A. More results for SigLiT

In section 4.1, we use the same precomputed embeddings for the images using a ViT-g vision model from [55]. Only resize augmentation is applied, to a fixed 288×288 resolution. We train a standard base size text tower, using the ScalingViT-Adafactor optimizer [54] with $\beta_1 = 0.9$ and $\beta_2 = 0.95$. We use 0.001 learning rate with a linear warmup schedule for the first 200 M examples seen, and then the learning rate is decayed to zero with a cosine learning rate schedule. Weight decay is set to 0.0001 for all the experiments. The text is tokenized by a 32 k vocabulary sentencepiece tokenizer [27] trained on the English C4 dataset [35], and a maximum of 16 text tokens are kept. Table 8 shows results with multiple train examples seen and batch sizes, for both the sigmoid loss and the softmax loss baseline.

For training SigLiT in under a day with 4 chips (Section 4.4), we used the LION optimizer with peak learning rate 1×10^{-4} and weight decay 1×10^{-7}. The learning rate was warmed linearly to the peak in 6.5 k steps, then cosine decayed to zero for the remaining 58.5 k steps.

B. More results for SigLIP

In Table 5, we present more results for SigLIP with multiple train examples seen: 3 billion examples and 9 billion examples respectively.

Batch Size	3 B		9 B	
	sigmoid	softmax	sigmoid	softmax
512	**51.5**	47.7	-	-
1 k	**57.3**	53.2	-	-
2 k	**62.1**	59.3	-	-
4 k	**65.3**	63.8	**68.4**	66.6
8 k	**68.6**	66.6	**70.6**	69.4
32 k	**69.9**	**69.9**	**73.4**	72.9
98 k	69.5	**69.7**	73.0	**73.2**
307 k	-	-	71.6	**72.6**

Table 5: **SigLIP zeor-shot accuracy (%) on the ImageNet benchmark.** Both the sigmoid loss and the softmax loss baseline are presented. Experiments are performed on multiple train examples seen (3 B, 9 B) and train batch sizes (from 512 to 307 k). When trained for 9 B examples, the peak of the sigmoid loss comes earlier at 32 k than the peak of the softmax loss at 98 k. Together with the memory efficient advantage for the sigmoid loss, it allows one to train the best language-image model with much fewer amount of accelerators.

BS	Default	Best	Best LR	Best WD
8 k	70.1	70.1	0.001	0.0001
16 k	70.0	70.0	0.001	0.0001
32 k	68.2	69.0	0.0003	0.00003

Table 6: Default hyperparameters across different batch sizes, perform either the best or close to the best hyperparameter from a sweep. Zero-shot accuracy on ImageNet is reported. BS=batch size, LR=learning rate, WD=weight decay.

C. Robustness of SigLIP results

Hyperparameters for different batch sizes. Sigmoid loss doesn't require tuning hyperparameters for different batch sizes. For example, in both the SigLiP and SigLiT setup, we only used default 0.001 learning rate and 0.0001 weight decay across a wide range of batch sizes (from 512 to 1024k). We further performed a sweep of 9 hyperparameters across 3 batch sizes on the from-scratch SigLIP setup for 3B seen examples: learning rate $\{0.0003, 0.001, 0.003\}$ $\times$ weight decay $\{0.00003, 0.0001, 0.0003\}$ $\times$ batch size $\{8\text{ k}, 16\text{ k}, 32\text{ k}\}$. We observed in Table 6 that the default LR/WD is either the best or close to the best.

Standard deviation. We repeat SigLIP training five times, using the recommended 32k batch size and 3B seen examples. We report the average and std in Table 7. The std of the five runs is very small for both sigmoid and softmax.

Alternative optimizers. We repeat the same experiment with AdamW optimizer five times and got very similar results and std as reported in Table 7. We tested a linear learning rate scheduler instead of the default cosine learning rate scheduler, it achieves 69.9% accuracy.

D. More results for mSigLIP

We present the crossmodal retrieval results on the Crossmodal-3600 dataset, across all the 36 langauges in Figure 8 and Table 9.

Loss	Optimizer	Results (%)
Softmax	ViT-Adafactor	69.9 ± 0.1
Sigmoid	ViT-Adafactor	70.1 ± 0.2
Sigmoid	AdamW	70.3 ± 0.1

Table 7: Mean and standard deviation of five repeated experiments. Zero-shot accuracy on ImageNet is reported.

