A Transfer and finetuning details

Few-shot evaluation We use the linear adaptation protocol and evaluation sets from [68, 70], reporting the 10-shot classification accuracy. Specifically, we rely on the pre-logits layer for CLIP* and GAP of the encoder output sequence for our captioning models. For every combination of data set and model we run the 10-shot adaptation three times and report the mean (and standard deviation for key results).

LiT decoder and T5 decoder To train a multi-task decoder from scratch on top of the frozen representation for classification, captioning and VQA, we precisely follow the setup and hyper parameters from [2] except for the data mixing strategy, for which we set to "concat image-question pairs" ([2, Sec. 5.3]). For all encoders, we use the full feature sequence before pooling (including the class token for the evaluation of CLIP). Throughout, we rely on a B-sized transformer decoder [60] with 12 layers.

We also tried fine-tuning the image encoder along with the decoder for both CLIP* and Cap/CapPa models and did not obtain an improvement for any of the models. This is consistent with prior work which did not observe an improvement either for CLIP-style models when fine-tuning with the same decoder-based setup, see [2, Sec. 5.7].

For the T5 decoder we keep all the parameters frozen but reinitialize and train the cross-attention layers. We perform a small sweep around the default learning rate and weight decay of the setup used for training from scratch, while keeping the other hyperparameters unchanged.

Linear and non-linear ImageNet-1k probes (frozen transfer) When performing linear and non-linear probes on ImageNet-1k, we run a wide hyper-parameter optimization sweep for all types of probes (linear, MLP, MAP) in order to get solid, trustworthy conclusions. Specifically, for each image encoder and probe combination, we sweep the full cross-product over the following hyper-parameters: **epochs**: (1, 3, 10, 30, 100); **image cropping**: `resize(256)|random_crop(224)` or `inception_crop(224)`; **learning rate**: (0.001, 0.0003, 0.0001) plus earlier runs showing 0.003 and 0.01 to perform much worse; **weight decay**: (0.0001, lr * 0.1, 0.0); **hidden dimension**: 0 or 1024; **loss**: sigmoid or softmax cross-entropy. The head weights are always initialized to 0 and its bias to -6.9 in the sigmoid case.

For each result shown in Fig. 3, we select the best setting using 1% of the training data that was held-out for this purpose, and report its accuracy on the 50 000 images in the validation set. For completeness, we further compute various ImageNet test-set variants and report full results in Table 9.

Broad MAP-head transfers (fine-grained) We run the same sweep as described above for each individual dataset and model combination, but only using the MAP-head probe. For each dataset, we either use a provided held-out validation set for selecting the best settings, or hold out 20% of the training set if none is provided. Full numeric results are provided in Table 10. Note that we selected and classified the datasets as coarse- or fine-grained solely by looking at the datasets and their classes, before running any single experiment on them, and never revisited this selection.

Fine-tuning on the full ImageNet-1k data set When fine-tuning on the full ImageNet-1k dataset, we attach a fresh MAP head to the pretrained encoder and run full fine-tuning using the AdaFactor optimizer modified for ViTs in [68]. In each setting (B/16, B/16_{384}, L/14_{336}), we run the exact same sweep for CLIP*, CapPa, and Cap models. Notably, our exploration is significantly smaller than that of [14] and unlike for CLIP [50], ImageNet was fully de-duplicated from our pre-training dataset. In all cases, we select the best model on a held-out 2% of the training data and report that model's performance on the 50 000 image validation set without re-training.

For the B/16 models, we sweep over three **learning rates**: (0.0001, 0.00003, 0.00001); two **layer-wise learning-rate decays**: (None, 0.8); 2 **RandAugment** parameters: (10, 15); 3 **Mixup**: (0.0, 0.2, 0.5); and five **Polyak (EMA) averaging factors**: (None, 0.9, 0.999, 0.99999, 0.9999999) with a batch size of 2048 and 100 epochs. The best setting uses learning rate 0.00001, layer-wise decay 0.8, Mixup 0.5 and no Polyak averaging.

For the L/14 models at 336 px resolution, we sweep over three **learning rates**: (0.001, 0.0003, 0.0001), three **layer-wise learning-rate decays**: (None, 0.9, 0.8), and five **Polyak (EMA) averaging factors**:

(None, 0.9, 0.999, 0.99999, 0.9999999). Note that the latter does not require re-training for each setting and hence is cheap. We fix rand-augment to (2, 10), Mixup to 0.2, and training duration to 50 000 steps with batch-size 512, without revisiting these choices. Besides that, we mostly follow [15, 68]. The best setting uses learning rate 0.0001, layer-wise decay 0.9, and Polyak 0.99999 for both models.

B Additional Results

B.1 Probing and LiT tuning results

Table 9 shows the classification accuracy on different ImageNet-1k evaluation sets, when probing the frozen representation with different probes (linear and non-linear), extending the numerical results from Fig. 3.

Table 10 presents transfer results of the frozen representation to fine- and coarse-grained classification tasks (using a MAP head). This complements the results from Fig. 4 (Right).

Table 11 expands Table 4 in the main paper and shows frozen transfer for zero-shot classification and retrieval via LiT [70].

