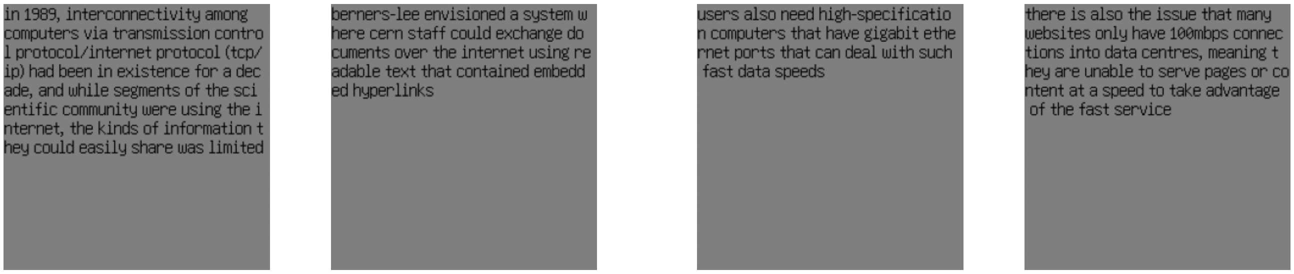
CLIPPO: Image-and-Language Understanding from Pixels Only—Supplementary Material

A. Example input images

Fig. 6 shows two examples of consecutive sentences from the C4 corpus, rendered using our Unifont renderer. The alt-texts for contrastive pretraining are rendered in the same way.

Fig. 7 shows example images from the VQAv2 training set [25] with rendered text in the format we use to adapt CLIPPO (and our baselines) to VQA. The question is rendered with line height of 16px (which is identical to the line height used during pretraining) and the image is resized as to fill the remaining space (with a total image size of 224×224 or 384×384).



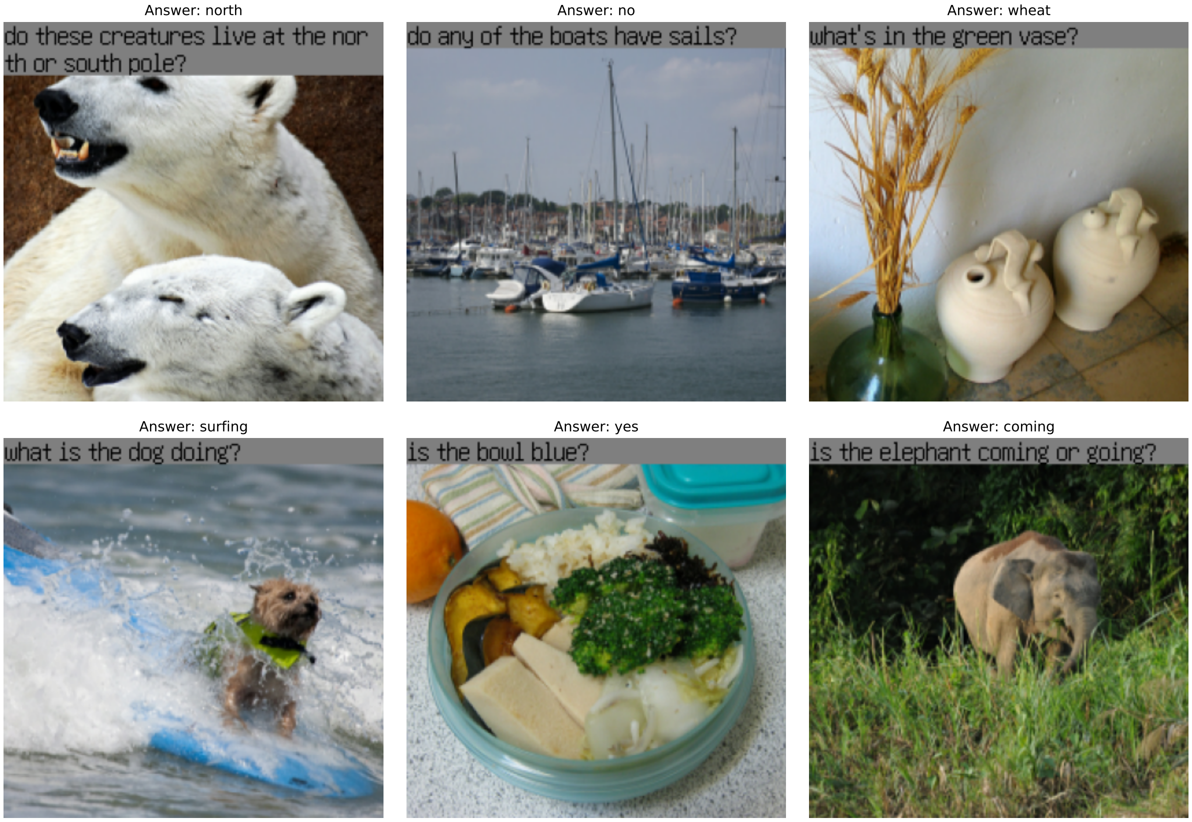
in 1989, interconnectivity among
computers via transmission contro
l protocol/internet protocol (tcp/
ip) had been in existence for a dec
ade, and while segments of the sci
entific community were using the i
nternet, the kinds of information t
hey could easily share was limited

berners-lee envisioned a system w
here cern staff could exchange do
cuments over the internet using re
adable text that contained embedd
ed hyperlinks

users also need high-specificatio
n computers that have gigabit ethe
rnet ports that can deal with such
 fast data speeds

there is also the issue that many
websites only have 100mbps connec
tions into data centres, meaning t
hey are unable to serve pages or co
ntent at a speed to take advantage
 of the fast service

Figure 6. Two examples for rendered consecutive sentences from C4 (image size 224×224). The rendering is identical for alt-texts.



Answer: north

do these creatures live at the nor
th or south pole?

Answer: no

do any of the boats have sails?

Answer: wheat

what's in the green vase?

Answer: surfing

what is the dog doing?

Answer: yes

is the bowl blue?

Answer: coming

is the elephant coming or going?

Figure 7. Example training images with rendered questions (black letters on gray background) from the VQAv2 dataset (image size 224×224). After fine-tuning CLIPPO on VQAv2 it can process images and question jointly in this form. Note that the answers (on white background) are not part of the image.

B. Training details

We rely on a single training setup for all our baselines and visual text models. This setup was tuned to to produce good results for standard image/alt-text contrastive training as in [56] (using exactly the same loss function as [56], following the pseudocode in [56, Fig. 3]) and we found that it readily transfers to 1T-CLIP and CLIPPO (including variants with text/text co-training).

Our default architecture is a ViT-B/16 [16] and we perform a subset of experiments with a ViT-L/16 architecture to study the effect of scale (we equip both models a MAP head [40] to pool embeddings). In all cases, the representation dimension used for the contrastive loss is 768. We set the batch size to 10,240 and train the main models for 250k steps, using a minimum 100k training steps for ablations. For models co-trained with a certain percentage of text/text data, we scale the number of iterations such that the number of image/alt-text pairs seen matches the number of iterations of the corresponding model without text/text data (e.g. when 50% of the data is text/text pairs we increase the number of iterations from 250k to 500k). The contrastive loss is computed across the full batch.

