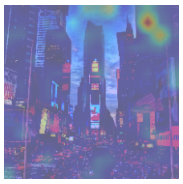
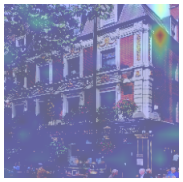
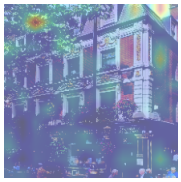
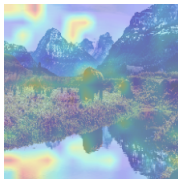
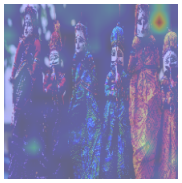
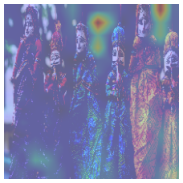
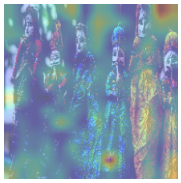
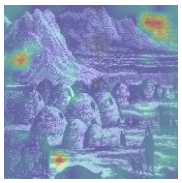
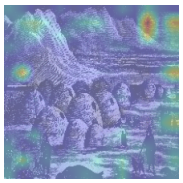
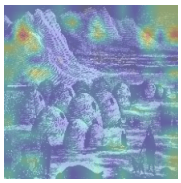
A. Qualitative Examples

Table 1. The attention map visualization of the ViT-L/16 models trained on different scales of data. Images are selected to represent cultures in Western-centric countries and countries where low-resource languages are spoken.

Concept	Image	1B	10B	100B
Street (New York) [1]				
Pub (London) [2]				
Bison (Yellowstone) [3]				
Igorot Dance (Igorot) [4]				
Kathputli Kala Chitra (Hindi) [5]				
Igloo (Inuit) [6]				

[1] By Terabass, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=134418052

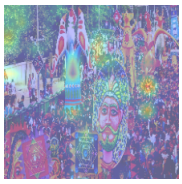
[2] By Ricardalovesmonuments - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=122810839

[3] Source: Yellowstone National Park, https://www.yellowstonenationalparklodges.com/connect/yellowstone-hot-spot/yellowstone-where-the-bison-roam/

[4] Source: Itogon, https://itogon.wordpress.com/2012/04/26/book-goes-to-heart-of-igorot-people/

[5] Source: The Better India, https://thebetterindia.com/57220/journey-indian-handicraft-landscape/

[6] Drawn by unknown artist based on sketches by C.F. Hall and photographed from the book by User:Finetooth - Arctic Researches and Life Among the Esquimaux: Being the Narrative of an Expedition in Search of Sir John Franklin in the Years 1860, 1861, and 1862 by Charles Francis Hall (1865), New York: Harper and Brothers., Public Domain, https://commons.wikimedia.org/w/index.php?curid=3648025

Concept	Image	1B	10B	100B
Pohela Boishakh (Bengali) [7]				

[7] Source: EyeNews, https://www.eyenews.news/english/Today-is-Pahela-Baishakh-the-first-day-of-Bengal-1430/757

B. Scaling Law

Table 2. Evaluations and scaling laws on Western-centric benchmarks, where scaling from 10B to 100B examples shows limited benefits.

Model	Metric (err%)	Value @ 100B ex			Scaling Laws					
					exponent			limit		
		1B	10B	100B	1B	10B	100B	1B	10B	100B
					Zero-shot classification					
B	ImageNet	41.2	39.4	**39.0**	-0.58	-0.97	-0.65	40.1	38.5	**37.9**
	CIFAR100	36.6	**35.9**	36.8	-0.26	-0.23	-0.24	33.8	**32.5**	33.7
	Pet	25.4	23.7	**22.3**	-0.43	-0.45	-0.37	22.3	21.7	**18.4**
L	ImageNet	31.2	29.7	**28.5**	-0.92	-0.91	-0.82	30.7	29.0	**27.1**
	CIFAR100	25.0	23.8	**23.4**	-0.26	-0.32	-0.43	22.7	**20.7**	21.1
	Pet	14.4	12.5	**9.5**	-0.61	-0.57	-0.51	12.3	9.6	**7.0**
H	ImageNet	29.6	25.6	**24.9**	-0.36	-0.64	-0.52	26.7	24.5	**23.3**
	CIFAR100	23.5	**19.8**	21.4	-0.25	-0.36	-0.29	20.6	18.0	**17.6**
	Pet	10.3	7.5	**7.2**	-0.45	-0.42	-0.50	8.1	5.3	**4.6**
					Retrieval @1					
B	COCO I2T@1	56.5	**51.6**	53.4	-0.24	-0.49	-0.30	52.4	**49.9**	50.7
	COCO T2I@1	70.9	**68.8**	70.0	-0.34	-0.39	-0.69	69.6	**67.1**	69.5
	Flickr I2T@1	24.2	21.2	**21.1**	-0.24	-0.34	-0.23	21.5	18.1	**17.0**
	Flickr T2I@1	43.1	**40.3**	40.4	-0.32	-0.42	-0.30	40.9	37.5	**36.7**
L	COCO I2T@1	49.7	47.2	**45.3**	-0.24	-0.41	-0.30	45.8	44.7	**42.9**
	COCO T2I@1	68.2	64.3	**62.5**	-0.19	-0.42	-0.41	64.2	62.6	**60.5**
	Flickr I2T@1	20.4	**15.5**	16.6	-0.21	-0.45	-0.21	16.5	14.1	**13.4**
	Flickr T2I@1	39.9	**32.3**	32.5	-0.10	-0.42	-0.42	34.6	**30.7**	**30.7**
H	COCO I2T@1	48.6	**42.0**	42.5	-0.21	-0.62	-0.47	44.6	**40.3**	40.6
	COCO T2I@1	64.9	60.3	**59.3**	-0.30	-0.55	-0.43	62.8	58.9	**57.3**
	Flickr I2T@1	16.8	**13.5**	13.9	-0.23	-0.40	-0.23	12.2	11.4	**11.3**
	Flickr T2I@1	34.3	28.5	**28.0**	-0.23	-0.56	-0.46	29.6	26.8	**25.9**
					10-shot					
B	Imagenet	46.6	45.6	**44.7**	-0.82	-0.61	-0.49	46.2	44.4	**43.3**
	Birds	53.8	**53.5**	53.9	-0.34	-0.40	-0.51	**51.5**	51.6	52.8
	Caltech	8.4	8.3	**8.2**	-0.30	-0.24	-0.23	7.1	7.2	**6.8**
	Cars	18.3	**16.8**	17.6	-0.63	-0.68	-0.60	17.1	**15.5**	16.3
	CIFAR100	38.7	**38.6**	39.0	-0.19	-0.22	-0.20	35.2	**34.9**	35.9
	Colorectal	**26.5**	29.2	27.0	-0.02	-0.06	-0.16	**20.2**	22.6	24.4
	Pet	22.9	23.2	**22.1**	-1.77	-0.62	-0.77	21.6	21.3	**20.6**
	DTD	**29.7**	30.9	30.9	-0.28	-0.24	-0.19	27.9	28.3	**27.2**
L	Imagenet	35.1	35.0	**33.7**	-0.67	-0.68	-0.63	34.1	34.0	**32.5**
	Birds	**44.0**	45.3	44.3	-0.51	-0.43	-0.51	**42.1**	43.2	42.7
	Caltech	**6.4**	7.4	7.5	-0.43	-0.17	-0.18	5.9	**4.8**	**4.8**
	Cars	**11.1**	11.3	11.5	-0.54	-0.49	-0.41	10.1	**9.7**	9.9
	CIFAR100	27.5	26.7	**25.5**	-0.24	-0.29	-0.41	24.0	23.7	**22.9**
	Colorectal	24.0	23.5	**22.6**	-0.18	-0.20	-0.27	**18.8**	20.2	20.5
	Pet	12.3	12.5	**11.8**	-0.70	-0.65	-0.53	11.3	11.4	**10.3**
	DTD	28.5	**27.1**	27.9	-0.22	-0.25	-0.23	25.2	**25.1**	25.5
H	Imagenet	32.4	29.8	**29.3**	-0.41	-0.73	-0.79	30.3	29.0	**28.3**
	Birds	41.6	39.1	**36.3**	-0.67	-0.52	-0.47	40.6	37.4	**33.9**
	Caltech	**5.7**	6.0	8.9	-0.21	-0.08	-0.11	4.3	**3.7**	4.6
	Cars	11.3	10.3	**9.6**	-0.27	-0.88	-0.44	9.1	10.1	**8.3**
	CIFAR100	25.8	**23.8**	24.2	-0.22	-0.25	-0.24	21.4	21.1	**19.7**
	Colorectal	**25.2**	26.2	25.9	-0.22	-0.20	-0.15	19.7	**17.9**	20.7
	Pet	10.8	9.1	**8.7**	-0.92	-0.48	-0.46	10.3	7.6	**6.5**
	DTD	29.2	**26.1**	26.8	-0.16	-0.23	-0.23	25.0	**23.8**	24.8

Table 3. Evaluations and scaling laws on culture diversity benchmarks, where scaling from 10B to 100B examples shows larger benefits.