Table 12 presents the performance of frozen Cap/Cap and CLIP* encoders when combined via cross-attention with a *frozen* T5 decoder. This represents the data from Fig. 4 (Left) in the main paper in tabular form.

Table 9: Extended numerical results for Fig. 3, i.e. linear and non-linear ImageNet-1k probes on top of the frozen models. While the *linear* separability of CLIP models is higher, the gap between CLIP* and Cap models is mostly closed when the probe also learns how to pool (*map*).

Model	Head	Top-1	ReaL	-v2	-R(endition)	-A(dvers.)	ObjectNet
CLIP* (8k)	linear	79.8	85.6	69.0	71.9	38.0	49.8
	mlp	80.4	86.1	69.6	74.4	39.3	50.9
	map	82.2	87.3	71.5	72.9	34.3	49.0
	map+mlp	82.2	87.4	71.7	71.8	34.8	48.3
CLIP* (16k)	linear	80.2	85.9	69.2	73.2	40.3	51.3
	mlp	80.9	86.1	70.3	71.4	37.3	49.8
	map	82.6	87.5	72.4	73.9	37.2	50.0
	map+mlp	82.6	87.5	72.1	73.0	36.3	49.3
Cap	linear	77.7	84.1	67.1	68.2	24.1	44.2
	mlp	78.5	84.8	68.0	76.0	27.1	45.6
	map	81.6	87.0	71.3	76.2	32.3	45.8
	map+mlp	81.5	87.0	71.5	76.2	31.4	45.8
CapPa	linear	78.3	84.6	66.5	67.7	22.1	43.7
	mlp	79.4	85.6	68.6	77.2	25.6	46.2
	map	82.0	87.5	72.3	80.9	41.5	50.1
	map+mlp	82.1	87.3	72.0	79.4	39.1	49.5
CLIP* L/14	linear	84.2	88.4	75.0	83.8	59.1	60.2
	mlp	84.6	88.5	74.9	83.3	56.6	58.6
	map	85.9	89.3	76.7	84.9	57.4	58.2
	map+mlp	85.8	89.2	77.0	83.6	56.1	57.8
CapPa L/14	linear	83.0	87.7	73.1	81.1	41.6	53.8
	mlp	84.1	88.7	74.6	87.3	47.0	56.8
	map	85.8	89.3	76.8	86.1	54.5	56.6
	map+mlp	85.8	89.2	76.6	85.5	52.2	56.5

Table 10: Transfer of the frozen representation to fine- and coarse-grained classification tasks (using a MAP head). This extends the numerical results of Figure 4 (Right).

Dataset	Grain	CLIP* (8k)	CLIP* (16k)	Cap	CapPa	CLIP* L/14	CapPa L/14
Dogs [28]	Fine	77.5	77.9	79.6	81.2	85.0	86.0
Flowers [45]	Fine	85.1	89.5	94.0	97.0	96.6	98.9
Birds [61]	Fine	76.9	78.1	76.3	54.2	85.0	86.7
Pets [48]	Fine	91.2	91.2	91.5	94.4	94.4	95.4
Cars [32]	Fine	90.8	91.6	93.4	93.3	94.0	95.8
Food [3]	Fine	91.0	92.1	91.1	91.5	94.9	94.2
RESISC [8]	Coarse	92.5	96.4	90.5	96.1	97.1	96.9
Products [46]	Coarse	88.8	89.0	87.8	88.5	90.7	90.3
SUN397 [71]	Coarse	81.8	82.9	82.0	82.1	85.7	85.2
Caltech [19]	Coarse	93.5	93.1	89.0	86.5	93.8	93.2
STL-10 [12]	Coarse	98.0	98.5	97.7	98.1	99.2	99.1
Cat/Dog [18]	Coarse	99.9	99.9	99.7	99.7	99.8	99.9

Table 11: Frozen transfer for zero-shot classification and retrieval via LiT [70] as a function of the number of number of training examples seen by the text encoder (the vision encoder is pretrained and frozen, and equipped with a MAP head which is trained along with the text encoder). The text encoder mirrors the architecture of the vision encoder. Especially for the larger model, CapPa is competitve with CLIP* with comparable or fewer examples seen. The CLIP numbers are obtained by evaluating the image and text encoders released by [50] in our eval setup. We report these numbers for reference, no LiT tuning is done on top of the CLIP vision encoder. This table complements Table 4 in the main paper.

	ImageNet 0shot				COCO t2i r@1				COCO i2t r@1			
LiT pairs:	0	900M	3B	12B	0	900M	3B	12B	0	900M	3B	12B
Cap	-	65.9	67.8	69.0	-	35.3	37.5	39.1	-	50.3	53.9	54.8
CapPa	-	66.4	68.8	70.2	-	34.3	37.3	38.6	-	49.7	53.9	55.1
CLIP* (8k)	65.6	65.9	67.6	69.0	41.5	36.5	38.2	39.5	56.7	52.0	54.0	56.1
CLIP* (16k)	67.7	66.7	69.0	70.0	43.0	37.0	38.9	40.1	58.2	53.0	55.1	57.0
CLIP			68.3				32.3				52.8	
CapPa L/14	-	74.6	76.4	77.5	-	40.6	43.9	45.4	-	56.6	60.3	62.6
CLIP* L/14	74.8	74.5	75.8	76.6	48.1	42.7	44.7	46.3	63.7	57.7	60.7	62.3
CLIP L/14			75.1				36.5				56.6	

Table 12: Performance of frozen representations trained via image captioning (Cap/CapPa) and contrastive (CLIP*) objective, when combined via cross-attention with a *frozen* T5 decoder. Only the cross-attention weights are updated during the training. See Table 2 for the corresponding models that have the decoder trained from scratch.