We use the Adafactor optimizer [66] with a learning rate of 10^{-3}, parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and decoupled weight decay with weight 10^{-4}. Gradients are clipped to a norm of 1. We initialize the learned temperature parameter in the contrastive loss with a value of 10. We employ a reciprocal square root schedule with 10k steps linear warmup and 10k steps linear cooldown. This schedule has the advantage that it allows resuming training before cooldown to train a subset of models for more steps (unlike e.g. a cosine schedule which is scaled to a predefined target number of steps). Apart from the learning rate, the training setup is static for all models except for the CLIPPO L/16 models co-trained with 25% and 50% C4 data. To save compute, we do not co-train these with C4 from scratch, but we take the checkpoints pretrained for 150k steps without C4 and continue training these with mixed batches for 350k more steps (i.e. we deviate from the rule described above to adapt the number of training steps with mixed batches).

Following the above schedules and hyperparamteres we further train CLIPPO models on a mix of YFCC-100M [71] and C4 (some initialized with an ImageNet21k-pretrained checkpoint), and release them publicly [4]. We use the full YFCC-100M data set, sampling one of the available title/description/tag annotations at random for each each example. We drop non-descriptive annotations (e.g. descriptions consisting of digits only) following the filtering procedure outlined in [85, Appendix E]. Results for these models can be found in Tables 6 and 8.

For all CLIPPO and 1T-CLIP experiments with ViT B/16-scale architecture (i.e. the majority of experiments) we train on 64 Cloud TPUv2 chips. For larger models (CLIP* B/16 and CLIPPO/1T-CLIP L/16) we use 64 Cloud TPUv3 or Cloud TPUv4 chips to accommodate the increased memory requirements.

B.1. Fine-tuning details for VQA tasks

Our fine-tuning protocol is inspired by the one described in [17, Sec. 4.1.1]. After replacing the last linear layer of the model with a randomly initialized one with an appropriate number of outputs, we fine-tune for 8,000 steps on a combination of the VQAv2 training set and 90% of the validation set, using the remaining 10% for learning rate selection (recall that we report results on the test-dev set). We rely on SGD with momentum 0.9 and a cosine schedule with 800 linear warmup steps, selecting the learning rate for each model from $\{0.03, 0.1, 0.2\}$. The learning rate for the parameters of the freshly initialized head is multiplied by a factor of 10. Gradients are clipped to a norm of 1.

As it is common in the VQA literature to perform evaluation at high resolution, we also evaluate our models on 384×384 images (rendering the question at the top of the image following the same strategy as for 224×224 images, see Appendix A). To adapt the models to this resolution before fine-tuning, we train a subset of models for 30k iterations at a resolution of 384px, starting from the corresponding 224px checkpoints stored right before cooldown.

[4] https://github.com/google-research/big_vision/blob/main/big_vision/configs/proj/clippo/README.md

C. Additional results

C.1. Results on LAION-400M

In Tables 4 and 5 show results on vision and vision-language benchmarks as well as the GLUE benchmark, for the most important CLIPPO and 1T-CLIP models trained on the publicly available LAION-400M dataset [64] (see Appendix C.2 for these results in the context of all other results in the paper). We also show the corresponding models trained on WebLI.

For all the benchmarks/metrics, models trained on LAION-400M exhibit the same ranking as the models trained on WebLI. The ImageNet-1k zero shot and 10-shot results are a few percentage points lower for the models trained on LAION-400M compared to the models trained on WebLI, but the retrieval results on MS-COCO and Flickr30k are consistently a few points better. The GLUE average scores seem largely independent of whether WebLI or LAION-400M is used as a pretraining data set, except for 1T-CLIP, where WebLI-based pretraining leads to a better GLUE score.

	#param.	training dataset	I1k 10s.	I1k 0s.	C I→T	C T→I	F I→T	F T→I
1T-CLIP	118M	WebLI	50.9	60.1	46.2	28.2	76.1	55.2
CLIPPO	93M	WebLI	49.7	58.0	44.9	29.0	73.1	55.4
CLIPPO	93M	WebLI + 25%C4	49.4	55.4	40.2	25.3	69.0	50.5
CLIPPO	93M	WebLI + 50%C4	45.6	51.1	34.3	21.7	61.7	43.2
1T-CLIP	118M	LAION	46.0	54.3	49.0	31.5	77.5	59.7
CLIPPO	93M	LAION	45.3	53.6	46.7	30.3	76.9	58.9
CLIPPO	93M	LAION + 25%C4	44.9	50.6	41.8	27.2	71.1	53.7
CLIPPO	93M	LAION + 50%C4	41.4	46.0	38.2	24.3	66.3	49.0

Table 4. Vision and vision-language cross-modal results obtained when training on LAION-400M [64], along with the corresponding models trained on WebLI. We report ImageNet-1k 10-shot linear transfer validation accuracy (I1k 10s.), ImageNet-1k zero-shot transfer validation accuracy (I1k 0s.), image-to-text and text-to-image retrieval recall@1 on MS-COCO (C I→T and C T→I) and on Flickr30k (F T→I and F I→T). All models have a ViT B/16 architecture (with separate text embedding for 1T-CLIP) and are trained for 100k iterations (with adapted number of steps for models co-trained with C4, see Sec. B).

	training dataset	MNLI-M/MM	QQP	QNLI	SST-2	COLA	STS-B	MRPC	RTE	avg
1T-CLIP text enc.	WebLI	71.6 / 71.5	83.5	80.5	85.0	0.0	74.1	82.8	54.2	67.0
CLIPPO	WebLI	72.2 / 72.5	84.0	81.2	86.7	0.0	81.0	84.0	57.8	68.8
CLIPPO	WebLI + 25%C4	77.0 / 76.7	85.4	82.8	90.9	20.1	83.1	83.6	54.5	72.7
CLIPPO	WebLI + 50%C4	78.8 / 78.3	86.0	84.8	92.0	34.4	83.1	84.2	58.8	75.6
1T-CLIP text enc.	LAION	72.2 / 72.8	84.1	79.8	86.9	0.0	38.0	81.4	54.2	63.3
CLIPPO	LAION	73.2 / 73.5	84.2	80.9	86.5	0.0	75.3	82.2	53.8	67.7
CLIPPO	LAION + 25%C4	77.0 / 77.0	85.5	83.3	91.1	22.0	83.3	84.6	57.0	73.4
CLIPPO	LAION + 50%C4	78.8 / 78.7	86.1	84.3	92.2	38.3	83.7	83.9	55.2	75.7

Table 5. Results for the GLUE benchmark (dev set) when training on LAION-400M [64], along with the corresponding models trained on WebLI. The metric is accuracy except for the performance on QQP and MRPC, which is measured using the F_1 score, CoLA which uses Matthew's correlation, and STS-B which evaluated based on Spearman's correlation coefficient. "avg" corresponds to the average across all metrics. All models have a ViT B/16 architecture (with separate text embedding for 1T-CLIP) trained for 100k iterations (with adapted number of steps for models co-trained with C4, see Sec. B).