Model	Metric (err %)	Value @ 100B ex			Scaling Laws					
					exponent			limit		
		1B	10B	100B	1B	10B	100B	1B	10B	100B
10-shot Geolocalization										
B	Dollar Street	77.7	75.8	**72.1**	-0.38	-0.36	-0.37	76.3	73.7	**70.2**
	GeoDE-Country	72.8	71.5	**71.4**	-0.35	-0.31	-0.37	70.8	69.6	**68.9**
	GeoDE-Region	61.1	60.8	**59.2**	-0.26	-0.22	-0.29	58.8	**57.0**	57.3
L	Dollar Street	63.6	64.1	**58.3**	-1.09	-0.38	-0.94	63.2	60.1	**57.5**
	GeoDE-Country	61.9	62.3	**57.8**	-0.40	-0.30	-1.11	58.8	58.0	**56.6**
	GeoDE-Region	54.2	53.6	**48.3**	-0.15	-0.16	-0.39	49.9	46.9	**46.3**
H	Dollar Street	64.6	59.1	**53.7**	-0.30	-0.56	-0.64	61.0	56.4	**52.5**
	GeoDE-Country	56.9	50.2	**47.6**	-0.23	-0.78	-0.62	52.2	49.4	**46.1**
	GeoDE-Region	54.6	47.6	**44.7**	0.00	-0.38	-0.31	50.1	45.3	**41.0**
Zero-shot classification										
B	Dollar Street	52.0	51.9	**51.6**	-0.38	-0.25	-0.28	50.4	**49.7**	**49.7**
	GeoDE	**7.8**	8.3	8.7	-0.24	-0.26	-0.25	6.1	6.7	**5.4**
	GLDv2	65.0	61.0	**59.4**	-0.46	-0.72	-0.51	61.6	59.3	**56.8**
L	Dollar Street	50.2	**48.1**	49.0	-0.22	-0.35	-0.17	46.9	**46.2**	**46.2**
	GeoDE	6.0	5.9	**4.9**	-0.29	-0.17	-0.25	4.7	4.3	**3.3**
	GLDv2	50.4	46.4	**45.7**	-0.53	-0.93	-0.89	48.5	44.8	**44.1**
H	Dollar Street	50.0	48.6	**47.4**	-0.15	-0.13	-0.20	43.9	44.2	**44.1**
	GeoDE	6.0	4.9	**4.8**	-0.19	-0.22	-0.24	**3.3**	**3.3**	3.5
	GLDv2	48.1	40.1	**38.8**	-0.52	-1.34	-0.80	46.0	39.0	**36.8**

C. Association Bias

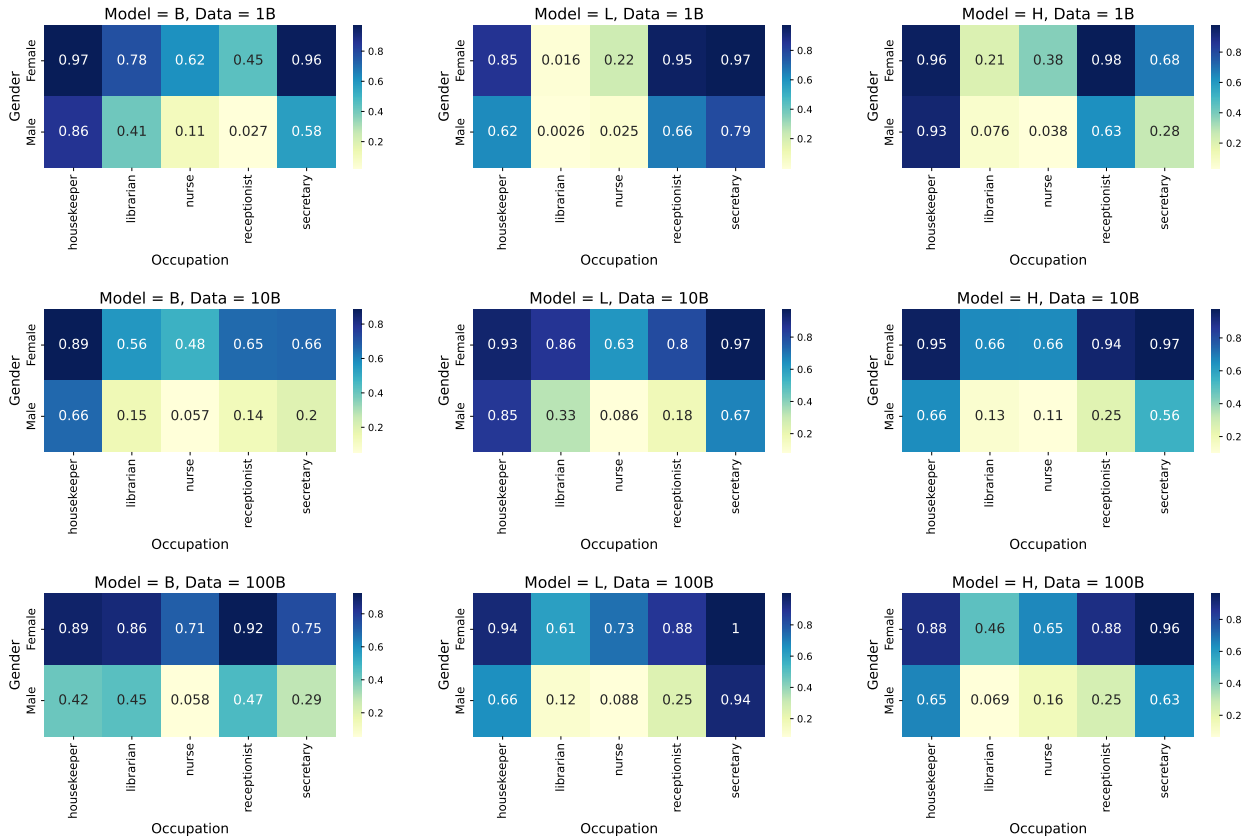


Figure 1. Association bias between gender and occupation, evaluated in scaled models and data.

D. Performance Disparity

Table 4. Performance disparity results for various SigLIP models pretrained on 100 billion seen examples of 1B, 10B, and 100B datasets. Here, disparity corresponds to the maximum gap across subgroups in Dollar Street (by income level) and GeoDE (by geographic region). Pretraining on 100B examples tends to improve disparity overall.

Model	**Data Scale**	**Performance per Subgroup**						**Disparity**
		0-shot Dollar Street						
		0-200	**200-685**	**685-1998**	**>1998**			
B	1B	29.4	43.9	56.5	62.0			32.5
B	10B	31.6	44.0	55.4	61.5			29.9
B	100B	32.0	44.3	56.3	61.0			**29.0**
L	1B	33.7	44.7	57.3	63.4			**29.7**
L	10B	35.7	47.8	58.7	65.5			29.8
L	100B	33.7	46.6	59.5	64.1			30.4
H	1B	32.3	44.9	58.4	64.5			32.2
H	10B	33.9	46.3	58.6	66.9			33.0
H	100B	34.1	48.2	62.2	66.1			**32.1**
		0-shot GeoDE						
		Africa	**Americas**	**East-Asia**	**Europe**	**South-East Asia**	**West Asia**	
B	1B	89.4	92.1	91.8	94.1	92.5	93.4	4.7
B	10B	88.4	91.8	91.4	94.0	92.2	93.0	5.5
B	100B	88.8	91.4	91.0	93.3	91.7	92.2	**4.4**
L	1B	92.0	94.0	94.0	95.2	94.2	94.9	3.2
L	10B	91.8	94.4	94.0	95.8	94.2	94.7	4.0
L	100B	93.5	95.1	95.4	96.2	95.0	95.8	**2.8**
H	1B	91.5	94.4	94.7	95.2	94.1	94.5	3.6
H	10B	93.4	95.4	95.0	96.5	95.1	95.6	3.0
H	100B	93.6	95.1	95.3	96.3	95.2	95.8	**2.7**

E. Evaluations of Data Scaling

Table 5. Detailed evaluation results of ViT-B/L/H models on 1/10/100 billion scale datasets. All metrics are measured by error rate, with the exception of "Representation Bias", which is measured by disparity.

