On Scaling up a Multilingual Vision and Language Model

Supplementary Material

A. Additional Model Details and Examples

A.1. PaLI-X Architecture Illustration

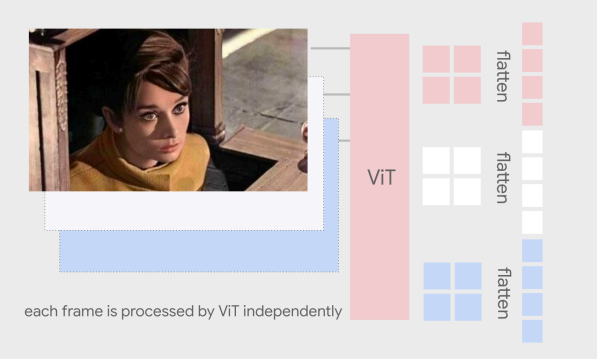

Figure 3. Visual input for videos: each frame is independently processed by ViT; patch embeddings are flattened and concatenated together to form the visual representation. (The example input image is in the public domain).

A.2. Tuning ViT-22B for better OCR capabilities

The vision encoder's ability to understand text is crucial to several downstream tasks and general usability. JFT-based pre-training is insufficient to cover this, and so we tuned ViT-22B on WebLI-OCR data. In order to stay true to the original discriminative classification-based objective used for ViT-22B, we turn OCR into a bag-of-words prediction task. OCR texts are tokenized using the mT5 tokenizer [83] across all languages, and the model is trained to predict whether or not a given token occurs in an image. This is treated as multilabel classification, with an expanded classification head.

In the ablation study shown in Table 20, we confirm that this this extra tuning step indeed has a significant improvement on Scene-Text understanding capabilities, demonstrated by the performance on ST-VQA and TextVQA. Meanwhile, the performance on regular VQA tasks such as those in the VQAv2 benchmark also improves.

A.3. Illustrative PaLI-X Examples

Table 10 shows representative examples of PaLI-X, illustrating improved abilities related to counting (both of the simple and complex variety), in context text-reading capabilities, and spatial awareness.

B. Additional results: Image Captioning and VQA

B.1. Information of Downstream Image Benchmarks

B.2. Extended Tables of Image Benchmarks

An extended table of results on some Image Benchmarks is shown as Table 12.

B.3. Multi-lingual Captioning

Multilingual captioning on XM-3600 The Crossmodal-3600 (XM3600) benchmark contains a geo-diverse set of 3600 images with human-annotated reference captions in 36 languages [23]. Table 13 presents multilingual results for both PaLI (current SoTA on XM-3600) and PaLI-X, both finetuned with 224×224 resolution. Overall, PaLI-X improves on the SoTA performance across 5 of the 7 languages we report here (and for 14 of the total 35 languages considered); notably, the performance on English is 4 CIDEr points lower compared to PaLI. The 35-language average CIDEr score is in the same ballpark between PaLI and PaLI-X, with a slight +0.5 advantage for PaLI.

B.4. TallyQA and the emergence of complex counting capability

We present in Table 14 the performance of similar models across a wide range of capacity – from 700M parameters to 55B parameters for PaLI-X. The graphs in Fig. 4 illustrate how simple counting appears to follow a more linear progression as parameter-size increases, while complex counting appears to show emergence somewhere before the datapoint provided by the performance of PaLI 17B. This corresponds to our intution that complex counting is a true multimodal task that requires additional capabilities from a model, in terms of the alignment that is required between the visual information and the prompt specification.

B.5. Details on Few-shot Modeling

B.5.1 Few-shot Formulation

Figure 5 illustrates the network flow of a few shot model. The text and prompt part of each shot is embedded and concatenated as text features for the PaLI-X model. Each shot's images and the target image are independently encoded by the ViT component, and the ViT features are concatenated along the sequence axis as visual features. Conditioned on that sequence, the PaLI-X decoder autoregressively makes the predictions for the target image.