C.2. All image, vision-language, and language understanding results

Image classification and retrieval Table 6 shows the full set of image classification and image/text retrieval results, including models trained for 100k and 250k steps.

In addition to the results presented in the main paper, we also show results for pretraining with multilingual alt-texts. In this context, CLIP*, 1T-CLIP, and CLIPPO all obtain a somewhat worse scores on these English-based metrics, but perform much better when evaluated on multilingual image/text retrieval.

We also show results CLIPPO models that were initialized with a ViT trained for image classification. We observe that this improves ImageNet-1k-based classification metrics, but cannot prevent the image and image/text metrics from degrading when co-training with C4 data.

	lan.	#param.	training dataset	steps	I1k 10s.	I1k 0s.	C I→T	C T→I	F I→T	F T→I
CLIP*	EN	203M	WebLI	100k	52.9	62.8	47.2	29.7	76.8	57.2
1T-CLIP	EN	118M	WebLI	100k	50.9	60.1	46.2	28.2	76.1	55.2
CLIPPO	EN	93M	WebLI	100k	49.7	58.0	44.9	29.0	73.1	55.4
CLIPPO untied	EN	186M	WebLI	100k	52.4	61.8	47.2	29.5	76.6	55.0
CLIPPO	EN	93M	WebLI + 25%C4	133k	49.4	55.4	40.2	25.3	69.0	50.5
CLIPPO	EN	93M	WebLI + 50%C4	200k	45.6	51.1	34.3	21.7	61.7	43.2
CLIP* L/16	EN	652M	WebLI	100k	59.0	67.2	49.6	32.1	79.3	60.1
1T-CLIP L/16	EN	349M	WebLI	100k	58.0	65.6	49.5	31.6	80.2	57.8
CLIPPO L/16	EN	316M	WebLI	100k	56.6	64.9	50.2	33.0	77.0	61.5
CLIP*	ML	203M	WebLI	100k	50.8	59.0	43.6	27.4	71.1	53.2
1T-CLIP	ML	118M	WebLI	100k	49.2	55.2	41.6	25.4	70.9	51.0
CLIPPO	ML	93M	WebLI	100k	47.3	52.0	38.9	24.4	67.7	48.3
CLIPPO JFT init	EN	93M	WebLI	100k	57.1	59.9	43.9	29.2	71.1	55.0
CLIPPO JFT init	EN	93M	WebLI + 25%C4	133k	54.5	56.3	37.0	24.3	64.4	47.3
CLIPPO JFT init	EN	93M	WebLI + 50%C4	200k	50.9	51.8	34.3	22.1	60.5	45.1
1T-CLIP	EN	118M	LAION	100k	46.0	54.3	49.0	31.5	77.5	59.7
CLIPPO	EN	93M	LAION	100k	45.3	53.6	46.7	30.3	76.9	58.9
CLIPPO	EN	93M	LAION + 25%C4	133k	44.9	50.6	41.8	27.2	71.1	53.7
CLIPPO	EN	93M	LAION + 50%C4	200k	41.4	46.0	38.2	24.3	66.3	49.0
CLIP*	EN	203M	WebLI	250k	55.8	65.1	48.5	31.3	79.2	59.4
1T-CLIP	EN	118M	WebLI	250k	53.9	62.3	48.0	30.3	77.5	58.2
CLIPPO	EN	93M	WebLI	250k	53.0	61.4	47.3	30.1	76.4	57.3
CLIPPO	EN	93M	WebLI + 25%C4	333k	52.1	57.4	40.7	26.7	68.9	51.8
CLIPPO	EN	93M	WebLI + 50%C4	500k	48.0	53.1	35.2	23.4	64.8	47.2
CLIP* L/16	EN	652M	WebLI	250k	62.0	70.1	51.3	34.1	80.5	62.9
1T-CLIP L/16	EN	349M	WebLI	250k	60.8	67.8	50.7	32.5	81.0	61.0
CLIPPO L/16	EN	316M	WebLI	250k	60.3	67.4	50.6	33.4	79.2	62.6
CLIPPO L/16	EN	316M	WebLI + 25%C4	500k	60.5	66.0	44.5	29.8	72.9	57.3
CLIPPO L/16	EN	316M	WebLI + 50%C4	500k	56.8	61.7	39.7	27.3	70.1	54.7
1T-CLIP 384px	EN	118M	WebLI	270k	57.8	66.2	51.5	32.7	81.7	63.0
CLIPPO 384px	EN	93M	WebLI	270k	57.2	64.7	51.0	32.9	79.9	61.9
CLIPPO 384px	EN	93M	WebLI + 25%C4	350k	56.0	61.0	44.3	27.9	73.4	55.0
1T-CLIP L/16 384px	EN	349M	WebLI	270k	64.5	70.9	52.6	34.8	81.6	63.8
CLIPPO L/16 384px	EN	317M	WebLI	270k	63.9	70.5	54.4	35.3	83.6	64.9
CLIPPO L/16 384px	EN	317M	WebLI + 25%C4	520k	64.2	69.0	47.5	31.9	76.2	59.7
CLIP*	ML	203M	WebLI	250k	53.7	62.1	46.9	29.4	76.9	57.8
1T-CLIP	ML	118M	WebLI	250k	52.6	58.4	44.9	27.7	72.2	53.7
CLIPPO	ML	93M	WebLI	250k	51.1	56.1	42.5	26.6	69.9	52.9
CLIPPO	ML	93M	YFCC-100M	250k	38.2	43.4	34.7	19.7	64.7	40.6
CLIPPO I21k init	ML	93M	YFCC-100M	250k	44.7	47.4	36.1	21.3	66.0	42.3
CLIPPO I21k init	ML	93M	YFCC-100M + 25%C4	333k	43.8	44.8	33.3	19.4	61.0	37.8
CLIPPO I21k init	ML	93M	YFCC-100M + 50%C4	500k	41.2	42.0	31.4	17.8	58.4	36.8
CLIPPO I21k init	ML	93M	YFCC-100M + 75%C4	500k	34.5	33.4	26.6	14.6	53.1	31.0

Table 6. We report ImageNet-1k 10-shot linear transfer validation accuracy (I1k 10s.), ImageNet-1k zero-shot transfer validation accuracy (I1k 0s.), image-to-text and text-to-image retrieval recall@1 on MS-COCO (C I→T and C T→I) and on Flickr30k (F T→I and F I→T). "CLIPPO untied" is a two tower model where two separate ViT B/16 models (i.e. with separate parameters) are used to encode images and rendered alt-texts. "CLIPPO JFT init" and "CLIPPO I21k init" are CLIPPO models that were initialized with the parameters of ViT B/16 from [16] trained on JFT-300M and ImageNet-21k, respectively. Models with the suffix "384px" are models trained for 30k iterations at a resolution of 384px, starting from the corresponding 224px checkpoints stored right before cooldown.