		ViT-B/16			ViT-L/16			ViT-H/14		
Metric	Category	1B	10B	100B	1B	10B	100B	1B	10B	100B
ImageNet 0-shot Classification		41.21	39.35	39.04	31.23	29.70	28.49	29.60	25.60	24.90
Cifar100 0-shot Classification		36.62	35.87	36.80	25.02	23.75	23.36	23.49	19.79	21.42
Pet 0-shot Classification		25.40	23.71	22.27	14.36	12.46	9.46	10.33	7.47	7.17
ImageNet 10-shot Classification		46.65	45.63	44.74	35.11	34.95	33.71	32.44	29.76	29.34
Cifar100 10-shot Classification		38.73	38.63	39.02	27.50	26.70	25.49	25.76	23.79	24.21
Pet 10-shot Classification		22.95	23.19	22.08	12.32	12.48	11.80	10.85	9.13	8.67
Bird 10-shot Classification		53.80	53.47	53.90	44.05	45.25	44.29	41.65	39.13	36.31
Caltech 10-shot Classification	Western	8.37	8.33	8.23	6.41	7.40	7.53	5.70	6.02	8.93
Cars 10-shot Classification		18.29	16.79	17.60	11.14	11.33	11.47	11.32	10.30	9.60
Colorectal Histology 10-shot Classification		26.53	29.23	27.00	24.00	23.53	22.57	25.17	26.17	25.87
DTD 10-shot Classification		29.73	30.85	30.90	28.46	27.07	27.93	29.20	26.12	26.76
COCO Image-Text 0-shot Retrieval		56.46	51.62	53.44	49.70	47.18	45.28	48.62	42.04	42.48
COCO Text-Image 0-shot Retrieval		70.90	68.84	70.01	68.16	64.32	62.51	64.86	60.32	59.29
Flickr Image-Text 0-shot Retrieval		24.20	21.20	21.10	20.40	15.50	16.60	16.80	13.50	13.90
Flickr Text-Image 0-shot Retrieval		43.12	40.26	40.42	39.94	32.32	32.52	34.26	28.46	28.00
Dollar Street 0-shot Classification		52.04	51.88	51.60	50.23	48.10	49.03	50.00	48.58	47.35
Dollar Street 10-shot Classification		77.69	75.81	72.12	63.56	64.09	58.29	64.60	59.10	53.69
GeoDE 0-shot Classification	Culture	7.85	8.27	8.65	6.01	5.90	4.88	5.99	4.87	4.81
GeoDE (country) 10-shot Classification		72.75	71.47	71.36	61.94	62.31	57.85	56.94	50.22	47.55
GeoDE (region) 10-shot Classification		61.09	60.80	59.18	54.21	53.59	48.29	54.56	47.63	44.68
GLDv2 0-shot Classification		65.05	60.96	59.40	50.39	46.37	45.72	48.05	40.08	38.78
Representation Bias		33.15	34.54	35.21	38.18	36.35	35.51	36.76	35.01	36.61
Income 0-200 Classification		70.57	68.43	67.97	66.30	64.35	66.30	67.69	66.11	65.92
Income 200-285 Classification		56.07	55.98	55.70	55.33	52.18	53.38	55.14	53.66	51.81
Income 285-685 Classification		43.45	44.57	43.73	42.71	41.32	40.48	41.60	41.41	37.79
Income >1998 Classification		38.05	38.51	38.98	36.56	34.51	35.91	35.53	33.12	33.86
GeoDE: Africa		10.58	11.56	11.15	7.99	8.24	6.55	8.46	6.56	6.40
GeoDE: Americas	Fairness	7.94	8.16	8.58	6.03	5.57	4.92	5.60	4.57	4.86
GeoDE: EastAsia		8.15	8.57	8.99	5.98	5.96	4.56	5.30	5.01	4.68
GeoDE: Europe		5.92	6.02	6.75	4.81	4.20	3.75	4.83	3.53	3.75
GeoDE: SouthEastAsia		7.51	7.81	8.26	5.78	5.78	5.02	5.86	4.89	4.76
GeoDE: WestAsia		6.57	7.01	7.85	5.11	5.30	4.19	5.50	4.42	4.19
Perceived Gender		9.05	8.34	8.11	5.25	6.06	4.97	5.53	4.59	4.60
Perceived Race		41.44	40.52	46.87	44.57	46.88	46.04	45.47	45.67	49.84
XM3600 Image-Text: Arabic		61.78	53.42	53.36	53.58	45.00	44.56	52.25	41.64	41.00
XM3600 Image-Text: Bengali		95.69	80.64	77.06	90.81	66.36	63.75	88.17	61.22	56.69
XM3600 Image-Text: Czech		60.78	51.89	50.83	52.31	43.81	42.22	49.94	40.11	39.44
XM3600 Image-Text: Danish		55.58	45.39	45.75	45.08	35.06	31.00	43.03	29.92	28.75
XM3600 Image-Text: German		39.47	31.53	31.78	30.61	24.28	24.03	29.17	22.75	21.89
XM3600 Image-Text: Greek		74.36	63.00	61.86	67.86	53.64	50.14	65.67	49.50	47.33
XM3600 Image-Text: English		56.53	55.03	55.50	54.14	52.42	51.67	53.22	51.42	49.64
XM3600 Image-Text: Spanish		49.17	42.94	44.22	41.56	38.44	35.81	40.03	33.89	34.28
XM3600 Image-Text: Persian		58.94	51.17	51.58	49.64	38.97	40.17	46.61	33.72	34.06
XM3600 Image-Text: Finnish		70.64	53.83	53.61	59.25	42.67	39.06	57.39	34.83	32.86
XM3600 Image-Text: Filipino	Multiling	87.86	82.06	81.92	82.72	72.86	71.36	81.31	66.14	63.03
XM3600 Image-Text: French		47.08	38.92	39.06	39.08	31.78	29.92	36.58	28.53	28.19
XM3600 Image-Text: Hindi		83.53	74.78	72.39	77.67	65.67	63.47	76.92	62.33	60.64
XM3600 Image-Text: Croatian		64.53	53.28	51.33	53.08	37.94	35.78	47.81	32.44	30.44
XM3600 Image-Text: Hungarian		64.50	49.06	47.53	53.81	38.64	34.42	51.22	32.67	30.36
XM3600 Image-Text: Indonesian		44.81	38.14	37.08	35.83	28.47	28.53	33.39	24.86	25.33
XM3600 Image-Text: Italian		48.58	41.00	40.86	38.42	33.33	30.97	36.47	29.64	28.89
XM3600 Image-Text: Hebrew		67.06	50.28	49.86	56.75	39.44	35.72	52.03	33.86	30.81
XM3600 Image-Text: Japanese		67.36	55.67	55.42	59.00	45.42	44.97	58.47	42.22	37.94
XM3600 Image-Text: Korean		58.64	49.61	49.53	50.75	40.33	38.31	46.81	35.39	35.08
XM3600 Image-Text: Maori		99.61	99.50	99.42	99.58	99.22	99.25	99.31	98.92	99.17
XM3600 Image-Text: Dutch		53.97	47.47	48.78	47.11	41.14	38.39	44.56	38.06	37.44
XM3600 Image-Text: Norwegian		56.56	46.78	47.89	45.33	36.11	34.28	43.39	31.81	30.19
XM3600 Image-Text: Polish		53.97	44.89	44.22	45.97	35.50	34.11	41.75	33.00	31.06
XM3600 Image-Text: Portuguese		51.03	44.19	44.39	43.33	36.03	34.56	41.14	32.69	32.28
XM3600 Image-Text: Quechua		95.53	94.08	93.89	94.64	93.53	93.92	94.58	93.06	92.78
XM3600 Image-Text: Romanian		64.56	51.39	52.03	52.19	38.31	35.39	47.92	32.36	30.11
XM3600 Image-Text: Russian		51.56	42.36	42.28	42.78	35.14	33.22	41.19	31.97	30.31
XM3600 Image-Text: Swedish		54.03	44.25	45.69	44.50	34.94	34.78	40.69	31.14	30.78
XM3600 Image-Text: Swahili		92.14	88.17	88.72	89.94	81.33	79.47	88.92	76.86	74.14
XM3600 Image-Text: Telugu		98.06	87.08	80.53	96.08	76.67	69.69	96.36	73.08	65.31
XM3600 Image-Text: Thai		79.33	68.67	67.47	72.61	59.47	58.86	71.25	56.86	52.78