	Classification					Captioning		OCR	Question Ans.	
	i1k	sun	food	res	pet	COCO	Flickr	VQA	VQAv2	GQA
Cap	79.0±0.1	81.3±0.1	89.3±0.0	92.4±0.1	92.3±0.3	119.7±0.6	72.2±0.9	57.7±0.0	64.6±0.1	52.1±0.2
CapPa	80.0±0.0	81.2±0.1	89.9±0.0	93.1±0.2	93.2±0.3	118.7±0.5	70.0±0.5	57.8±0.2	63.3±0.3	51.9±0.3
CLIP* (8k)	79.1±0.0	81.5±0.2	89.9±0.0	92.7±0.2	88.5±0.2	110.6±0.5	60.8±1.0	50.3±0.3	57.2±0.4	49.5±0.2
CLIP* (16k)	79.5±0.1	81.7±0.1	90.4±0.1	93.7±0.0	88.6±0.1	110.6±0.6	59.8±0.9	50.2±0.4	56.8±0.3	49.6±0.3

B.2 Scaling properties

Fig. 8 and 9 show the performance of frozen Cap, CapPa, and CLIP* encoders on a variety of tasks as a function of the number of training examples seen and the encoder model size, respectively. Specifically, we evaluate our models on classification, captioning, and VQA when combined with a decoder trained from scratch to solve all those tasks jointly (following [2]), and on 10-shot linear classification based on the pre-logit features.

Tables 13, 14 and 15, 16 show the data from Fig. 8 and 9, respectively, in tabular form. For completeness, in Table 16 we also present the ImageNet zero-shot accuracy (without prompts) of CapPa and CLIP* models obtained with their respective pretrained decoder and encoder. We emphasize that scoring-based zero-shot classification is not the focus of this paper, and we did not optimize the Cap/CapPa models for this.

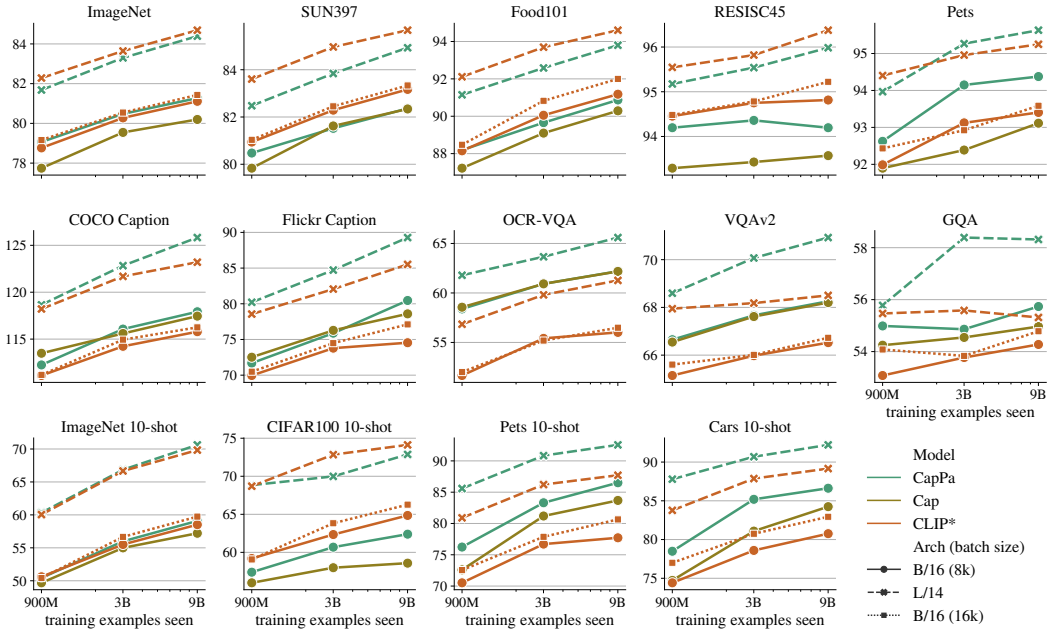

Figure 8: Performance of vision backbones pretrained with captioning (Cap/CapPa) and contrastive objective (CLIP*) as a function of the number of pretraining examples seen (expands the results in Fig. 2). **Top two rows:** Classification, captioning, and VQA performance with a decoder trained from scratch in multi-task fashion (see [2] for details). We use CIDEr for captioning, the VQAv2 weighted accuracy for VQAv2, and exact matching accuracy for all other tasks. **Bottom row:** 10-shot linear classification accuracy on the frozen pre-logit representation.