VQA Table 7 shows results for all our models and baselines on VQAv2 (test-dev set). In addition to what is discussed in the main paper, we observe that co-training with 50% C4 data does not lead to improvements over co-training with 25% C4 data. Further, the gap between 1T-CLIP and CLIPPO becomes narrow as the model size grows. Increasing the resolution form 224px to 384px leads to a substantial improvement across models.

	res.	yes/no	number	other	overall
ViT B/16 JFT	224	71.16	40.71	51.55	58.39
1T-CLIP	224	76.08	42.46	53.1	61.36
CLIP*	224	77.49	44.65	55.47	63.31
CLIPPO 50%C4	224	83.81	45.45	55.62	66.08
CLIPPO	224	83.01	46.36	56.55	66.29
CLIPPO 25%C4	224	84.48	46.18	56.27	66.74
CLIPPO L/16 50%C4	224	84.33	48.2	58.68	68.05
CLIPPO L/16	224	83.74	49.33	58.9	68.05
1T-CLIP L/16	224	84.03	49.41	59.53	68.48
CLIPPO L/16 25%C4	224	84.91	49.26	59.33	68.73
1T-CLIP	384	77.92	45.21	56.45	64.02
CLIPPO	384	84.22	47.94	58.62	67.95
CLIPPO 25%C4	384	86.91	49.34	60.52	70.12
CLIPPO L/16	384	86.26	51.91	61.89	70.79
1T-CLIP L/16	384	86.3	52.01	62.32	71.03
CLIPPO L/16 25%C4	384	86.85	53.57	63.05	71.78
METER CLIP B/32+BERT	224				69.56
ViLT B/32	384				70.33
Pythia CLIP B/16	600				62.72
MCAN CLIP B/32	600				65.40

Table 7. Results on the VQAv2 benchmark (test-dev set). Our 224px and 384px models and baselines are pretrained for 250k and 270k steps (or an appropriately adapted number of steps when co-trained with C4), respectively, and fine-tuned to VQAv2. In addition to CLIPPO and baselines produced in this work, we also compare to Pythia and MCAN models with ViT vision encoders from [67], and with comparably sized METER [17] and ViLT [36] models. "ViT B/16 JFT" is the model trained on JFT-300M from [16].

Language understanding Table 8 shows additional results for our models and baselines on the GLUE benchmark. We discuss a number of observations that were not discussed in the main paper.

First, it can be seen that a randomly initialized ViT performs much worse than all the other models, including the vision encoders of the different CLIP* and 1T-CLIP variants, which all perform similarly, independently on the precise training setup.

We further present results for models that were trained with multilingual image/alt-text pairs (note that GLUE contains only English tasks). When trained for 100k steps, CLIP*, 1T-CLIP and CLIPPO obtain a lower GLUE score than their counterparts trained on English-only alt-texts. The GLUE scores of these multilingual models improve when training for 250k steps. In particular, CLIPPO almost matches its English-only counterpart, whereas CLIP* and 1T-CLIP still lag a few points behind their English-only counterparts.

Moreover, the accuracy of CLIP* and 1T-CLIP vision encoders we observe for SST-2 is in agreement with what was reported in [56, Table 10] for CLIP with a ViT-B/16 image encoder. Note that CLIPPO obtains a significantly higher score. With frozen representations we obtained 71.6% for the CLIP* vision encoder vs. 78.3% for CLIPPO, so again CLIPPO performs better by a large margin.

	lan.	training dataset	steps	MNLI-M/MM	QQP	QNLI	SST-2	COLA	STS-B	MRPC	RTE	avg
BERT-Base	EN	Wiki + BC		84.0 / 84.2	87.6	91.0	92.6	60.3	88.8	90.2	69.5	83.1
PIXEL	EN	Wiki + BC		78.1 / 78.9	84.5	87.8	89.6	38.4	81.1	88.2	60.5	76.3
BiLSTM	EN			66.7 / 66.7	82.0	77.0	87.5	17.6	72.0	85.1	58.5	68.1
BiLSTM+Attn,ELMo	EN			72.4 / 72.4	83.6	75.2	91.5	44.1	56.1	82.1	52.7	70.0
ViT from scratch	EN			33.4 / 33.2	51.2	56.4	53.9	0.0	5.1	81.2	52.7	40.8
CLIP* img. enc.	EN	WebLI	100k	65.2 / 66.5	75.7	68.0	77.8	0.0	6.9	81.5	52.3	54.9
CLIP* text enc.	EN	WebLI	100k	70.6 / 71.0	80.6	71.1	85.9	0.0	62.4	82.1	54.9	64.3
1T-CLIP img. enc.	EN	WebLI	100k	64.4 / 65.5	74.2	65.8	74.5	0.0	12.0	81.6	53.8	54.7
1T-CLIP text enc.	EN	WebLI	100k	71.6 / 71.5	83.5	80.5	85.0	0.0	74.1	82.8	54.2	67.0
CLIPPO unt. img. enc.	EN	WebLI	100k	64.8 / 65.6	76.4	67.0	77.1	0.0	7.0	81.4	51.6	54.5
CLIPPO unt. text enc.	EN	WebLI	100k	65.2 / 65.1	83.7	74.8	86.6	3.1	56.1	81.8	54.9	63.5
CLIPPO	EN	WebLI	100k	72.2 / 72.5	84.0	81.2	86.7	0.0	81.0	84.0	57.8	68.8
CLIPPO	EN	WebLI + 25%C4	133k	77.0 / 76.7	85.4	82.8	90.9	20.1	83.1	83.6	54.5	72.7
CLIPPO	EN	WebLI + 50%C4	250k	78.8 / 78.3	86.0	84.8	92.0	34.4	83.1	84.2	58.8	75.6
CLIPPO	EN	C4	100k	79.3 / 78.8	86.4	85.4	93.2	47.7	84.2	83.7	59.6	77.6
CLIPPO	ML	WMT19	100k	72.9 / 72.9	80.8	74.5	88.6	4.0	19.6	81.9	55.6	61.2
CLIPPO	EN	WMT19 BT	100k	70.0 / 70.3	80.5	80.1	84.6	10.8	65.7	81.6	56.0	66.6
1T-CLIP L/16	EN	WebLI	100k	72.8 / 73.3	84.3	81.4	88.5	0.0	79.1	82.3	53.4	68.3
CLIPPO L/16	EN	WebLI	100k	67.4 / 66.9	84.9	76.7	86.5	0.0	81.5	82.9	53.1	66.6
CLIP* img. enc.	ML	WebLI	100k	63.3 / 64.4	73.8	65.9	75.6	0.0	7.0	81.7	54.5	54.0
CLIP* text enc.	ML	WebLI	100k	63.1 / 63.1	79.2	70.6	75.6	4.4	34.8	81.2	49.8	58.0
1T-CLIP img. enc.	ML	WebLI	100k	62.9 / 64.3	73.5	63.8	71.9	0.0	6.5	81.3	53.1	53.0
1T-CLIP text enc.	ML	WebLI	100k	64.9 / 64.8	80.5	74.7	78.6	4.2	66.0	81.5	50.2	62.8
CLIPPO	ML	WebLI	100k	72.0 / 72.2	82.1	80.4	85.0	0.0	16.1	81.6	50.9	60.0
1T-CLIP img. enc.	EN	LAION	100k	66.8 / 67.6	77.9	73.3	78.8	0.0	12.9	81.7	55.2	57.1
1T-CLIP text enc.	EN	LAION	100k	72.2 / 72.8	84.1	79.8	86.9	0.0	38.0	81.4	54.2	63.3
CLIPPO	EN	LAION	100k	73.2 / 73.5	84.2	80.9	86.5	0.0	75.3	82.2	53.8	67.7
CLIPPO	EN	LAION + 25%C4	133k	77.0 / 77.0	85.5	83.3	91.1	22.0	83.3	84.6	57.0	73.4
CLIPPO	EN	LAION + 50%C4	250k	78.8 / 78.7	86.1	84.3	92.2	38.3	83.7	83.9	55.2	75.7
CLIP* img enc.	EN	WebLI	250k	66.4 / 67.5	78.6	69.4	78.6	0.0	5.2	81.2	52.7	55.5
CLIP* text enc.	EN	WebLI	250k	71.8 / 72.5	82.7	73.0	86.2	6.6	65.0	81.4	53.8	65.9
1T-CLIP text enc.	EN	WebLI	250k	72.6 / 73.0	83.8	80.7	84.9	0.0	79.6	83.3	57.0	68.3
CLIPPO	EN	WebLI	250k	73.0 / 72.6	84.3	81.2	86.8	1.8	80.5	84.1	53.4	68.6
CLIPPO	EN	WebLI + 25%C4	333k	77.7 / 77.2	85.3	83.1	90.9	28.2	83.4	84.5	59.2	74.4
CLIPPO	EN	WebLI + 50%C4	500k	79.2 / 79.2	86.4	84.2	92.9	38.9	83.4	84.8	59.9	76.6
CLIPPO	EN	C4	250k	79.9 / 80.2	86.7	85.2	93.3	50.9	84.7	86.3	58.5	78.4
1T-CLIP L/16 text enc.	EN	WebLI	250k	74.3 / 74.7	85.1	81.6	86.6	8.0	82.5	83.1	57.4	70.4
CLIPPO L/16	EN	WebLI	250k	68.4 / 67.2	85.1	77.2	87.6	0.0	81.0	84.3	52.7	67.1
CLIPPO L/16	EN	WebLI + 25%C4	500k	76.6 / 75.5	87.1	79.9	93.2	48.2	84.1	84.6	56.0	76.1
CLIPPO L/16	EN	WebLI + 50%C4	500k	82.3 / 82.4	87.9	86.7	94.2	55.3	85.8	85.9	59.2	80.0
CLIPPO L/16	EN	C4	250k	83.9 / 83.6	87.9	89.1	94.7	62.0	87.1	87.0	62.5	82.0
CLIP* text enc.	ML	WebLI	250k	64.3 / 64.6	80.8	75.7	78.6	11.2	70.7	81.9	49.8	64.2
1T-CLIP text enc.	ML	WebLI	250k	65.8 / 65.7	80.9	75.0	80.7	0.0	71.1	81.9	51.6	63.6
CLIPPO	ML	WebLI	250k	71.1 / 71.2	82.8	79.6	85.2	0.0	78.3	83.1	53.1	67.1
CLIPPO	ML	YFCC-100M	250k	71.3 / 71.5	79.1	67.9	85.7	0.0	14.0	83.4	54.9	58.6
CLIPPO I21k init	ML	YFCC-100M	250k	70.0 / 70.1	83.7	81.6	86.1	0.0	18.5	83.0	53.1	60.7
CLIPPO I21k init	ML	YFCC-100M + 25%C4	333k	75.7 / 75.1	85.2	83.5	89.6	0.0	82.3	82.7	52.7	69.7
CLIPPO I21k init	ML	YFCC-100M + 50%C4	500k	77.4 / 77.4	86.0	83.9	91.7	34.5	84.5	85.1	56.3	75.2
CLIPPO I21k init	ML	YFCC-100M + 75%C4	500k	79.8 / 79.1	86.5	84.3	92.0	44.5	85.3	88.2	58.5	77.6