XM3600 Image-Text: Turkish		60.33	50.03	50.06	52.78	40.72	39.72	48.56	36.56	34.94
XM3600 Image-Text: Ukrainian		62.39	52.25	49.78	55.19	41.25	37.83	52.75	36.94	33.25
XM3600 Image-Text: Vietnamese		54.31	45.33	45.22	43.19	34.00	32.44	40.75	29.06	29.08
XM3600 Image-Text: Chinese		63.92	51.08	51.19	53.67	42.47	42.50	54.17	40.53	38.42
XM3600 Text-Image: Arabic		73.77	67.79	68.49	67.49	59.74	59.86	65.87	56.22	54.91
XM3600 Text-Image: Bengali		97.19	89.25	89.53	95.17	79.72	77.31	94.22	76.36	72.42
XM3600 Text-Image: Czech		71.81	64.49	65.48	65.52	58.57	58.18	63.59	55.79	55.07
XM3600 Text-Image: Danish		68.23	59.97	61.73	60.01	51.18	49.50	56.72	46.53	45.46
XM3600 Text-Image: German		55.15	47.80	49.18	45.85	39.88	39.75	43.80	36.56	36.99
XM3600 Text-Image: Greek		82.61	75.69	75.71	77.96	69.11	67.35	75.68	65.45	64.10
XM3600 Text-Image: English		62.32	59.41	60.78	58.97	57.57	56.32	58.15	56.40	55.82
XM3600 Text-Image: Spanish		57.35	52.74	55.49	52.64	49.06	48.31	51.24	47.27	46.62
XM3600 Text-Image: Persian		71.80	65.18	65.58	62.65	55.06	56.09	59.79	52.93	49.69
XM3600 Text-Image: Finnish		81.00	70.80	68.28	72.96	59.11	56.24	70.79	51.07	49.35
XM3600 Text-Image: Filipino		93.60	90.28	91.07	90.89	83.98	83.70	89.55	80.61	77.92
XM3600 Text-Image: French		56.70	50.23	50.57	48.33	43.31	42.10	47.52	40.48	39.96
XM3600 Text-Image: Hindi		91.01	86.55	86.09	87.43	81.38	80.01	87.71	79.21	78.22
XM3600 Text-Image: Croatian		75.52	67.53	66.85	66.68	54.42	54.22	63.21	50.71	48.53
XM3600 Text-Image: Hungarian		74.24	63.83	63.53	66.49	53.73	50.75	64.26	48.31	45.72
XM3600 Text-Image: Indonesian		60.08	52.90	53.96	50.28	44.05	43.97	49.27	41.45	40.81
XM3600 Text-Image: Italian		57.90	51.51	52.08	47.96	42.80	42.60	48.03	40.62	40.34
XM3600 Text-Image: Hebrew		76.50	64.76	62.76	69.11	56.25	54.14	65.88	51.49	49.99
XM3600 Text-Image: Japanese		76.74	69.20	68.99	69.56	62.34	58.44	69.16	57.06	54.78
XM3600 Text-Image: Korean		70.82	64.88	67.23	64.52	56.76	56.51	61.52	53.57	52.67
XM3600 Text-Image: Maori		99.78	99.78	99.78	99.73	99.56	99.62	99.75	99.67	99.51
XM3600 Text-Image: Dutch		63.50	59.25	59.05	57.41	52.02	51.48	55.49	49.88	49.10
XM3600 Text-Image: Norwegian		70.36	63.58	63.44	61.54	53.81	52.99	60.04	49.16	48.20
XM3600 Text-Image: Polish		63.73	57.39	57.71	56.06	47.92	47.09	53.28	45.05	44.49
XM3600 Text-Image: Portuguese		62.16	57.16	57.93	54.54	49.48	48.72	52.44	47.48	46.64
XM3600 Text-Image: Quechua		98.46	97.94	97.85	97.88	98.14	98.04	98.18	98.28	98.26
XM3600 Text-Image: Romanian		74.48	65.48	65.11	65.20	54.05	52.41	61.69	48.77	47.09
XM3600 Text-Image: Russian		61.65	53.83	54.17	53.47	47.58	45.36	51.60	43.58	43.08
XM3600 Text-Image: Swedish		66.11	59.05	60.50	58.78	50.72	51.82	55.34	47.66	47.93
XM3600 Text-Image: Swahili		96.30	94.01	94.73	94.55	90.09	89.57	93.85	87.47	85.67
XM3600 Text-Image: Telugu		98.76	92.69	90.40	97.76	87.47	83.03	98.18	84.44	79.57
XM3600 Text-Image: Thai		86.81	80.38	79.47	81.83	74.60	73.67	82.21	73.31	69.67
XM3600 Text-Image: Turkish		72.31	65.24	65.17	65.21	55.12	56.70	62.35	53.59	52.19
XM3600 Text-Image: Ukrainian		75.01	66.08	65.35	68.84	57.74	55.32	66.07	54.18	50.84
XM3600 Text-Image: Vietnamese		70.38	64.82	64.64	61.84	54.00	53.39	58.46	50.29	48.76
XM3600 Text-Image: Chinese		73.98	64.78	64.96	64.87	59.03	57.33	65.25	56.15	56.68
Average Western 0-shot Classification	Western	34.41	32.98	32.70	23.54	21.97	20.44	21.14	17.62	17.83
Average Western 10-shot Classification		30.63	30.77	30.43	23.62	23.59	23.10	22.76	21.30	21.21
Average Western 0-shot Retrieval		48.67	45.48	46.24	44.55	39.83	39.23	41.13	36.08	35.92
Average Western Classification		31.66	31.37	31.05	23.60	23.15	22.37	22.32	20.30	20.29
Average Dollar Street Classification	Culture	64.87	63.85	61.86	56.89	56.09	53.66	57.30	53.84	50.52
Average GeoDE Classification		47.23	46.85	46.39	40.72	40.60	37.01	39.16	34.24	32.35
Average Income Classification	Fairness	52.03	51.87	51.59	50.22	48.09	49.02	49.99	48.57	47.35
Average Geographic Classification		7.78	8.19	8.59	5.95	5.84	4.83	5.92	4.83	4.77
Average Demography Classification		25.24	24.43	27.49	24.91	26.47	25.50	25.50	25.13	27.22
Average Multiling: Low-Resource Lang	Multiling	91.22	84.27	83.16	87.73	77.14	75.01	86.58	73.69	70.93
Average Multiling: High-Resource Lang		63.66	55.42	55.53	55.54	46.75	45.43	53.38	43.11	41.81
Average Western		36.20	35.13	35.10	29.19	27.60	26.87	27.34	24.51	24.46
Average Culture		56.08	54.87	53.72	47.72	46.72	44.01	46.69	41.75	39.48
Average Fairness		25.44	25.46	26.08	23.87	23.36	23.01	23.88	22.80	22.70
Average Multiling		65.23	56.09	55.61	57.52	47.23	45.40	55.38	43.33	41.63

F. Evaluations of Transferability to Generative Models

The downstream tasks in Table 6 are categorized as the following groups and reported in Table **??**:

1. **Semantics**: "COCOcap", "NoCaps", "COCO-35L (en)", "XM3600 (en)", "OKVQA", "AOKVQA-MC (val)", "AOKVQA-DA (val)", "GQA", "NLVR2", "MARVL (avg5)", "VizWizVQA (val)", "TallyQA (simple)", "TallyQA (complex)", "CountBenchQA", "RefCOCO (testA)", "RefCOCO (testB)", "RefCOCO+ (testA)", "RefCOCO+ (testB)", "RefCOCOg (test)"
2. **OCR**: "DocVQA (val)", "OCR-VQA", "ChartQA (avg)", "ChartQA (human)", "ChartQA (aug)", "SciCap", "AI2D", "ScienceQA", "InfoVQA (val)", "TextCaps", "TextVQA (val)", "ST-VQA (val)", "Screen2Words", "WidgetCap"
3. **Multilinguality**: "xGQA (avg8)", "XM3600 (avg36)", "COCO-35L (avg35)"
4. **Remote Sensing**: "RSVQA-lr", "RSVQA-hr (test)", "RSVQA-hr (test2)"

Table 6. Detailed evaluation results of the transferability of contrastively trained vision models (ViT-L/16) to generative vision-language models (PaliGemma), with both frozen and unfrozen setups. Task-specific Numbers are reported for vision models trained on 1 billion, 10 billion and 100 billion raw data respectively, using PaliGemma's default fine-tuning configuration.