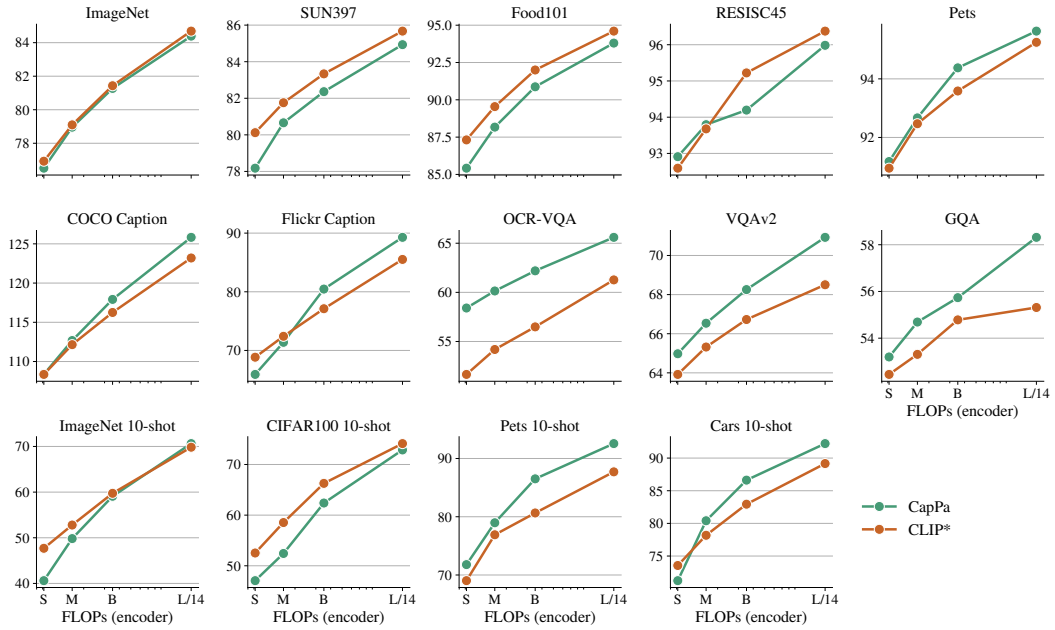

Figure 9: Performance of vision backbones pretrained with captioning (CapPa) and contrastive objective (CLIP*) as a function of the model size/FLOPs (we compare ViT-S/16, M/16, B/16, and L/14; this expands the results in Fig. 2). **Top two rows:** Classification, captioning, and VQA performance with a decoder trained from scratch in multi-task fashion (see [2] for details). We use CIDEr for captioning, the VQAv2 weighted accuracy for VQAv2, and exact matching accuracy for all other tasks. **Bottom row:** 10-shot linear classification accuracy on the frozen pre-logit representation.

Table 13: Data corresponding to Fig. 8 (top two rows) in tabular form. See caption of Fig. 8 for details on the metrics.

			Classification					Captioning		OCR	Question Ans.	
ex. seen	model	arch	i1k	sun	food	res	pet	COCO	Flickr	VQA	VQAv2	GQA
900M	Cap	B/16 (8k)	77.7	79.8	87.2	93.3	91.9	113.5	72.5	58.6	66.5	54.2
	CapPa	B/16 (8k)	79.1	80.5	88.2	94.2	92.6	112.2	71.7	58.4	66.6	55.0
		L/14	81.7	82.5	91.1	95.2	94.0	118.7	80.2	61.8	68.6	55.8
	CLIP*	B/16 (8k)	78.8	80.9	88.2	94.5	92.0	111.1	70.0	51.7	65.1	53.1
		B/16 (16k)	79.2	81.0	88.5	94.5	92.4	111.2	70.5	52.0	65.6	54.1
		L/14	82.3	83.6	92.1	95.5	94.4	118.2	78.6	56.8	67.9	55.5
3B	Cap	B/16 (8k)	79.5	81.6	89.1	93.4	92.4	115.6	76.3	60.9	67.6	54.5
	CapPa	B/16 (8k)	80.5	81.5	89.7	94.4	94.1	116.1	75.9	60.9	67.7	54.9
		L/14	83.3	83.8	92.6	95.5	95.3	122.8	84.7	63.7	70.1	58.4
	CLIP*	B/16 (8k)	80.3	82.3	90.1	94.8	93.1	114.2	73.8	55.4	66.0	53.8
		B/16 (16k)	80.5	82.5	90.8	94.8	92.9	114.9	74.5	55.2	66.0	53.8
		L/14	83.6	85.0	93.7	95.8	95.0	121.7	82.1	59.8	68.2	55.6
9B	Cap	B/16 (8k)	80.2	82.3	90.3	93.6	93.1	117.5	78.6	62.2	68.2	55.0
	CapPa	B/16 (8k)	81.3	82.4	90.9	94.2	94.4	117.9	80.5	62.2	68.3	55.7
		L/14	84.4	84.9	93.8	96.0	95.6	125.8	89.3	65.6	70.9	58.3
	CLIP*	B/16 (8k)	81.1	83.2	91.2	94.8	93.4	115.8	74.5	56.0	66.5	54.3
		B/16 (16k)	81.4	83.3	92.0	95.2	93.6	116.3	77.1	56.5	66.7	54.8
		L/14	84.7	85.7	94.6	96.4	95.2	123.2	85.5	61.3	68.5	55.3