Table 8. Complete results for the GLUE benchmark (dev set). The metric is accuracy except for the performance on QQP and MRPC, which is measured using the F_1 score, CoLA which uses Matthew's correlation, and STS-B which evaluated based on Spearman's correlation coefficient. "avg" corresponds to the average across all metrics. The results for BERT-Base and PIXEL are from [60, Table 3], and BiLSTM and BiLSTM+Attn, ELMo from [73, Table 6]. All encoders considered here have a transformer architecture comparable to BERT-Base (up to the text embedding layer), except for CLIPPO L/16 which uses a ViT L/16, and the two BiLSTM model variants. Wiki and BC stand for (English) Wikipedia and Bookcorpus [86] data, respectively. "ViT from scratch" is a randomly initialized, untrained ViT B/16. "CLIPPO unt." is a two tower model where two separate ViT B/16 models (i.e. with separate parameters) are used to encode images and rendered alt-texts. All models process rendered text except for "CLIP* text enc." and "1T-CLIP text enc." which process tokenized text. "CLIPPO I21k init" are CLIPPO models that were initialized with the parameters of ViT B/16 trained on ImageNet-21k.

C.3. Multilingual vision-language understanding

Multilingual image/text retrieval Fig. 8 shows the per-language retrieval performance on Crossmodal3600 [70] of CLIP*, 1T-CLIP, and CLIPPO. CLIP* obtains a slightly better performance than the other two methods which is not surprising given it uses about double the trainable parameters of the other models and separate text and image encoders. CLIPPO matches or outperforms 1T-CLIP on average, despite having fewer trainable parameters. Overall, the performance per-language correlates strongly across all models, with Japanese and Korean showing the biggest differences between CLIPPO and the other models.