	Frozen ViT			Unfrozen ViT		
Metric	1B Data	10B Data	100B Data	1B Data	10B Data	100B Data
COCOcap	134.6	132.9	134.4	135.0	132.1	134.0
NoCaps	114.1	110.5	112.8	113.4	111.4	113.3
COCO-35L (avg35)	107.6	105.9	108.0	107.7	106.8	107.8
COCO-35L (avg34)	106.9	105.2	107.3	107.0	106.0	107.1
COCO-35L (en)	130.6	130.4	133.4	132.4	132.5	133.4
XM3600 (en)	75.5	74.9	75.2	75.3	75.4	76.0
XM3600 (avg36)	37.9	36.9	38.0	37.7	37.5	38.0
Screen2Words	108.9	107.5	109.9	105.0	105.3	105.5
TextCaps	86.5	79.3	93.2	87.6	81.8	83.8
SciCap	149.7	146.9	150.0	146.1	144.6	147.1
WidgetCap	120.1	109.6	117.9	113.3	108.4	114.9
VQAv2 (minival)	79.4	78.8	79.8	79.2	78.6	78.6
OKVQA	60.4	59.6	59.7	59.6	59.7	59.9
AOKVQA-MC (val)	74.2	72.7	73.0	73.0	72.7	74.2
AOKVQA-DA (val)	58.5	56.8	57.3	59.1	57.7	57.9
GQA	63.4	63.5	63.6	63.8	63.0	63.5
NLVR2	87.5	86.7	87.2	86.4	86.4	87.0
MARVL (avg5)	76.7	76.2	76.6	76.3	76.8	77.0
AI2D	69.8	70.0	70.6	68.2	68.5	68.6
ScienceQA	95.4	94.9	94.4	94.5	92.9	94.7
RSVQA-lr	93.0	92.4	92.3	93.6	92.8	93.0
RSVQA-hr (test)	92.5	92.5	92.7	92.6	92.6	92.6
RSVQA-hr (test2)	90.4	90.4	90.5	90.5	90.4	90.6
ChartQA (avg)	45.1	43.6	45.0	41.4	40.3	42.5
ChartQA (human)	31.8	31.8	32.6	29.8	28.3	30.5
ChartQA (aug)	58.5	55.4	57.4	53.0	52.3	54.5
VizWizVQA (val)	72.3	71.2	72.8	72.0	71.6	71.9
TallyQA (simple)	76.6	75.7	75.9	76.6	75.7	76.9
TallyQA (complex)	65.0	65	65.5	65.4	64.5	65.3
CountBenchQA	68.2	69.0	67.3	60.6	61.2	63.7
OCR-VQA	68.3	67.5	68.2	66.9	66.0	67.1
TextVQA (val)	44.5	41.4	44.7	41.2	40.4	41.2
DocVQA (val)	25.0	23.5	25.8	23.4	21.7	23.1
InfoVQA (val)	22.3	22.2	23	21.4	22.0	22.1
ST-VQA (val)	46.6	42.8	46.7	43.5	40.1	43.2
xGQA (avg8)	55.2	55.2	55	55.6	54.5	54.8
xGQA (avg7)	54.1	54.0	53.8	54.5	53.3	53.6
RefCOCO (testA)	67.4	67.5	67.9	64.5	64.2	65.1
RefCOCO (testB)	62.7	62.0	63.8	60.2	59.6	60.9
RefCOCO+ (testA)	63	62.7	63.5	60.2	59.9	60.3
RefCOCO+ (testB)	55.6	54.9	56.2	53.2	52.5	53.3
RefCOCOg (test)	59.1	58.9	60	56.5	56.1	57.2
Avg Semantics	77.1	76.4	77.2	76.0	75.4	76.4
Avg OCR	69.5	66.9	70.0	66.8	65.2	67.0
Avg Multilinguality	66.9	66.0	67.0	67.0	66.3	66.9
Avg Remote Sensing	92.0	91.8	91.8	92.3	91.9	92.1
Avg	75.1	73.7	75.3	73.6	72.7	73.9

G. Evaluations of Data Quality Filtering

Table 7. Detailed evaluation results of data quality filtering on ViT-L/16 models. All evaluations are conducted on datasets of 5 billion image-text pairs and across different number of seen examples. All metrics are measured by error rate, with the exception of "Representation Bias", which is measured by disparity.

Metric	Filter	1B	5B	10B	20B	30B
ImageNet 0-shot Classification	Baseline (en)	34.67	28.17	26.68	26.15	24.32
	CLIP filtered	31.18	26.76	25.14	24.39	23.90
	Other filtered	34.50	29.52	28.13	26.70	26.45
Cifar100 0-shot Classification	Baseline (en)	33.05	26.08	24.37	24.52	23.99
	CLIP filtered	31.69	26.96	25.37	24.68	25.76
	Other filtered	36.07	35.27	29.95	32.58	30.78
Pet 0-shot Classification	Baseline (en)	17.25	11.99	11.69	9.13	8.72
	CLIP filtered	13.68	10.49	8.78	8.59	8.23
	Other filtered	14.04	9.62	8.99	7.28	6.62
ImageNet 10-shot Classification	Baseline (en)	42.41	35.25	33.17	33.17	30.68
	CLIP filtered	38.57	32.53	30.60	29.20	28.72
	Other filtered	38.32	32.32	30.42	29.05	28.46
Cifar100 10-shot Classification	Baseline (en)	36.61	30.02	27.39	27.23	26.82
	CLIP filtered	32.83	28.44	28.04	26.20	27.40
	Other filtered	35.30	35.56	31.18	32.26	31.79
Pet 10-shot Classification	Baseline (en)	22.95	16.93	15.32	15.26	11.72
	CLIP filtered	17.31	11.72	10.44	8.97	8.83
	Other filtered	14.15	10.38	9.08	7.63	7.52
Bird 10-shot Classification	Baseline (en)	41.18	31.69	29.91	29.60	27.37
	CLIP filtered	32.38	25.20	23.85	22.21	21.95
	Other filtered	34.57	27.01	26.30	24.65	23.73
Caltech 10-shot Classification	Baseline (en)	10.45	9.94	9.34	9.63	9.60
	CLIP filtered	11.18	10.68	10.44	10.50	10.50
	Other filtered	8.97	9.25	9.01	8.30	9.06
Cars 10-shot Classification	Baseline (en)	16.47	11.03	10.16	10.05	8.94
	CLIP filtered	13.07	9.70	8.89	7.75	8.01
	Other filtered	16.84	13.07	12.52	11.30	11.30
Colorectal Histology 10-shot Classification	Baseline (en)	27.80	27.17	24.77	27.03	25.33
	CLIP filtered	25.97	22.90	20.80	24.23	27.13
	Other filtered	24.53	24.70	25.47	27.10	26.53
DTD 10-shot Classification	Baseline (en)	31.12	26.91	26.33	26.97	26.86
	CLIP filtered	29.20	25.69	25.37	23.51	23.72
	Other filtered	28.09	26.81	24.73	24.52	23.56
COCO Image-Text 0-shot Retrieval	Baseline (en)	46.80	40.28	39.30	39.18	37.04
	CLIP filtered	41.06	36.04	36.48	34.84	34.02
	Other filtered	42.92	38.32	36.80	35.96	36.24
COCO Text-Image 0-shot Retrieval	Baseline (en)	62.26	56.78	54.78	55.22	53.20
	CLIP filtered	59.11	55.27	54.45	53.12	53.03
	Other filtered	60.53	56.01	54.60	53.23	53.27
Flickr Image-Text 0-shot Retrieval	Baseline (en)	16.70	11.30	11.30	11.30	10.90
	CLIP filtered	14.80	9.90	9.70	9.60	8.90
	Other filtered	16.70	13.80	12.60	13.10	12.00
Flickr Text-Image 0-shot Retrieval	Baseline (en)	32.26	24.78	24.74	24.90	22.66
	CLIP filtered	29.52	24.98	23.34	22.12	22.02
	Other filtered	32.84	27.18	26.48	24.82	24.32
Dollar Street 0-shot Classification	Baseline (en)	54.67	50.44	49.81	49.98	49.37
	CLIP filtered	53.71	52.58	51.88	50.63	51.44
	Other filtered	50.23	47.63	47.86	47.45	47.08
Dollar Street 10-shot Classification	Baseline (en)	84.87	79.27	77.18	76.21	72.54