Table 14: Data corresponding to Fig. 8 (bottom row) in tabular form. 10-shot linear classification accuracy on the frozen pre-logit features.

ex. seen	model	arch	ImageNet	CIFAR100	Pets	Cars
900M	Cap	B/16 (8k)	49.7	56.0	72.6	74.7
	CapPa	B/16 (8k)	50.4	57.4	76.2	78.5
		L/14	60.3	68.8	85.6	87.8
	CLIP*	B/16 (8k)	50.6	59.2	70.5	74.4
		B/16 (16k)	50.4	59.1	72.6	77.0
		L/14	60.0	68.7	80.9	83.8
3B	Cap	B/16 (8k)	55.0	58.0	81.2	81.1
	CapPa	B/16 (8k)	56.0	60.7	83.3	85.2
		L/14	66.9	70.0	90.8	90.7
	CLIP*	B/16 (8k)	55.5	62.3	76.7	78.6
		B/16 (16k)	56.7	63.8	77.9	80.7
		L/14	66.7	72.8	86.2	87.9
9B	Cap	B/16 (8k)	57.2	58.6	83.7	84.2
	CapPa	B/16 (8k)	59.1	62.4	86.5	86.6
		L/14	70.6	72.9	92.6	92.2
	CLIP*	B/16 (8k)	58.5	64.9	77.7	80.8
		B/16 (16k)	59.7	66.3	80.6	82.9
		L/14	69.8	74.1	87.7	89.2

Table 15: Data corresponding to Fig. 9 (top two rows) in tabular form. See caption of Fig. 9 for details on the metrics

			Classification					Captioning		OCR	Question Ans.	
arch	FLOPs	model	i1k	sun	food	res	pet	COCO	Flickr	VQA	VQAv2	GQA
S/16	9.2G	CapPa	76.5	78.2	85.4	92.9	91.2	108.4	65.9	58.4	65.0	53.2
		CLIP*	76.9	80.1	87.3	92.6	91.0	108.4	68.8	51.7	63.9	52.4
M/16	16.0G	CapPa	79.0	80.7	88.2	93.8	92.7	112.7	71.4	60.1	66.5	54.7
		CLIP*	79.1	81.8	89.5	93.7	92.5	112.2	72.4	54.2	65.3	53.3
B/16	35.1G	CapPa	81.3	82.4	90.9	94.2	94.4	117.9	80.5	62.2	68.3	55.7
		CLIP*	81.4	83.3	92.0	95.2	93.6	116.3	77.1	56.5	66.7	54.8
L/14	161.8G	CapPa	84.4	84.9	93.8	96.0	95.6	125.8	89.3	65.6	70.9	58.3
		CLIP*	84.7	85.7	94.6	96.4	95.2	123.2	85.5	61.3	68.5	55.3

Table 16: Data corresponding to Fig. 9 (bottom row) in tabular form. 10-shot linear classification accuracy on the frozen pre-logit features. We also show the ImageNet zero-shot classification accuracy (without prompts) when using the pretrained text encoder (CLIP*) or text decoder with scoring (CapPa) for reference (last column).

arch	FLOPs	model	ImageNet	CIFAR100	Pets	Cars	ImageNet zs.
S/16	9.2G	CapPa	40.6	47.1	71.8	71.2	35.1
		CLIP*	47.7	52.5	69.0	73.6	52.8
M/16	16.0G	CapPa	49.8	52.4	79.0	80.4	43.0
		CLIP*	52.8	58.5	76.9	78.2	58.7
B/16	35.1G	CapPa	59.1	62.4	86.5	86.6	52.7
		CLIP*	59.7	66.3	80.6	82.9	64.1
L/14	161.8G	CapPa	70.6	72.9	92.6	92.2	63.8
		CLIP*	69.8	74.1	87.7	89.2	71.2

B.3 Attribution, relation, ordering

Table 17 shows extended results for different models on the ARO benchmark [67] (see Table 6 in the main paper). In addition to the clear superiority of Cap/CapPa over CLIP* models discussed in the main paper, it can be observed that increasing the model capacity form B/16 to L/14 leads to an overall improvement for CapPa, while this is not the case for CLIP*.

Table 17: Results on the Attribute, Relation and Order (ARO) benchmark [67]. Cap and CapPa models clearly outperform all CLIP* and CLIP variants across all data sets, even when training the model to be sensitive to word ordering and attribution as in NegCLIP [67]. Values for "ARO Best" are taken from [67]. "Blind dec." corresponds to Cap without vision encoder, i.e. the vision encoder features fed to the decoder are replaced with all zeros.