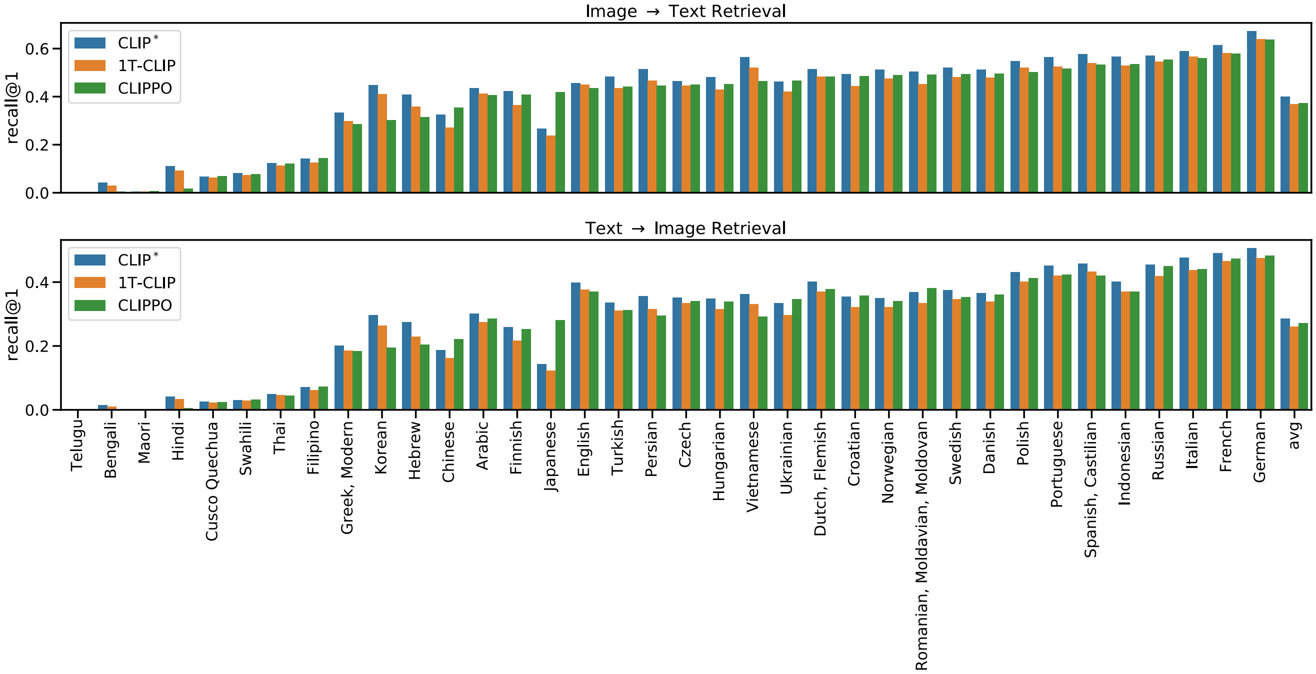


Figure 8. Per-language and average image-to-text and text-to-image recall@1 on the Crossmodal3600 data set. All the models are trained for 250k iterations on WebLI with multilingual alt-texts. CLIP* and 1T-CLIP use a SentecePiece tokenizer with vocabulary size 32,000 built from 300M randomly sampled WebLI alt-texts, whereas CLIPPO is tokenizer-free by design.

Tokenizers We use the following open-source tokenizers in our experiments:

- *T5-en* [57]: `gs://t5-data/vocabs/cc_all.32000/sentencepiece.model`
- *T5-all* [57]: `gs://t5-data/vocabs/cc_en.32000/sentencepiece.model`
- *mT5* [79]: `gs://t5-data/vocabs/mc4.250000.100extra/sentencepiece.model`

We take the first 32,000 pieces of the mc4 vocabulary to create a vocabulary of equal size to the others.

Tokenizer efficiency Fig. 9 shows the average sequence length on 20,000 samples of different languages from C4. CLIPPO obtains a balanced average performance across the selected languages.

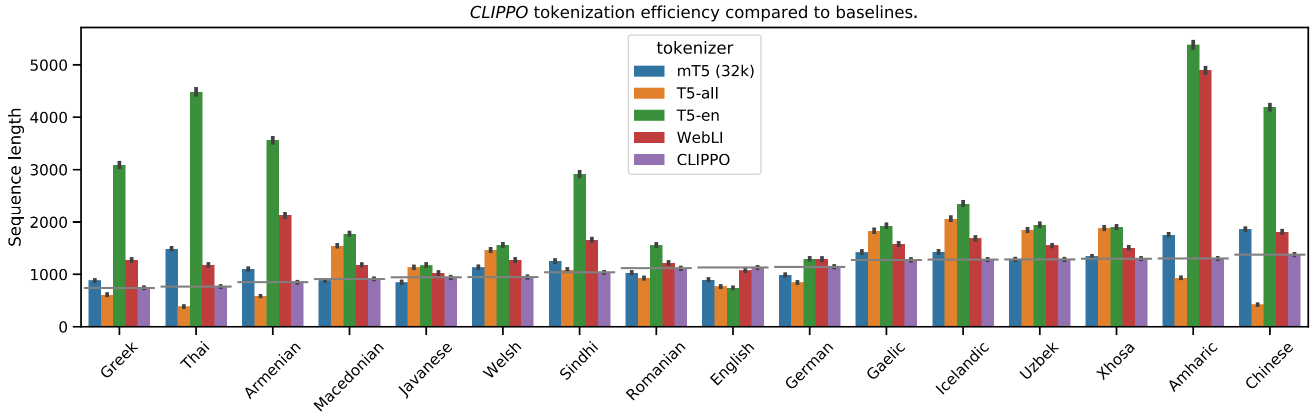


Figure 9. Sequence length of SentencePiece tokenziers derived from different corpora. All non-CLIPPO tokenizers have a vocabulary size of 32,000.

D. Ablations and analysis

D.1. Impact of weight sharing

To better understand whether a modality-shared patch embedding or modality-shared heads are degrading the performance of CLIPPO we train different models with separate embeddings and/or heads for image and (rendered) text inputs. The results in Table 9 show that neither of the variants with separate embeddings and/or heads leads to a consistent improvement in image classification or retrieval metrics compared to the default CLIPPO variant where both the embedding and head are shared. For comparison we also show a variant with two ViT B/16 models (i.e. with separate parameters) to separately encode images and rendered alt-texts, which mostly matches the CLIP* baseline.

	#param.	shared	separated	I1k 10s.	I1k 0s.	C I→T	C T→I	F I→T	F T→I
CLIP*	203M	-	all	52.9	62.8	47.2	29.7	76.8	57.2
CLIPPO untied	186M	-	all	52.4	61.8	47.2	29.5	76.6	55.0
1T-CLIP	118M	encoder, heads	embeddings	50.9	60.1	46.2	28.2	76.1	55.2
CLIPPO	93M	all	-	49.7	58.0	44.9	29.0	73.1	55.4
CLIPPO	94M	encoder, embeddings	heads	49.2	58.1	45.0	28.7	71.8	56.5
CLIPPO	94M	encoder, heads	embeddings	49.8	58.4	44.5	28.6	73.7	56.4
CLIPPO	94M	encoder	embeddings, heads	48.9	57.6	44.5	26.8	72.9	53.7

Table 9. We report ImageNet-1k 10-shot linear transfer validation accuracy (I1k 10s.), ImageNet-1k zero-shot transfer validation accuracy (I1k 0s.), image-to-text and text-to-image retrieval recall@1 on MS-COCO (C I→T and C T→I) and on Flickr30k (F T→I and F I→T). All models are trained for 100k iterations. "CLIPPO untied" is a two tower model where two separate ViT B/16 models (i.e. with separate parameters) are used to encode images and rendered alt-texts.

D.2. Impact of the text location

As we train CLIPPO with text rendered at the top left of the image, it is interesting to see how the performance changes when the text is rendered at different locations at inference time. To this end, we repeat the transfer VQAv2 experiment with text rendered in the middle and at the bottom of the image. We observe a drop for the middle/bottom locations, but this drop can be fixed simply by multiplying learning rate for the positional embedding by 3 during fine-tuning on the VQAv2 training set. Multiplying the learning rate of the positional embedding CLIP* and 1T-CLIP during fine-tuning does not affect their performance on VQAv2.

text location	top	middle	bottom
no LR scaling	66.29	60.00	61.53
3× LR for pos. embedding	66.36	66.50	66.04

Table 10. The impact of the text location on the VQAv2 test-dev score.

