VQA	xAI [101]
CVBench-2D	Vision-Centric VQA	Tong et al. [91]
CVBench-3D	Vision-Centric VQA	Tong et al. [91]
ImageNet-1k	Image Classification	Deng et al. [24]
ADE-20k	Image Segmentation	Zhou et al. [111]
NYU Depth v2	Depth Estimation	Silberman et al. [79]

Table 19. **List of benchmarks used**

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

Table 20. **LVD-142M Data Sources.** In contrast to LVD-142M, which relies on highly curated data sources drawn from distributions closely aligned with various downstream evaluation tasks (see the table above from Oquab et al. [73]), our data curation approach adopts the methodology from MetaCLIP [104], utilizing web data collected from 15 snapshots of CommonCrawl (CC) spanning January 2021 through January 2023.