	CLIP filtered	88.86	84.59	84.73	82.80	82.80
	Other filtered	90.16	89.46	87.91	88.72	87.77
GeoDE 0-shot Classification	Baseline (en)	8.98	6.48	6.43	6.26	6.23
	CLIP filtered	9.64	8.54	8.02	7.42	7.22
	Other filtered	9.50	7.69	7.50	7.50	7.53
GeoDE (country) 10-shot Classification	Baseline (en)	84.29	77.28	73.22	73.37	68.85
	CLIP filtered	85.82	81.98	80.11	78.08	78.24
	Other filtered	91.37	89.52	88.30	87.65	86.76
GeoDE (region) 10-shot Classification	Baseline (en)	66.67	61.66	57.71	58.77	55.78
	CLIP filtered	70.68	68.16	66.99	64.81	63.68
	Other filtered	75.82	72.39	72.95	72.13	71.27
GLDv2 0-shot Classification	Baseline (en)	65.50	53.18	50.13	49.48	44.16
	CLIP filtered	61.15	52.46	49.55	47.41	46.37
	Other filtered	80.87	74.06	72.37	72.37	70.17
Representation Bias	Baseline (en)	33.89	28.22	36.00	33.52	30.96
	CLIP filtered	11.46	19.14	20.03	26.57	14.05
	Other filtered	39.31	36.44	39.01	40.57	35.51
Income 0-200 Classification	Baseline (en)	71.31	67.22	68.34	67.50	67.04
	CLIP filtered	69.36	69.36	68.71	66.67	67.87
	Other filtered	69.36	67.97	65.65	66.11	66.67
Income 200-285 Classification	Baseline (en)	60.15	55.33	54.87	54.49	55.33
	CLIP filtered	58.48	57.46	56.63	54.59	56.63
	Other filtered	54.22	50.88	52.64	51.16	51.16
Income 285-685 Classification	Baseline (en)	46.61	42.99	41.04	42.43	40.76
	CLIP filtered	46.43	44.75	44.20	42.90	43.45
	Other filtered	40.95	39.09	39.37	39.37	37.70
Income >1998 Classification	Baseline (en)	40.56	36.19	34.98	35.44	34.33
	CLIP filtered	40.56	38.70	37.95	38.33	37.77
	Other filtered	36.37	32.56	33.77	33.12	32.74
Africa	Baseline (en)	11.51	8.19	7.88	7.72	7.85
	CLIP filtered	11.00	9.74	9.37	9.28	8.44
	Other filtered	12.04	9.97	9.51	9.85	9.88
Americas	Baseline (en)	8.59	6.74	6.15	6.37	6.27
	CLIP filtered	9.57	8.60	8.30	7.29	7.16
	Other filtered	9.63	7.68	7.32	7.53	7.48
EastAsia	Baseline (en)	9.90	7.10	7.37	7.29	6.71
	CLIP filtered	10.45	9.34	8.88	7.72	7.67
	Other filtered	10.52	8.92	8.63	8.21	8.48
Europe	Baseline (en)	6.75	4.82	5.29	5.01	5.17
	CLIP filtered	7.71	6.89	6.52	5.52	6.01
	Other filtered	7.29	5.62	5.57	5.45	5.51
SouthEastAsia	Baseline (en)	8.69	6.23	6.00	5.77	6.01
	CLIP filtered	9.74	8.47	7.40	7.74	7.32
	Other filtered	8.89	7.28	7.47	7.16	7.11
WestAsia	Baseline (en)	8.14	5.61	5.64	5.17	5.08
	CLIP filtered	9.24	8.16	7.59	6.75	6.52
	Other filtered	8.32	6.34	6.17	6.47	6.35
Perceived Gender	Baseline (en)	8.41	6.42	5.78	5.98	5.64
	CLIP filtered	8.43	7.63	8.08	7.56	6.35
	Other filtered	13.08	10.64	11.13	11.02	9.55
Perceived Race	Baseline (en)	37.87	44.74	43.93	48.30	44.95
	CLIP filtered	33.08	40.63	38.98	41.89	43.04
	Other filtered	53.52	52.46	53.21	52.83	56.43
Average Western 0-shot Classification	Baseline (en)	28.33	22.08	20.91	19.93	19.01
	CLIP filtered	25.52	21.40	19.76	19.22	19.30
	Other filtered	28.20	24.81	22.36	22.18	21.28

Average Western 10-shot Classification	Baseline (en)	28.62	23.62	22.05	22.37	20.92
	CLIP filtered	25.06	20.86	19.80	19.07	19.53
	Other filtered	25.10	22.39	21.09	20.60	20.25
Average Western 0-shot Retrieval	Baseline (en)	39.50	33.29	32.53	32.65	30.95
	CLIP filtered	36.12	31.55	30.99	29.92	29.49
	Other filtered	38.25	33.83	32.62	31.78	31.46
Average Western Classification	Baseline (en)	28.54	23.20	21.74	21.70	20.40
	CLIP filtered	25.19	21.01	19.79	19.11	19.47
	Other filtered	25.94	23.05	21.43	21.03	20.53
Average Dollar Street Classification	Baseline (en)	69.77	64.86	63.50	63.09	60.96
	CLIP filtered	71.29	68.58	68.30	66.71	67.12
	Other filtered	70.19	68.55	67.89	68.08	67.42
Average GeoDE Classification	Baseline (en)	53.32	48.48	45.79	46.13	43.62
	CLIP filtered	55.38	52.89	51.71	50.10	49.71
	Other filtered	58.90	56.54	56.25	55.76	55.18
Average Income Classification	Baseline (en)	54.66	50.43	49.81	49.97	49.36
	CLIP filtered	53.71	52.57	51.87	50.62	51.43
	Other filtered	50.22	47.62	47.86	47.44	47.07
Average Geographic Classification	Baseline (en)	8.93	6.44	6.39	6.22	6.18
	CLIP filtered	9.62	8.53	8.01	7.39	7.19
	Other filtered	9.45	7.63	7.44	7.45	7.47
Average Demography Classification	Baseline (en)	23.14	25.58	24.86	27.14	25.30
	CLIP filtered	20.76	24.13	23.53	24.72	24.70
	Other filtered	33.30	31.55	32.17	31.93	32.99
Average Western	Baseline (en)	31.47	25.89	24.62	24.62	23.21
	CLIP filtered	28.10	23.82	22.78	21.99	22.14
	Other filtered	29.22	25.92	24.42	23.90	23.44
Average Culture Diversity	Baseline (en)	60.83	54.72	52.41	52.34	49.49
	CLIP filtered	61.64	58.05	56.88	55.19	54.96
	Other filtered	66.33	63.46	62.82	62.64	61.76
Average Fairness	Baseline (en)	26.54	24.30	23.94	24.29	23.76
	CLIP filtered	26.17	25.81	25.22	24.69	24.85
	Other filtered	27.02	24.95	25.04	24.86	24.92

H. Evaluations of Language Rebalancing

Table 8. Detailed evaluation results of the rebalancing of low-resource languages on ViT-L/16 models and datasets of 1/10/100 billion scales, with 100 billion examples seen in training. All metrics are measured by error rate, with the exception of "Representation Bias", which is measured by disparity.

	1B Data		10B Data		100B Data	
Metric	Before	After	Before	After	Before	After
ImageNet 0-shot Classification	31.23	31.39	29.70	30.47	28.49	28.80
Cifar100 0-shot Classification	25.02	24.96	23.75	24.04	23.36	23.51
Pet 0-shot Classification	14.36	13.00	12.46	12.05	9.46	11.23
ImageNet 10-shot Classification	35.11	34.94	34.95	34.99	33.71	33.89
Cifar100 10-shot Classification	27.50	27.82	26.70	26.50	25.49	25.05
Pet 10-shot Classification	12.32	13.71	12.48	15.59	11.80	13.46
Bird 10-shot Classification	44.05	42.75	45.25	45.29	44.29	42.89
Caltech 10-shot Classification	6.41	8.09	7.40	8.97	7.53	8.35
Cars 10-shot Classification	11.14	11.34	11.33	11.54	11.47	11.21
Colorectal Histology 10-shot Classification	24.00	25.50	23.53	24.43	22.57	28.00
DTD 10-shot Classification	28.46	29.31	27.07	27.39	27.93	29.04
COCO Image-Text 0-shot Retrieval	49.70	52.92	47.18	50.28	45.28	45.90
COCO Text-Image 0-shot Retrieval	68.16	67.50	64.32	63.60	62.51	62.16
Flickr Image-Text 0-shot Retrieval	20.40	24.30	15.50	20.30	16.60	16.40
Flickr Text-Image 0-shot Retrieval	39.94	37.88	32.32	32.64	32.52	33.30
Dollar Street 0-shot Classification	50.23	51.16	48.10	49.42	49.03	49.23
Dollar Street 10-shot Classification	63.56	65.04	64.09	65.51	58.29	59.42
GeoDE 0-shot Classification	6.01	6.03	5.90	5.97	4.88	5.42
GeoDE (country) 10-shot Classification	61.94	59.79	62.31	60.52	57.85	53.34
GeoDE (region) 10-shot Classification	54.21	53.99	53.59	53.30	48.29	48.05
GLDv2 0-shot Classification	50.39	51.82	46.37	47.73	45.72	44.29
Representation Bias	38.18	35.21	36.35	32.61	35.51	32.74
Income 0-200 Classification	66.30	67.32	64.35	65.83	66.30	65.37
Income 200-285 Classification	55.33	54.22	52.18	53.48	53.38	53.20
Income 285-685 Classification	42.71	44.75	41.32	42.80	40.48	40.76
Income >1998 Classification	36.56	38.33	34.51	35.53	35.91	37.58
Africa	7.99	8.34	8.24	7.81	6.55	7.46
Americas	6.03	5.51	5.57	5.84	4.92	5.20
EastAsia	5.98	6.07	5.96	5.90	4.56	5.27
Europe	4.81	4.41	4.20	4.23	3.75	4.00
SouthEastAsia	5.78	6.21	5.78	6.15	5.02	5.50
WestAsia	5.11	5.30	5.30	5.67	4.19	4.79
Perceived Gender	5.25	5.27	6.06	5.96	4.97	5.03
Perceived Race	44.57	49.02	46.88	45.89	46.04	47.35
Crossmodal-3600 Image-Text Retrieval: Arabic	53.58	56.44	45.00	45.89	44.56	44.78
Crossmodal-3600 Image-Text Retrieval: Bengali	90.81	76.03	66.36	63.53	63.75	61.47
Crossmodal-3600 Image-Text Retrieval: Czech	52.31	52.81	43.81	43.36	42.22	41.61
Crossmodal-3600 Image-Text Retrieval: Danish	45.08	45.22	35.06	34.81	31.00	32.53
Crossmodal-3600 Image-Text Retrieval: German	30.61	32.00	24.28	24.36	24.03	23.11
Crossmodal-3600 Image-Text Retrieval: Greek	67.86	70.17	53.64	53.42	50.14	51.94
Crossmodal-3600 Image-Text Retrieval: English	54.14	54.58	52.42	51.58	51.67	50.89
Crossmodal-3600 Image-Text Retrieval: Spanish	41.56	43.50	38.44	38.00	35.81	35.89
Crossmodal-3600 Image-Text Retrieval: Persian	49.64	55.33	38.97	41.97	40.17	38.11
Crossmodal-3600 Image-Text Retrieval: Finnish	59.25	60.11	42.67	42.42	39.06	40.28
Crossmodal-3600 Image-Text Retrieval: Filipino	82.72	72.56	72.86	62.72	71.36	60.22
Crossmodal-3600 Image-Text Retrieval: French	39.08	39.72	31.78	31.47	29.92	29.61
Crossmodal-3600 Image-Text Retrieval: Hindi	77.67	71.67	65.67	65.44	63.47	63.53
Crossmodal-3600 Image-Text Retrieval: Croatian	53.08	53.72	37.94	38.86	35.78	35.64
Crossmodal-3600 Image-Text Retrieval: Hungarian	53.81	54.61	38.64	37.81	34.42	34.78
Crossmodal-3600 Image-Text Retrieval: Indonesian	35.83	37.47	28.47	30.94	28.53	28.42
Crossmodal-3600 Image-Text Retrieval: Italian	38.42	40.69	33.33	33.50	30.97	31.03
Crossmodal-3600 Image-Text Retrieval: Hebrew	56.75	47.75	39.44	37.39	35.72	34.19
Crossmodal-3600 Image-Text Retrieval: Japanese	59.00	61.58	45.42	45.78	44.97	46.69
Crossmodal-3600 Image-Text Retrieval: Korean	50.75	53.06	40.33	40.00	38.31	38.58
Crossmodal-3600 Image-Text Retrieval: Maori	99.58	97.94	99.22	95.00	99.25	96.08
Crossmodal-3600 Image-Text Retrieval: Dutch	47.11	48.06	41.14	41.42	38.39	39.94