	Arch	VG Attribution	VG Relation	Flickr Order	COCO Order
Blind dec.	-	83.7	86.2	98.8	98.7
Cap	B/16	88.9	86.6	99.1	99.0
CapPa	B/16	85.7	86.7	99.2	98.8
CapPa	L/14	89.3	86.0	99.3	99.0
CLIP* (8k)	B/16	55.4	39.8	43.7	32.8
CLIP* (16k)	B/16	53.2	39.7	45.5	37.0
CLIP* (16k)	L/14	57.8	35.9	40.2	31.5
CLIP	B/32	63.2	59.1	59.4	47.3
CLIP	B/16	62.7	58.7	57.9	49.5
ARO Best	-	88.0	73.0	60.0	46.0
NegCLIP	B/32	71.0	81.0	91.0	86.0

B.4 SugarCrepe

We provide the full breakdown of results across all our models on SugarCrepe in Table 18. The numbers for OpenCLIP are taken from [21] and represent the best, largest contrastive model that was benchmarked on SugarCrepe to date. Even the small ViT-B/16 Cap model significantly outperforms it on all but the "Replace Object" task, which is a task that matches contrastive's "bag of word"-style of learning well.

Table 18: Full results on the SugarCrepe [21] benchmark suite.

Training	Arch	Replace			Swap		Add	
		Object	Attribute	Relation	Object	Attribute	Object	Attribute
Cap	B/16	91.10	88.32	85.21	79.27	**88.74**	98.59	99.28
CapPa	B/16	89.95	88.71	84.35	80.49	85.74	98.84	**99.42**
CapPa	L/14	92.01	**90.10**	**87.34**	**82.11**	88.44	**98.93**	**99.42**
CLIP* (8k)	B/16	93.70	82.36	66.29	61.79	67.12	83.46	76.01
CLIP* (16k)	B/16	94.07	84.64	67.14	60.98	65.47	86.37	77.46
CLIP* (16k)	L/14	95.70	84.26	69.06	65.04	68.02	86.76	78.32
OpenCLIP	G/14	**96.67**	88.07	74.75	62.20	74.92	92.19	84.54

We further show qualitative examples in Tables 22–24. The examples are manually picked to be representative (we show wins and losses), while avoiding uninteresting (i.e. seemingly random), too cluttered, or too verbose examples. Thus, the examples are cherry-picked to be presentable, but are meant to be representative. All images are from the COCO validation set.

Each image comes with a positive and a (hard) negative caption, and a model's prediction is deemed correct when it scores the positive caption higher than the negative one. For the CapPa model, we score each caption using the log-likelihood, meaning negative numbers closer to zero correspond to a higher score (i.e. a score of -20 means the caption fits the image more than a score of -110). For the CLIP model, we score each caption using the dot-product of normalized embedding similarity as is usual, but we multiply the resulting score by 100 for readability.

Table 19: Impact of decoder architecture design choices in Cap on 10-shot linear classification accuracy: **Left:** Effect of sharing the embedding between decoder input and output and removing biases from decoder layers. **Right:** Effect of the number of decoder layers.

share emb.	dec. bias	ImageNet	CIFAR100	Pets	Cars
yes	no	47.8	55.8	71.5	71.7
no	no	49.7	56.0	72.6	74.7
yes	yes	48.3	54.6	74.4	70.2
no	yes	49.3	56.6	72.7	71.9

dec. layers	ImageNet	CIFAR100	Pets	Cars
3	48.7	53.7	73.5	73.7
6	49.7	56.0	72.6	74.7
12	48.7	54.8	74.4	73.8

We noticed that for the **Add** scenarios, where CapPa performs almost perfectly, the only losses are due to typos in the positive caption ("toliet' instead of "toilet" and "bridge" instead of "bride"), so we also provide the score for the corrected caption in the *Pos (fixed)*, which confirms the typos are the reason for the model failure.

B.5 Ablations: Decoder architecture

While following the original transformer decoder architecture [60] closely, we investigate several modifications that have become common in the literature [52, 11]. Specifically, we ablate the effect of removing biases in decoder layers, as well as sharing the decoder input and output embeddings. Table 19 (left) shows that not sharing the embeddings leads to overall better 10-shot accuracy than sharing them, and additionally removing the decoder biases does not hurt. Furthermore, we observed significantly improved stability across encoder architectures, scales and training schedules when removing the decoder biases.

Table 19 (right) reveals that the overall best 10-shot classification accuracy is obtained when using a 6 layer decoder. This decoder depth also leads to a total parameter count comparable to the corresponding CLIP* model (Table 1).

B.6 Further Ablations

Table 20 compares the performance of the CapPa and CLIP* vision encoders with a ViT-B/16 pretrained in supervised fashion on ImageNet-21k when combined with transformer decoder.

Table 21 represents the data from Fig. 6 in tabular form.

Table 20: Comparison of CapPa and CLIP* with a ViT-B/16 pretrained in supervised fashion on ImageNet-21k (we use the checkpoint from Steiner et al. 2021) when combined with a transformer decoder trained from scratch for classification, captioning, and VQA [2]. CLIP* and CapPa outperform the model pretrained in supervised fashion.