Image Credit: Wikimedia Commons [CC BY-SA 4.0]

Q: how many coins are there?
A: 12

Q: how many one-dollar coins are there?
A: 2

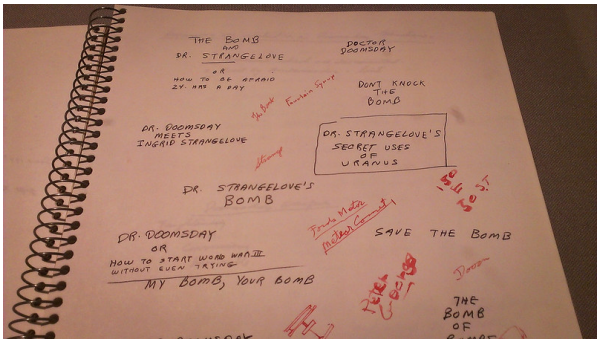

Image Credit: ChrisGoldNY (flickr) [CC BY-NC 2.0]

Q: what is written inside the box?
A: dr. strangelove's secret uses of uranus

Q: what is written on the top-left corner of the page?
A: the bomb and dr. strangelove

Q: what is written on the top-right corner of the page?
A: doctor doomsday

Public Domain Image

Captioning:
a movie poster for rebecca starring laurence olivier and joan fontaine.

Q: who directed this movie?
A: alfred hitchcock

Q: who produced this movie?
A: david o. seznick

Table 10. Examples of counting, text reading capabilities with context and spatial awareness. Results are generated by the multi-task-finetuned models using the model's inherent OCR capabilities (i.e., without the use of an external OCR system).

Benchmark	Visual Domain	Description	Metric
COCO Captions	Natural Images	Captioning of natural images	CIDEr
NoCaps		Captioning of natural images	CIDEr
TextCaps		Captioning of natural images containing text	CIDEr
VizWiz-Cap		Captioning of photos taken by people who are blind	CIDEr
VQAv2		VQA on natural images	VQA accu.
OKVQA		VQA on natural images requiring outside knowledge	VQA accu.
TextVQA		VQA on natural images containing text	VQA accu.
VizWiz-QA		VQA on photos taken by people who are blind	VQA accu.
ST-VQA		VQA on natural images containing text	ANLS
TallyQA		VQA with counting questions	EM
OVEN		VQA on natural images for visual entity recognition	EM
InfoSeek		VQA on natural images for visual info-seeking questions	Relaxed EM
OCR-VQA	Illustrations	VQA on images of book covers	EM
ChartQA		VQA on images of charts	RA
AI2D		VQA on images of scientific diagrams	EM
DocVQA	Documents	VQA on images of scanned documents	ANLS
InfographicsVQA		VQA on images of infographics	ANLS
Screen2Words	UIs	Captioning a UI screen to describe functionality	CIDEr
Widget Captioning		Captioning a UI component on a screen	CIDEr

Table 11. Summary of Image Captioning and VQA benchmarks used for evaluating PaLI-X

	COCO	NoCaps		VQAv2		OKVQA	TallyQA	
Model	Karp.-test	val	test	test-dev	test-std	val	simple	complex
SimVLM	143.3	112.2	110.3	80.03	80.34	-	-	-
CoCa (2.1B)	143.6	122.4	120.6	82.3	82.3	-	-	-
GIT (0.7B)	144.8	125.5	123.4	78.56	78.81	-	-	-
GIT2 (5.1B)	145.0	126.9	**124.8**	81.74	81.92	-	-	-
OFA (0.9B)	145.3	-	-	82.0	82.0	-	-	-
Flamingo (80B)	138.1	-	-	82.0	82.1	57.8*	-	-
BEiT-3 (1.9B)	147.6	-	-	84.2	84.0	-	-	-
PaLM-E (562B)	138.7	-	-	80.0	-	**66.1**	-	-
MoVie	-	-	-	69.26	-	-	74.9	56.8
PaLI (17B)	149.1	**127.0**	124.4	84.3	84.3	64.5	81.7	70.9
PaLI-X (55B)	**149.2**	126.3	124.3	**86.0**	**86.1**	**66.1**	**86.0**	**75.6**

Table 12. Results on COCO Captions (Karpathy split), NoCaps, VQAv2, OKVQA, and TallyQA with end-to-end modeling without OCR pipeline input. The "simple" and "complex" are test subsplits.

Model	en	fr	hi	iw	ro	th	zh	35-lang avg.
PaLI	**98.1**	75.5	31.3	46.8	35.8	72.1	**36.5**	**53.6**
PaLI-X	94.2	**78.7**	**32.0**	**46.9**	**36.9**	**75.3**	36.1	53.1

Table 13. CIDEr scores on image captioning for the Crossmodal-3600 benchmark for seven diverse languages (English, French, Hindi, Hebrew, Romanian, Thai, and Chinese), as well as the average of the 35 languages covered by the benchmark. Both models are finetuned with 224×224 resolution.