Batch Size	450 M		900 M		3 B		18 B	
	sigmoid	softmax	sigmoid	softmax	sigmoid	softmax	sigmoid	softmax
512	72.5	69.5	75.0	72.8	77.2	74.6	-	-
1 k	75.5	73.6	77.2	76.0	79.6	77.9	-	-
2 k	77.1	76.3	79.3	78.1	81.3	80.1	82.2	81.2
4 k	79.2	78.3	80.8	79.8	82.4	81.2	83.0	82.0
8 k	80.8	79.7	82.0	81.0	83.1	82.6	83.6	83.1
16 k	81.2	81.2	82.7	82.1	83.8	83.5	84.2	84.1
32 k	81.9	81.4	83.1	82.7	84.2	84.0	84.6	84.4
64 k	81.6	81.6	83.0	82.8	84.3	84.1	84.7	84.4
128 k	80.5	80.0	83.1	83.2	84.2	84.4	84.7	84.6
256 k	72.8	72.2	82.1	81.7	84.3	84.2	84.7	84.6
1024 k	-	-	-	-	-	-	84.7	-

Table 8: **SigLiT zero-shot accuracy (%) on the ImageNet benchmark.** Both the sigmoid loss and the softmax loss baseline are presented. Extensive experiments are performed on multiple train examples seen (450 M, 900 M, 3 B, 18 B) and train batch sizes (from 512 to 1 M).

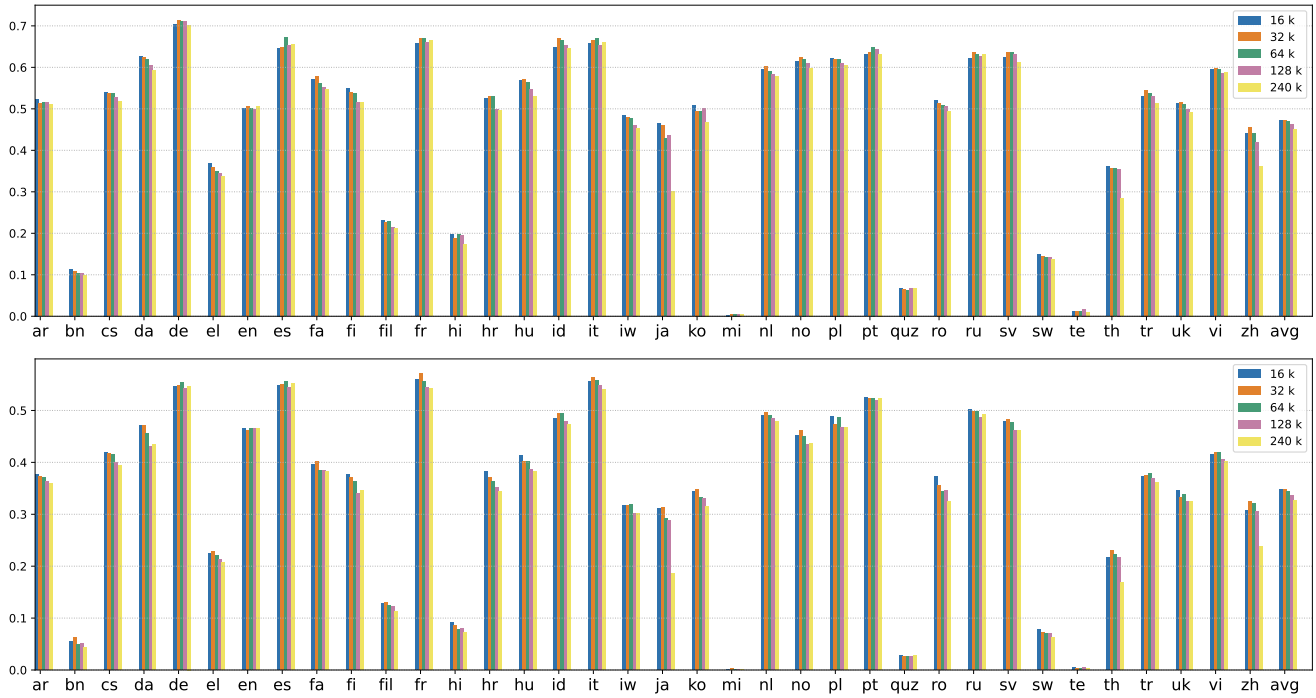

Figure 8: **Zero-shot retrieval on Crossmodal-3600.** Top: Image to text. Bottom: text to image. Colors are batch sizes.