	Classification					Captioning		OCR	Question Ans.	
	i1k	sun	food	res	pet	COCO	Flickr	VQA	VQAv2	GQA
ViT-B/16 (i21k)	73.1±0.1	72.8±0.2	81.2±0.2	86.2±0.1	85.6±0.3	95.0±0.4	52.3±0.1	39.1±0.1	57.6±0.3	50.1±0.1
CapPa	81.3±0.1	82.4±0.1	90.9±0.1	94.2±0.2	94.4±0.1	117.9±0.6	80.5±0.2	62.2±0.0	68.3±0.1	55.7±0.2
CLIP*	81.4±0.1	83.3±0.1	92.0±0.1	95.2±0.2	93.6±0.2	116.3±0.7	77.1±0.7	56.5±0.1	66.7±0.1	54.8±0.6

C Societal impact

Our models fit in the broader context of large scale vision-language pretraining and as such share many of the benefits and issues of related models such as [50, 26, 66, 40, 63]: They produce versatile vision models which obtain strong performance on natural images, on OCR-related tasks, and also when combined with a generative language decoder. These capabilities enable many useful applications (e.g. assistive technologies, medical imaging), but also potentially harmful ones (e.g. surveillance). We generally recommend either employing the CapPa vision encoder with a new, task-specific prediction head, or using the pretrained decoder for scoring only. We do not recommend the pretrained decoder for downstream image captioning applications without further refinement, as

Table 21: Ablation results representing Fig. 6 in tabular form. **Left:** 10-shot linear classification accuracy based on the frozen encoder representation as a function of the fraction of training batches for which parallel prediction is performed in CapPa. **Right:** 10-shot linear classification accuracy and zero-shot classification accuracy as a function of the vision encoder architecture.

fraction	INet	C100	Pets	Cars
0%	49.7	56.0	72.6	74.7
25%	46.7	52.9	71.9	72.8
50%	49.8	57.8	76.8	76.9
75%	50.4	57.4	76.2	78.5
90%	49.0	59.5	73.1	79.0

arch	model	INet	C100	Pets	Cars	INet zs.
R50	CLIP* (8k)	39.8	33.5	49.2	60.9	43.6
	Cap	37.8	33.3	48.6	52.4	28.5
ViT-B/32	CLIP* (8k)	44.1	57.7	64.7	68.1	48.3
	Cap	41.0	53.7	64.0	58.7	35.4
ViT-B/16	CLIP* (8k)	50.6	59.2	70.5	74.4	52.2
	Cap	49.7	56.0	72.6	74.7	43.8

it is trained on a large number of alt-texts from the web. Harmful biases should be carefully assessed in the context of the concrete downstream application and prediction head used. For example, when combining the encoder with a (potentially pretrained) decoder for captioning or VQA, an assessment of hallucinations, attribute binding issues and stereotypical attribution should be done.

Table 22: Representative examples of CapPa L/14 wins (first three) and losses (fourth) over CLIP L/14 on the different **replace** categories of the SugarCrepe hard negatives benchmark suite. See text for example selection criteria. Higher score means better: for CapPa this is the log-likelihood, so closer to 0 is better, while for CLIP this is the matching score (multiplied by 100) so closer to 100 is better.

		CapPa wins over CLIP			**CLIP wins over CapPa**
Replace Object	Positive	CapPa:**-28.6** CLIP:**23.6** *Street signs on the corner of Gladys and Detroit*	CapPa:**-45.4** CLIP:**13.2** *A run down building with two planters outside the door*	CapPa:**-42.4** CLIP:**16.8** *A brown bird has a small yellow head.*	CapPa:**-54.6** CLIP:**13.7** *The model toys are positioned on the table.*
	Negative	CapPa:**-53.8** CLIP:**13.9** *Street signs on the corner of Gladys and Chicago.*	CapPa:**-59.2** CLIP:**14.8** *A run down building with a statue outside the door.*	CapPa:**-46.2** CLIP:**18.1** *A brown bird has a small yellow beak.*	CapPa:**-51.7** CLIP:**8.8** *The books are positioned on the table.*
					
Replace Attribute	Positive	CapPa:**-45.0** CLIP:**14.7** *A plate of food with a fried egg and colorful vegetables.*	CapPa:**-47.3** CLIP:**13.1** *A bunch of different foods on display on a counter.*	CapPa:**-17.3** CLIP:**15.7** *A large black truck in a parking lot.*	CapPa:**-115.9** CLIP:**14.1** *Two large trucks are travelling along a tree-lined roadway.*
	Negative	CapPa:**-59.5** CLIP:**15.1** *A plate of food with a fried egg and monochromatic vegetables.*	CapPa:**-53.8** CLIP:**14.8** *A bunch of similar foods on display on a counter.*	CapPa:**-38.3** CLIP:**16.1** *A small black truck in a parking lot.*	CapPa:**-61.7** CLIP:**12.5** *Two large trucks are travelling along a deserted roadway.*
					
Replace Relation	Positive	CapPa:**-20.0** CLIP:**18.5** *A fire hydrant in a grassy field next to a bush*	CapPa:**-48.2** CLIP:**18.6** *A cell phone on top of a calculator near a computer keyboard.*	CapPa:**-29.1** CLIP:**24.4** *A red fire hydrant on a city sidewalk.*	CapPa:**-55.6** CLIP:**17.6** *A train driving over a small bridge on a green hillside.*
	Negative	CapPa:**-56.1** CLIP:**21.5** *A fire hydrant in a grassy field far from a bush.*	CapPa:**-56.1** CLIP:**19.2** *A cell phone underneath a calculator near a computer keyboard.*	CapPa:**-35.6** CLIP:**25.6** *A red fire hydrant beside a city sidewalk.*	CapPa:**-54.7** CLIP:**17.4** *A train passing under a small bridge on a green hillside.*
					

Table 23: Representative examples of CapPa L/14 wins (first three) and losses (fourth) over CLIP L/14 on the different **add** categories of the SugarCrepe hard negatives benchmark suite. See text for example selection criteria. Higher score means better: for CapPa this is the log-likelihood, so closer to 0 is better, while for CLIP this is the matching score (multiplied by 100) so closer to 100 is better.