Encoder shot and Decoder shots While images for all few-shot examples and target example are given as input to the model, text information can be provided in different ways. During inference time, all text information related to

Model	TallyQA simple	TallyQA complex	Weighted average
PaLI (700M)	66.9	55.6	62.4
PaLI (3B)	72.0	56.7	65.9
PaLI (17B)	76.2	65.5	71.9
PaLI-X (55B)	81.3	71.0	77.2

Table 14. Performance on TallyQA splits for simple and complex questions. All models use 224×224 image resolution.

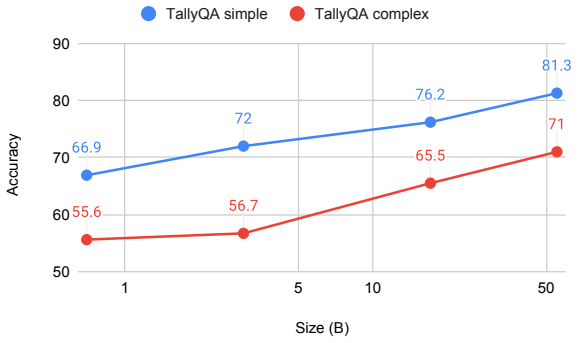

Figure 4. Performance on TallyQA splits for simple and complex using PaLI variants and PaLI-X. All models use 224×224 image resolution. The emergent behavior on complex counting beyond the 3B size is made clear with PaLI-X.

the few-shot examples is given to the encoder; in the case of a Multi-answer VQA task, for example, this includes both the prompts that contain the questions, and the expected answers. Prompt for the target example is also given to the encoder, and the decoder is tasked with generating an answer for the target example. During training, however, we increase the training efficiency by making the model predict answers for both the target example and selected shots (the *decoder shots*). That is, we partition the N shots in two sets: encoder shots ($N_e > 0$) and decoder shots ($N_d \geq 0$), such that $N_e + N_d \leq N$. We use up to 4 shots in total during pre-training (i.e. $N = 4$), and sample N_e uniformly at random from 1 to N. Text input for encoder shots contain both prompts and answers. The decoder shots, however, act as if they were target examples: their text input to the encoder contains only the prompt, and the decoder needs to predict answers for the decoder shots in addition to the target example.

Attention re-weighting Increasing the number of shots turned out to be challenging, potentially due to cross-attention to target example input tokens getting diluted by the large number of shots. To address this, we introduce an attention re-weighting mechanism. As shown in Figure 6, we explicitly boost the weights for cross attention between decoder tokens and encoded tokens from the target example (that is, the target image and the target text prompt).

Specifically, if there are N shots in total, when decoding each token we multiply the cross attention weights by N for the target image and text tokens from the encoder outputs. We observe this attention re-weighting technique is especially helpful when we provide the model with many shots (Table 15). [84] introduces a technique along similar lines to manipulate attention weights when gathering them from different threads of encoded shots at inference time.

Setup	4-shot	8-shot	16-shot
w/ re-weighting	81.5	82.1	82.8
w/o re-weighting	81.2	75.4	67.5

Table 15. Effect of attention re-weighting on num_shot more than 4 based on a 3B (mT5-large + ViT-G/14) PaLI-X model.

B.5.2 Additional Few-shot Results

Multilingual captioning results Table 16 reports the CIDEr scores for 7 languages and an average over 35 languages to demonstrate PaLI's multilingual captioning capabilities on the XM3600 benchmark in teh few-shot setting. The pre-trained model (no few-shot finetuning) achieves an average score of 22.7. The PaLI-X model achieves an average score of 45.1 for 4 shots and 47.1 for 32 shots. Note that the 32-shot PaLI-X average CIDEr score is only 6 points behind the fully finetuned model, which uses roughly 600k training examples per language (while the few-shot approach does not update the model parameters).

Qualitative results Figure 7 shows 3 examples on few-shot captioning and VQA tasks for qualitative analysis. The first row shows captions for the images using the images' original language, demonstrating the cross multilingual transfer of the few-shot capability. The second row captions the images with a country's popular food, showing that the few-shot approach can access the model's world knowledge. The last row shows a VQA with an explanation-like scenario where we ask if the technologies in the images are "new". Generally speaking, the shown personal computer was produced more than 40 years ago and could be regarded as old technology considering the fast pace of the current high-tech