E. Label noise experiments

All models had an M/16 image tower and a M text tower. They were trained from random initialisation for 3.6B examples seen, with a batch size of 16384. A cosine learning rate schedule was used, with an initial linear warmup for 10% of steps up to a peak learning rate of 0.001.

Lang.	Image-to-text					Text-to-image				
	16 k	32 k	64 k	128 k	240 k	16 k	32 k	64 k	128 k	240 k
ar	52.39	51.33	51.50	51.53	51.06	37.61	37.37	37.14	36.31	35.98
bn	11.39	10.78	10.44	10.31	9.89	5.53	6.25	4.94	5.14	4.36
cs	54.08	53.69	53.67	52.75	51.78	41.82	41.64	41.46	39.93	39.38
da	62.72	62.42	62.00	60.42	59.33	47.01	47.01	45.55	43.03	43.47
de	70.33	71.42	71.19	71.11	70.25	54.70	54.83	55.36	54.31	54.71
el	36.94	35.81	35.06	34.53	33.81	22.42	22.78	22.00	21.25	20.79
en	50.11	50.53	50.22	49.94	50.67	46.46	46.21	46.55	46.60	46.60
es	64.69	64.94	67.19	65.31	65.61	54.81	55.04	55.51	54.47	55.24
fa	57.03	57.75	56.06	55.28	54.64	39.61	40.15	38.43	38.36	38.30
fi	54.94	54.08	53.78	51.69	51.67	37.70	37.14	36.38	33.98	34.50
fil	23.22	22.75	22.92	21.39	21.22	12.83	12.93	12.41	12.19	11.34
fr	65.69	66.92	66.97	66.14	66.47	55.92	57.08	55.50	54.39	54.29
hi	19.86	18.81	19.89	19.53	17.36	9.09	8.55	7.86	8.06	7.28
hr	52.67	53.03	52.97	49.92	49.58	38.16	37.09	36.37	35.25	34.33
hu	56.97	57.11	56.33	54.83	53.03	41.37	40.20	40.22	38.55	38.25
id	64.83	67.06	66.56	65.39	64.72	48.53	49.42	49.49	47.82	47.29
it	65.86	66.42	67.11	65.25	66.08	55.52	56.39	55.85	54.75	54.11
iw	48.36	47.86	47.72	46.06	45.25	31.78	31.76	31.89	30.08	30.12
ja	46.42	45.94	42.89	43.72	30.17	31.04	31.32	29.21	28.87	18.50
ko	50.78	49.53	49.44	50.22	46.78	34.44	34.72	33.15	33.06	31.54
mi	0.36	0.42	0.58	0.56	0.42	0.16	0.22	0.19	0.19	0.19
nl	59.56	60.36	58.94	58.31	57.86	48.95	49.55	48.95	48.37	47.88
no	61.36	62.39	61.97	60.89	59.86	45.25	46.21	45.04	43.53	43.71
pl	62.19	62.03	61.97	61.11	60.50	48.80	47.36	48.70	46.79	46.72
pt	63.14	63.61	64.89	64.31	63.25	52.41	52.34	52.30	51.93	52.37
quz	6.78	6.42	6.36	6.64	6.67	2.74	2.57	2.67	2.69	2.79
ro	52.06	51.44	50.97	50.58	49.31	37.20	35.60	34.34	34.52	32.50
ru	62.22	63.64	63.08	62.69	63.08	50.11	49.89	49.71	48.61	49.31
sv	62.33	63.53	63.53	63.06	61.19	47.89	48.18	47.64	46.17	46.16
sw	14.83	14.42	14.31	14.17	13.78	7.81	7.17	7.11	6.94	6.34
te	1.25	1.25	1.19	1.69	1.06	0.40	0.29	0.32	0.47	0.32
th	36.11	35.78	35.64	35.56	28.33	21.58	23.08	22.22	21.62	16.76
tr	53.08	54.50	53.72	52.94	51.25	37.33	37.38	37.81	36.97	36.08
uk	51.42	51.50	51.17	49.86	49.22	34.54	33.21	33.79	32.49	32.39
vi	59.58	59.78	59.53	58.53	58.83	41.43	41.92	41.85	40.63	40.26
zh	44.11	45.67	44.11	41.92	36.08	30.74	32.45	32.05	30.61	23.72
avg	47.21	47.36	47.11	46.34	45.00	34.82	34.87	34.44	33.58	32.72

Table 9: **Image-to-text and text-to-image zero-shot retrieval results on all 36 languages of Crossmodal-3600**, with mSigLIP models trained at different batch sizes for 30 B examples seen.