		CapPa wins over CLIP			**CLIP wins over CapPa**
Add Object	Positive	CapPa:**-18.2** CLIP:**13.7** *A bathroom with a mirror and a sink.*	CapPa:**-30.2** CLIP:**14.3** *A two layered cake sits on a table top*	CapPa:**-27.3** CLIP:**14.0** *an image of a plate of food with meat and veggies*	CapPa:**-60.3** CLIP:**22.5** *A bridge and groom cutting their wedding cake that has fruit on top.*
	Negative	CapPa:**-150.3** CLIP:**13.7** *A bathroom with a mirror, sink, and shower.*	CapPa:**-64.9** CLIP:**15.5** *A two layered cake sits on a table top next to a vase of flowers.*	CapPa:**-148.6** CLIP:**14.6** *An image of a plate of food with meat, fruit, and veggies.*	CapPa:**-53.4** CLIP:**21.6** *A bride and groom cutting their wedding cake that has flowers and fruit on top.*
					
	Pos (fixed)				CapPa:**-46.1** CLIP:**22.7** *A bride and groom cutting their wedding cake that has fruit on top.*
Add Attribute	Positive	CapPa:**-43.8** CLIP:**21.0** *A little girl smiling for the camera with an umbrella behind her.*	CapPa:**-65.4** CLIP:**16.1** *A clock fastened to a brick store front reads 10 after 10*	CapPa:**-49.9** CLIP:**17.6** *A person frying some kind of food on a stove.*	CapPa:**-62.1** CLIP:**20.4** *There is a stuffed animal sitting on the toliet.*
	Negative	CapPa:**-121.5** CLIP:**21.0** *A little girl smiling for the camera with a polka-dotted umbrella behind her.*	CapPa:**-90.7** CLIP:**17.0** *A clock fastened to a lush brick store front reads 10 after 10.*	CapPa:**-115.3** CLIP:**19.5** *A person frying some curry-spiced food on a stove.*	CapPa:**-49.8** CLIP:**19.1** *There is a stuffed animal sitting on the decorated toilet.*
					
	Pos (fixed)				CapPa:**-36.8** CLIP:**21.3** *There is a stuffed animal sitting on the toilet.*

Table 24: Representative examples of CapPa L/14 wins (first three) and losses (fourth) over CLIP L/14 on the different **swap** categories of the SugarCrepe hard negatives benchmark suite. See text for example selection criteria. Higher score means better: for CapPa this is the log-likelihood, so closer to 0 is better, while for CLIP this is the matching score (multiplied by 100) so closer to 100 is better.

		CapPa wins over CLIP			**CLIP wins over CapPa**
Swap Object	Positive	CapPa:**-33.5** CLIP:**22.5** *a bright kitchen with tulips on the table and plants by the window*	CapPa:**-54.9** CLIP:**22.1** *A person cutting a pizza next to a salad and bottles of wine on wooden table.*	CapPa:**-38.1** CLIP:**15.6** *A close up of a sandwich with a drink in the back.*	CapPa:**-111.4** CLIP:**20.5** *Statues on the second floor of a building, sitting below a clock.*
	Negative	CapPa:**-56.8** CLIP:**22.8** *A bright kitchen with plants on the table and tulips by the window.*	CapPa:**-57.5** CLIP:**22.5** *A person cutting a salad next to a pizza and bottles of wine on wooden table.*	CapPa:**-45.6** CLIP:**16.2** *A close up of a drink with a sandwich in the back.*	CapPa:**-110.8** CLIP:**19.4** *A clock on the second floor of a building, sitting below statues.*
					
Swap Attribute	Positive	CapPa:**-32.6** CLIP:**15.4** *a white cake is by a bunch of flowers*	CapPa:**-45.9** CLIP:**19.4** *A blue tennis racket has a yellow tennis ball on it.*	CapPa:**-28.4** CLIP:**16.3** *a black bike rests against a brown bed*	CapPa:**-108.6** CLIP:**19.3** *All of the cows are poking their heads out, eating some hay.*
	Negative	CapPa:**-64.1** CLIP:**17.3** *A bunch of cakes are by a white flower.*	CapPa:**-54.9** CLIP:**19.5** *A yellow tennis racket has a blue tennis ball on it.*	CapPa:**-52.8** CLIP:**16.9** *a brown bike rests against a black bed.*	CapPa:**-107.1** CLIP:**18.2** *Some cows are poking their heads out, eating all of the hay.*