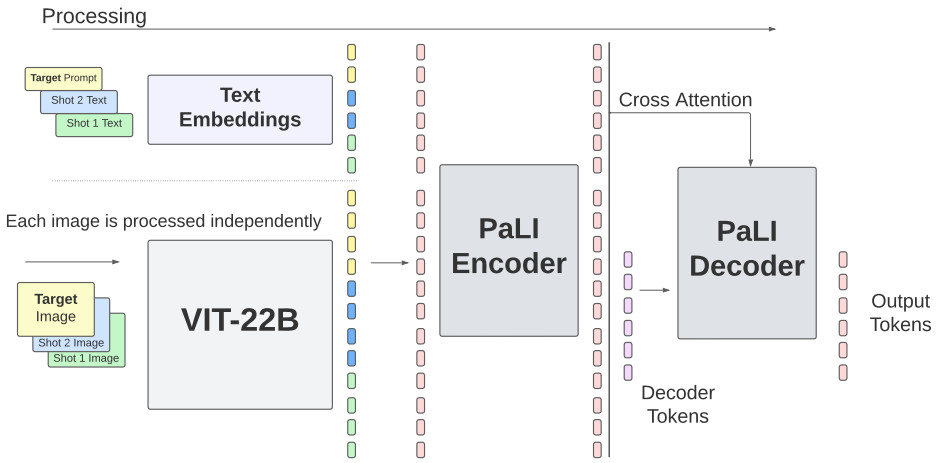

Figure 5. A detailed view on how the few-shot exemplars are fed to the model components.

	Crossmodal-3600 Captioning							
	en	fr	hi	iw	ro	th	zh	35-lang avg.
PaLI-X 0-shot	48.8	25.0	10.5	20.1	13.0	33.3	18.4	22.7
PaLI-X (2 text-only shots[5])	54.5	46.7	12.0	22.2	9.4	40.3	23.7	25.8
PaLI-X 4 shots	77.8	62.5	22.2	38.7	30.2	56.0	27.7	45.1
PaLI-X 32 shots	81.4	66.1	25.6	40.6	32.4	59.4	29.7	47.1
PaLI-X (finetuned)	94.2	78.7	32.0	46.9	36.9	75.3	36.1	53.1

Table 16. Few-shot performance of the PaLI-X model on multilingual captioning tasks.

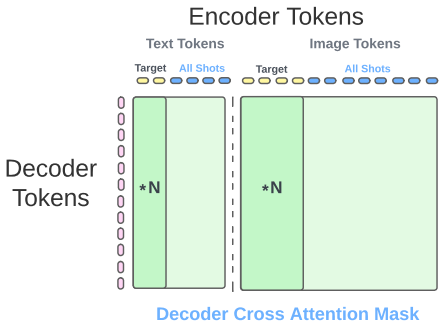

Figure 6. Re-weighted attention with few-shots.

development. However, the 3 input shots provide the detailed calibration for the concept of "new" and the few-shot model successfully take the context and output "new" with plausible explanation to the very old PC.

B.5.3 Few-shot ablation results

In this section, we present and discuss some ablation results for few-shot we explored in order to inform our final design choices on PaLI-X. Unless otherwise specified, we use a 700M-parameter model with the same encoder-decoder architecture, consisting of a ViT-B/16 vision encoder and a mT5-base encoder-decoder language model.

Benefit from using Episodic WebLI Table 17 shows that the Episodic WebLI dataset is essential for the model to develop few-shot capability.

Model components	Setup	COCO 4-shot (CIDEr)
mT5-base + ViT-B/16	full mixture	60.7
	w/o EW	11.5

Table 17. Effect of Episodoc WebLI on 4-shot COCO Captions.

Pooling vs not pooling image tokens To mitigate the computational burden that arises with many shots, we can pool (for example, average) the per-image tokens before concatenating all input tokens. This pooled image tokens model achieved a CIDEr score of 56.3 for 4-shots COCO captioning, which is substantially lower than the full model's CIDEr score of 61.7. This highlights the importance of keeping all the tokens coming out of the ViT encoder, despite the computational overhead.

Figure 7. Qualitative Results on few-shot captioning (first two rows) and VQA (the last row) tasks.

Limited-range Encoding Attention. We explore per-example image-text attention, as proposed and applied in [10]. Under this approach, the image query tokens for each example can only attend to its corresponding text tokens, while the text query tokens can attend to all tokens. By using this per-example attention model, we achieved a CIDEr score of 59.6, which is 2.1 points lower than the full attention model's CIDEr score of 61.7 for 4-shots COCO captioning.