Crossmodal-3600 Image-Text Retrieval: Norwegian	45.33	46.81	36.11	36.72	34.28	34.47
Crossmodal-3600 Image-Text Retrieval: Polish	45.97	45.81	35.50	35.61	34.11	34.33
Crossmodal-3600 Image-Text Retrieval: Portuguese	43.33	42.53	36.03	38.33	34.56	34.11
Crossmodal-3600 Image-Text Retrieval: Quechua	94.64	94.97	93.53	93.83	93.92	93.42
Crossmodal-3600 Image-Text Retrieval: Romanian	52.19	52.72	38.31	38.06	35.39	34.86
Crossmodal-3600 Image-Text Retrieval: Russian	42.78	45.00	35.14	35.11	33.22	33.42
Crossmodal-3600 Image-Text Retrieval: Swedish	44.50	46.19	34.94	36.06	34.78	34.19
Crossmodal-3600 Image-Text Retrieval: Swahili	89.94	75.06	81.33	67.64	79.47	65.81
Crossmodal-3600 Image-Text Retrieval: Telugu	96.08	81.00	76.67	67.78	69.69	66.33
Crossmodal-3600 Image-Text Retrieval: Thai	72.61	74.72	59.47	60.50	58.86	59.92
Crossmodal-3600 Image-Text Retrieval: Turkish	52.78	54.94	40.72	41.25	39.72	39.89
Crossmodal-3600 Image-Text Retrieval: Ukrainian	55.19	57.33	41.25	40.97	37.83	39.19
Crossmodal-3600 Image-Text Retrieval: Vietnamese	43.19	42.22	34.00	35.22	32.44	32.86
Crossmodal-3600 Image-Text Retrieval: Chinese	53.67	54.81	42.47	44.67	42.50	43.97
Crossmodal-3600 Text-Image Retrieval: Arabic	67.49	65.43	59.74	59.02	59.86	59.70
Crossmodal-3600 Text-Image Retrieval: Bengali	95.17	83.83	79.72	75.56	77.31	73.33
Crossmodal-3600 Text-Image Retrieval: Czech	65.52	65.19	58.57	59.19	58.18	57.56
Crossmodal-3600 Text-Image Retrieval: Danish	60.01	59.93	51.18	52.77	49.50	49.74
Crossmodal-3600 Text-Image Retrieval: German	45.85	47.48	39.88	40.72	39.75	39.50
Crossmodal-3600 Text-Image Retrieval: Greek	77.96	75.46	69.11	69.24	67.35	68.25
Crossmodal-3600 Text-Image Retrieval: English	58.97	56.93	57.57	57.52	56.32	56.51
Crossmodal-3600 Text-Image Retrieval: Spanish	52.64	52.79	49.06	49.90	48.31	48.76
Crossmodal-3600 Text-Image Retrieval: Persian	62.65	63.27	55.06	55.54	56.09	54.64
Crossmodal-3600 Text-Image Retrieval: Finnish	72.96	72.06	59.11	58.61	56.24	56.42
Crossmodal-3600 Text-Image Retrieval: Filipino	90.89	83.32	83.98	78.41	83.70	74.94
Crossmodal-3600 Text-Image Retrieval: French	48.33	49.81	43.31	44.62	42.10	42.34
Crossmodal-3600 Text-Image Retrieval: Hindi	87.43	83.45	81.38	80.96	80.01	79.22
Crossmodal-3600 Text-Image Retrieval: Croatian	66.68	65.73	54.42	56.10	54.22	53.60
Crossmodal-3600 Text-Image Retrieval: Hungarian	66.49	66.66	53.73	54.57	50.75	51.16
Crossmodal-3600 Text-Image Retrieval: Indonesian	50.28	49.62	44.05	44.58	43.97	44.30
Crossmodal-3600 Text-Image Retrieval: Italian	47.96	49.51	42.80	45.41	42.60	42.66
Crossmodal-3600 Text-Image Retrieval: Hebrew	69.11	60.25	56.25	55.62	54.14	51.65
Crossmodal-3600 Text-Image Retrieval: Japanese	69.56	71.62	62.34	63.34	58.44	61.42
Crossmodal-3600 Text-Image Retrieval: Korean	64.52	64.72	56.76	57.83	56.51	57.58
Crossmodal-3600 Text-Image Retrieval: Maori	99.73	97.92	99.56	96.30	99.62	96.19
Crossmodal-3600 Text-Image Retrieval: Dutch	57.41	58.78	52.02	53.88	51.48	51.82
Crossmodal-3600 Text-Image Retrieval: Norwegian	61.54	61.46	53.81	54.35	52.99	53.50
Crossmodal-3600 Text-Image Retrieval: Polish	56.06	56.43	47.92	49.96	47.09	47.16
Crossmodal-3600 Text-Image Retrieval: Portuguese	54.54	54.07	49.48	51.03	48.72	48.34
Crossmodal-3600 Text-Image Retrieval: Quechua	97.88	97.89	98.14	98.03	98.04	97.88
Crossmodal-3600 Text-Image Retrieval: Romanian	65.20	65.55	54.05	54.79	52.41	51.93
Crossmodal-3600 Text-Image Retrieval: Russian	53.47	53.75	47.58	48.43	45.36	46.83
Crossmodal-3600 Text-Image Retrieval: Swedish	58.78	59.12	50.72	52.50	51.82	50.97
Crossmodal-3600 Text-Image Retrieval: Swahili	94.55	84.91	90.09	80.20	89.57	78.20
Crossmodal-3600 Text-Image Retrieval: Telugu	97.76	87.85	87.47	82.04	83.03	80.15
Crossmodal-3600 Text-Image Retrieval: Thai	81.83	80.83	74.60	75.72	73.67	75.03
Crossmodal-3600 Text-Image Retrieval: Turkish	65.21	64.41	55.12	58.01	56.70	56.82
Crossmodal-3600 Text-Image Retrieval: Ukrainian	68.84	68.01	57.74	59.49	55.32	57.30
Crossmodal-3600 Text-Image Retrieval: Vietnamese	61.84	61.28	54.00	55.01	53.39	53.51
Crossmodal-3600 Text-Image Retrieval: Chinese	64.87	65.56	59.03	61.21	57.33	59.49
Average Western 0-shot Classification	23.54	23.12	21.97	22.18	20.44	21.18
Average Western 10-shot Classification	23.62	24.18	23.59	24.34	23.10	23.99
Average Western 0-shot Retrieval	44.55	45.65	39.83	41.70	39.23	39.44
Average Western Classification	23.60	23.89	23.15	23.75	22.37	23.22
Average Dollar Street Classification	56.89	58.10	56.09	57.46	53.66	54.33
Average GeoDE Classification	40.72	39.94	40.60	39.93	37.01	35.60
Average Income Classification	50.22	51.15	48.09	49.41	49.02	49.23
Average Geographic Classification	5.95	5.97	5.84	5.93	4.83	5.37
Average Demography Classification	24.91	27.14	26.47	25.93	25.50	26.19
Average Multilingual: Low-Resource Lang	87.73	78.82	77.14	72.04	75.01	70.10
Average Multilingual: High-Resource Lang	55.54	56.21	46.75	47.53	45.43	45.75
Average Western	29.19	29.69	27.60	28.54	26.87	27.55
Average Culture Diversity	47.72	47.97	46.72	47.07	44.01	43.29
Average Fairness	23.87	24.56	23.36	23.76	23.01	23.46
Average Multilingual	57.52	56.64	47.23	46.43	45.40	44.61

I. Distribution of Languages

We reuse the 35 languages[8] reported in Crossmodal-3600 benchmark [**?**] for multilingual experiments.