D.3. Typographic attacks

Prior works have identified that CLIP can be fooled by typographic attacks, whereby it reads scene text and zero-shot classifies an image according to this text text rather than the objects in the scene [23, 42, 50]. As CLIPPO shares processing for images and text, it is interesting to analyze whether the models are more prone to such typographic attacks.

We assess this on two ways: first, we test models on the real-world Typographic Attack dataset curated by Materzynska et al. [51]. The dataset was created from 20 objects. For each object there is a picture of the object without any adversarial attack, and 19 versions where a post-it note is stuck on top of the object. Written on the note is an "incorrect" label unrelated to the object. A contrastive model susceptible to typographic attacks would classify the object as one of these confounding labels. Secondly, we re-evaluate zero-shot classification accuracy on ImageNet, but for each image insert a randomly selected "incorrect" label using our Unifont renderer. A model which reads this label instead of observing the image would suffer a larger drop in ImageNet accuracy, and thus also be more susceptible to typographic attacks.

Table 11 (left) shows the accuracy with which models predict the correct label instead of the confounder on the post-it note. All models are largely able to ignore the typographic attack, and the CLIPPO models are on par with or better than the counterparts relying on a tokenizer. Table 11 (right) shows the drop in accuracy due to adversarial text labels, rendered at different locations using the CLIPPO Unifont renderer, in ImageNet classification. All models see a drop in accuracy of roughly similar magnitude, except for CLIPPO when text is positioned at the top (where it is during normal training). Here, the drop is lower, possibly indicating a distinction between "scene text" and the rendered-text inputs.

	CLIP*	1T-CLIP	CLIPPO
	Without prompts		
B/16	85.0%	89.4%	89.4%
L/16	89.4%	87.5%	93.8%
	With prompts		
B/16	87.5%	91.9%	92.5%
L/16	92.5%	88.7%	91.3%

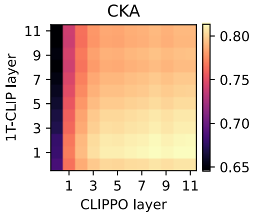
i1k acc	original	bottom	middle	top
CLIP*	65.1%	-1.3 ± 0.1%	-7.0 ± 0.1%	-2.0 ± 0.1%
1T-CLIP	61.4%	-1.4 ± 0.1%	-7.5 ± 0.2%	-2.3 ± 0.1%
CLIPPO	62.3%	-1.2 ± 0.1%	-7.3 ± 0.1%	-1.2 ± 0.1%

Table 11. Classification accuracy when exposed to typographic attacks. **Left:** The rate at which models correctly ignore real-world typographic attacks on the dataset of Materzynska et al. [51]. **Right:** The effect on the classification accuracy of adding adversarial text labels to ImageNet-1k using the CLIPPO unifont renderer (for B/16 models).

D.4. Modality gap and representation analysis

Fig. 10 shows additional modality gap visualizations, complementing those in the main paper (Sec. 4.6). In addition to a visualization for the WebLI validation set, we also show results on the MS-COCO validation set. The qualitative and quantitative trend across model variants on MS-COCO is similar to that observed for WebLI, except that the modality gap is somewhat larger for a given model variant (we use the formula from [45, Sec. 4.2] to compute the modality gap). This might be due to the fact that image/caption pairs from MS-COCO have a different distribution than the image/alt-text pairs from WebLI. We further observe that 1T-CLIP and CLIPPO models have a comparable modality gap, and adding more C4 data to the training data mix does not necessarily lead to a reduction in modality gap (going from 25% to 50% C4 data increases the modality gap for MS-COCO).

Since the modality gap measures the Euclidean distance between the image and alt-text mean embeddings it does not fully reflect how the pairwise Euclidean distance between embeddings of corresponding images and alt-texts changes. We plot histograms of the latter in Fig. 11 and observe that the average pairwise distance across models roughly follows the trend of the modality gap. However, the average pairwise distance remains larger than 0.5 even when the modality gap is smaller than 0.1, hence corresponding images and alt-text are not mapped to the same embedding.



Finally, to assess representation similarities between 1T-CLIP and CLIPPO beyond the final representation layer, and in particular to better understand the role of the tokenizer, we compute the centered kernel alignment (CKA) [38] between layer outputs for sentences from C4. Other than the first two layers, all CLIPPO layers are similar to 1T-CLIP layers.

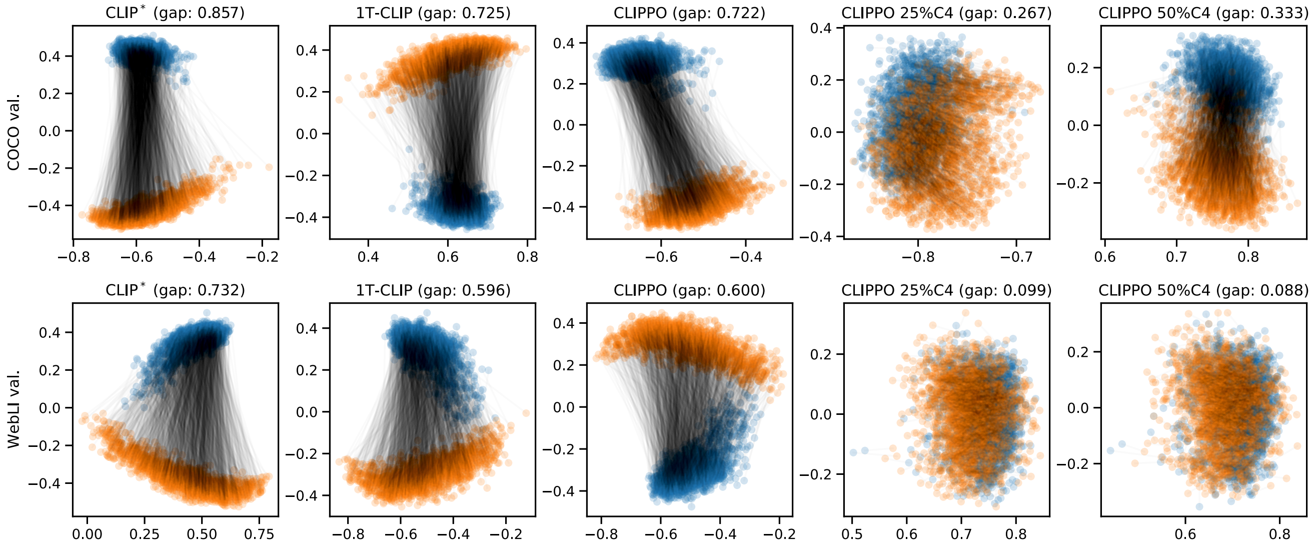


Figure 10. Visualization of the modality gap for examples from the WebLI and MS-COCO validation sets. The visualization follows the analysis from [45] and shows embedded images (blue dots) and corresponding alt-text (orange dots), projected to the first two principal components of the validation data matrix.

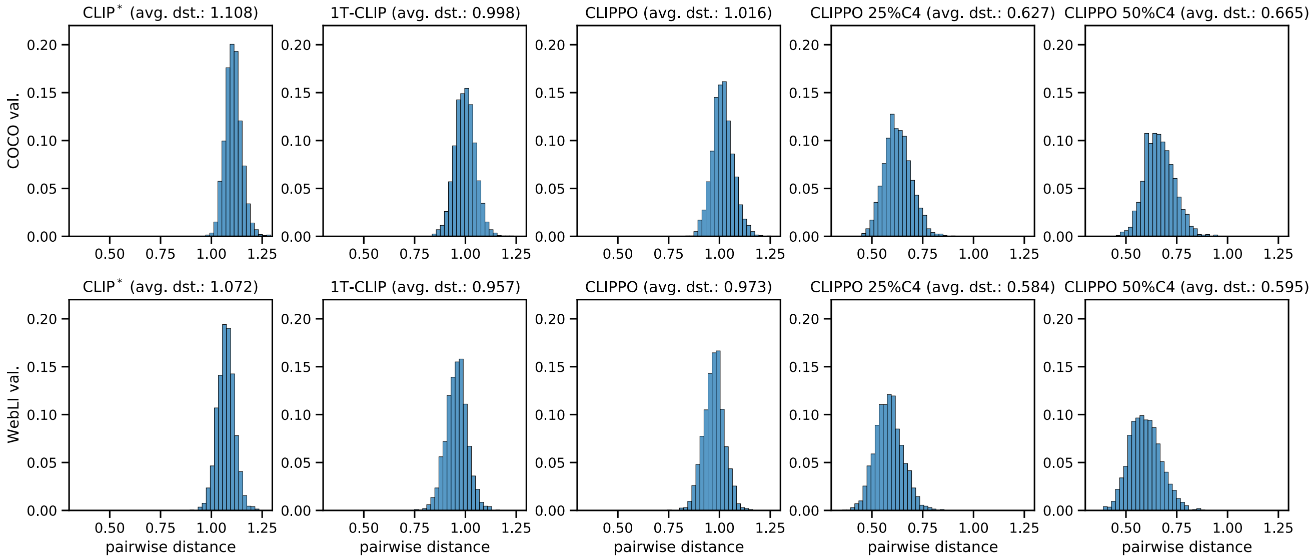


Figure 11. Histograms of the distribution of the Euclidean distance between corresponding image and alt-text embeddings. The average distance across models follows the trend of the modality gap, but the reduction in distance between embeddings when co-training with C4 is not as drastic as for the modality gap.

D.5. Patch embedding analysis

Following [16], we inspect the patch embedding of different CLIPPO variants and baselines. Concretely, we visualize the top 30 principal components of the patch embedding kernel in Fig. 12. Qualitatively, the top components for CLIP* and 1T-CLIP are similar to those for supervised ViT training in [16, Sec. 4.5], resembling a plausible basis for image patches. There seems to be no substantial visual difference between the patch embedding structure for English and multilingual variants of CLIP* and 1T-CLIP. By contrast, the top components for CLIPPO appear to contain more horizontal, high-frequency visual features than the other models, with these features becoming more pronounced as the fraction of C4 data in the training mix increases, or when multilingual alt-text is used. We speculate that this structure might be useful to represent letters and subwords with varying horizontal position as prevalent in the rendered text images fed to CLIPPO.

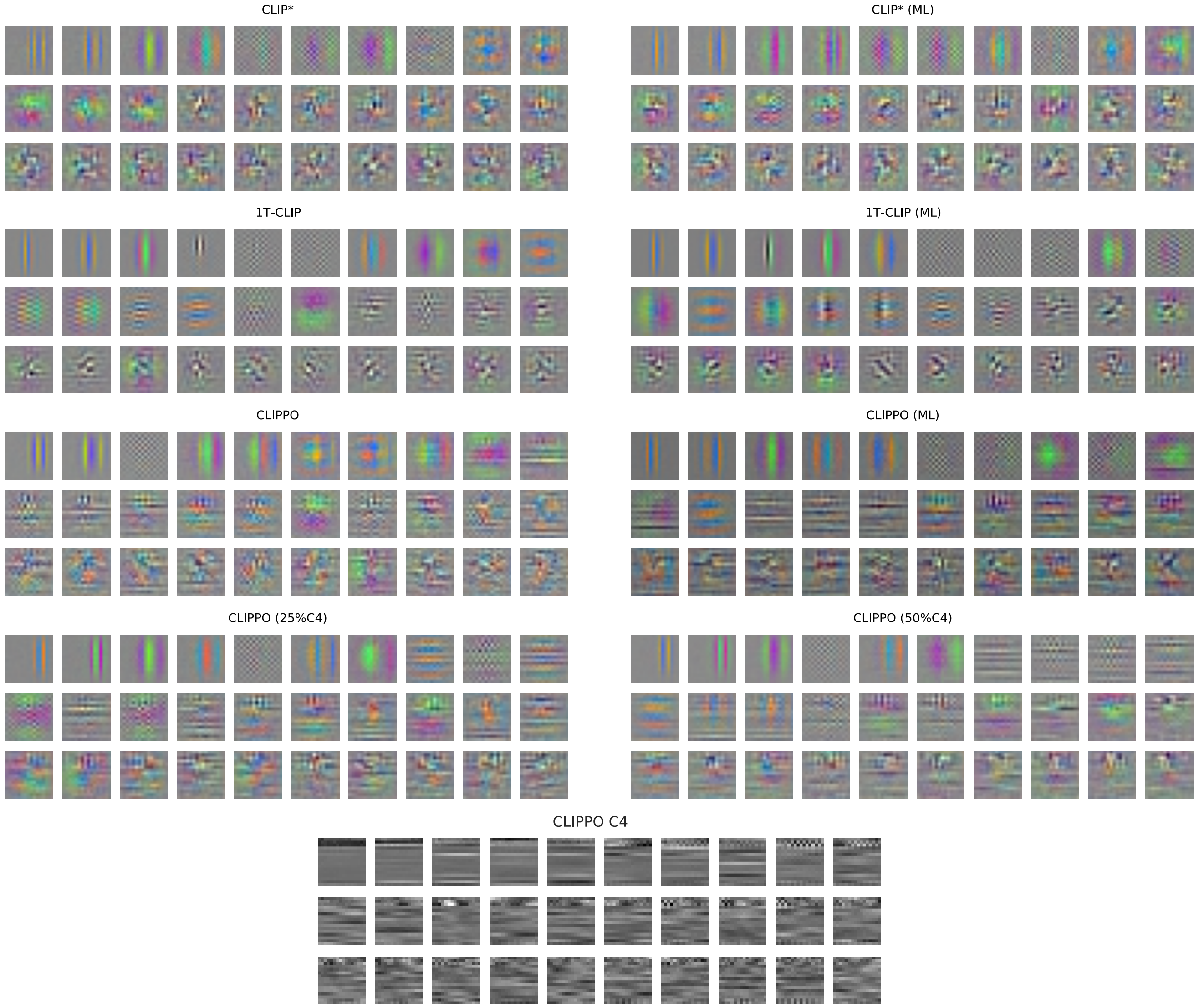


Figure 12. Visualization of the top 30 principal components of the patch embedding kernel for CLIPPO variants and baselines. The top components for CLIPPO appear to contain more horizontal, high-frequency visual features than the other models.