Table 9. Distribution of the 35 languages used in multilingual evaluations.

Language	Type	Pages (%)
Maori	Low-resource	0.001
Telugu	Low-resource	0.036
Swahili	Low-resource	0.046
Filipino	Low-resource	0.111
Bengali	Low-resource	0.113
Hebrew	Low-resource	0.240
Hindi	Low-resource	0.267
Croatian	High-resource	0.284
Norwegian	High-resource	0.290
Finnish	High-resource	0.296
Danish	High-resource	0.370
Hungarian	High-resource	0.378
Ukrainian	High-resource	0.476
Romanian	High-resource	0.489
Greek	High-resource	0.560
Swedish	High-resource	0.660
Czech	High-resource	0.727
Persian	High-resource	0.881
Thai	High-resource	1.167
Dutch	High-resource	1.173
Arabic	High-resource	1.258
Vietnamese	High-resource	1.337
Turkish	High-resource	1.554
Polish	High-resource	1.825
Italian	High-resource	1.964
Korean	High-resource	2.519
Portuguese	High-resource	3.054
Indonesian	High-resource	3.181
French	High-resource	3.354
Chinese	High-resource	3.544
German	High-resource	3.869
Russian	High-resource	6.981
Spanish	High-resource	8.214
Japanese	High-resource	8.752
English	High-resource	35.353
Low-resource All	Low-resource	0.814
High-resource All	High-resource	94.510

[8] "Quechua" is excluded as it is not supported by the language detection method we used.

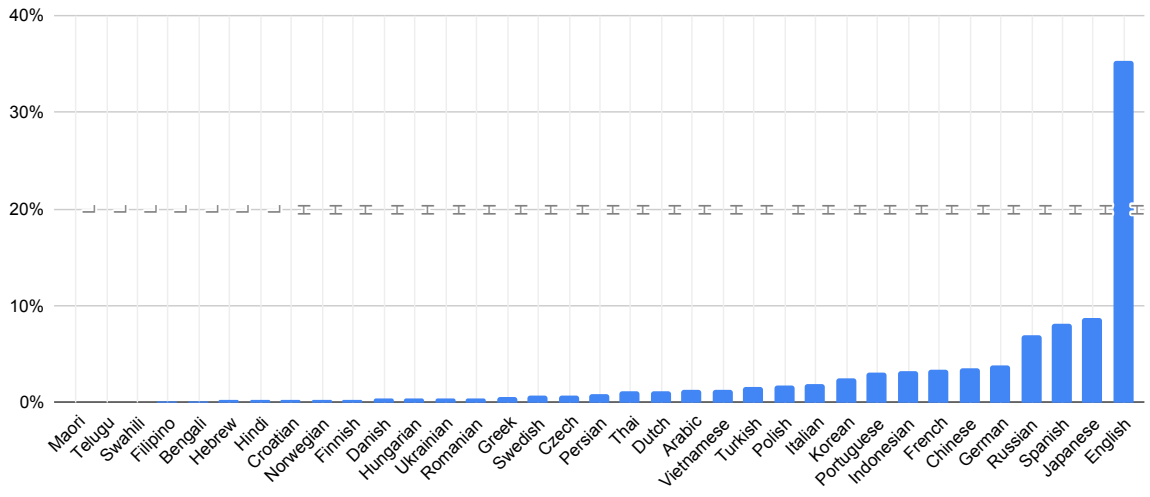


Figure 2. Visualization of the language distribution, where "L" and "H" denote low-resource and high-resource language respectively.

J. Computation-based Scaling Law

Table 10. ImageNet Zero-shot Classification.

Model	Data	3B	10B	33B	100B
B	100B	52.6	56.8	59.9	61
B	10B	52.1	56.2	59.3	60.6
B	1B	51.5	55.6	58	58.8
H	100B	66	70.5	73.4	75.1
H	10B	65.5	70.4	73	74.4
H	1B	64.3	68.5	71	70.4
L	100B	61.6	66.2	69.3	71.5
L	10B	61.7	66.1	69.1	70.3
L	1B	61.2	65.3	67.9	68.8

Table 11. COCO Image-To-Text Zero-shot Retrieval.

Model	Data	3B	10B	33B	100B
B	100B	41.2	44	45.7	46.6
B	10B	41.3	45.1	47	48.4
B	1B	40.2	43	44.7	43.5
H	100B	49.8	52.1	56.1	57.5
H	10B	49.8	53.1	55.8	58
H	1B	47.9	52	52	51.4
L	100B	48.4	50.2	52.5	54.7
L	10B	47	49.8	52.3	52.8
L	1B	46.4	48.9	50.6	50.3

Table 12. DollarStreet Geoloc 10-shot Retrieval.

Model	Data	3B	10B	33B	100B
B	100B	21.7	24.7	26	27.9
B	10B	18.3	20.6	22.1	24.2
B	1B	18.2	20.2	21.4	22.3
H	100B	35.2	42.2	44.6	46.3
H	10B	29.4	37.2	40.9	40.9
H	1B	29	34.1	36.1	35.4
L	100B	30.5	36.1	40.5	41.7
L	10B	28.1	34.5	36.1	35.9
L	1B	25.6	32.7	35.9	36.4

Table 13. Telugu Image-To-Text Zero-shot Retrieval.

Model	Data	3B	10B	33B	100B
B	100B	5.7	10.3	15.6	19.5
B	10B	6.6	9.4	13	12.9
B	1B	2.4	2.1	2	1.9
H	100B	7.5	17.5	27.4	34.7
H	10B	7.9	16.2	23	26.9
H	1B	3.8	4.2	3.4	3.6
L	100B	7.2	16.1	23.8	30.3
L	10B	7.2	14.9	19.4	23.3
L	1B	4.5	4.7	4.2	3.9

K. Absolute Performance Levels

Table 14. Crossmodal-3600 results spanning a wide performance spectrum.

Range of Absolute Error Rate	Task	10B	100B	Error Rate Reduction
0-20	GeoDE: EastAsia	5.01	4.68	0.33
0-20	Caltech 10-shot Classification	6.02	8.93	-2.91
0-20	GeoDE: Africa	6.56	6.4	0.16
0-20	Pet 0-shot Classification	7.47	7.17	0.3
20-40	Crossmodal-3600 Image-Text Retrieval: Turkish	36.56	34.94	1.62
20-40	Crossmodal-3600 Text-Image Retrieval: German	36.56	36.99	-0.43
20-40	Crossmodal-3600 Image-Text Retrieval: Ukrainian	36.94	33.25	3.69
20-40	Crossmodal-3600 Image-Text Retrieval: Dutch	38.06	37.44	0.62
40-60	Crossmodal-3600 Image-Text Retrieval: Arabic	41.6	41.0	0.6
40-60	COCO Image-Text 0-shot Retrieval	42.04	42.48	-0.44
40-60	Crossmodal-3600 Image-Text Retrieval: Japanese	42.22	37.94	4.28
40-60	Crossmodal-3600 Text-Image Retrieval: Russian	43.58	43.08	0.5
60-80	COCO Text-Image 0-shot Retrieval	60.32	59.29	1.03
60-80	Crossmodal-3600 Image-Text Retrieval: Bengali	61.22	56.69	4.53
60-80	Crossmodal-3600 Image-Text Retrieval: Hindi	62.33	60.64	1.69
60-80	Crossmodal-3600 Text-Image Retrieval: Greek	65.45	64.1	1.35
80-100	Crossmodal-3600 Text-Image Retrieval: Telugu	84.44	79.57	4.87
80-100	Crossmodal-3600 Text-Image Retrieval: Swahili	87.47	85.67	1.8
80-100	Crossmodal-3600 Image-Text Retrieval: Quechua	93.06	92.78	0.28
80-100	Crossmodal-3600 Image-Text Retrieval: Maori	98.92	99.17	-0.25