Attention re-weighting for large number of shots. We report the few-shot results on COCO captioning from early-stopped PaLI-2 3B models; in this case, we did not apply normalized attention in training. We provide the test results with and without attention re-weighting during *inference* for a different number of encoder shots. Attention re-weighting achieves increasing CIDEr scores of 82.1, 84.3 and 84.5 with 4, 8 and 16 shots respectively. On the other hand, the model achieves 83.4, 76.5 and 66.3 without attention re-weighting. The decreasing performance may suggest that the model fails to locate the target image and text prompt among the large number of shots, whereas the attention re-weighting helps the model to focus on the target features. Accordingly, we decided to include attention re-weighting during finetuning for PaLI-X.

Distributing shots between encoder and decoder. We explore the use of both encoder and decoder shots during pre-training. We pretrain the PaLI-2 700M model on PaLI-2 mixtures with varying number of encoder shots (between 1 and 4). The remaining shots (up to exactly 4) are used as decoder shots. Using only encoder shots leads to a 64.0 CIDEr score for 4 shots in COCO captioning. The best mix of encoder and decoder shots achieves a CIDEr score of 65.2. This suggests splitting shots leads to a more challenging pre-train task that helps the model learn more efficiently.

B.6. Finetuning hyperparameters

The hyperparamter choices for downstream finetuning experiments are summarized in Table 18. As mentioned in the Main Text, for all of the downstream finetuning experiments, we used a reduced set of hyperparameters, without heavy per-task optimization.

B.7. Multi-task finetuning

We deduplicated every training set mixture over the test sets of every task in order to prevent leakage of any test-set examples into the training set. The mixture is formed by putting the training examples of each subtask together, with heuristic adjustments for a better balance. Following the resolutions for the single-task finetuning, the multi-task captioning and VQA finetuning are done with 672 and 756 image reso-

Benchmark	learning rate schedule	Steps before LR decay to 0	batch size
COCO	linear decay from 1e-4	10k	256
VQAv2		20k	256
OCRVQA		20k	256
Multitask-VQA		20k	256
Multitask-Captioning		20k	256
All other		5k	128

Table 18. Hyperparameter used for finetuning PaLI-X.

lutions, respectively. The multitask finetuning covers just about 5M examples, which is 20k steps with a batch size of 256. For scene-text and document understanding tasks, the multi-task finetuning uses the end-to-end setting without OCR pipeline input.

The following aspects made multitask finetuning particularly challenging: (i) all tasks used the same prompt without task-specific indicators; the model is thus required to adapt to the style of multiple benchmarks simultaneously. 2) We do not perform per-task validation set optimization. All subtasks are evaluated using the same checkpoint, but tasks converge to their optimal value at a different pace.

B.8. Ablation studies

We first show in Table 20 the advantage brought by the OCR co-training stage of ViT-22B. We pair the vanilla ViT-22B and the ViT-22B with additional OCR co-training with a small language model mT5-base and pretrain these models on 40M of WebLI-OCR data with the splitOCR objective, before finetuning on ST-VQA. Co-training on image and OCR classification has a significant advantage on ST-VQA and TextVQA. In the meantime, the performance on VQAv2, which is not very scene-text heavy, is improved as well. Moreover, we found that making the top left patch white, which helped the co-training of image classification and ocr classification on ViT-22B, is not required for the subsequent training of PaLI-X.

For ablation of the PaLI-X training procedure, we used a 5B model with UL2-3B and ViT-G with 2B parameters, which is roughly a 10:1 down-scale of the PaLI-X 55B model. For stage 1 training, we show in Table 21 that adding image token generation does not harm the performance on the main image+language understanding tasks.

We performed additional studies around the scaling behavior. Due to that for stage 1 training with 224×224 image resolution, PaLI-X's mixture is similar to PaLI's, the comparison for results after stage 1 training should reflect the benefit of further scaling from PaLI's 17B parameters. Besides the TallyQA comparison shown in Table 14, in Table 22 we further expand this comparison to cover VQAv2, COCO captions and TextVQA, showing the benefit of scale across the board, especially for more difficult tasks that require fine-grained understanding such as TextVQA and TallyQA.

C. Additional results: Video Captioning and QA

Below we give a brief description of each video data set we used for evaluation. Note that we freshly collected the data when performing the experiments, which led to different effective numbers of videos in different splits in some cases, see Table 23.

These descriptions refer to the original dataset size, but we train on (sometimes significantly) fewer videos — the exact numbers are given in Table 23. This is because not all videos in the datasets were available online at the time of writing (e.g., due to user deletion).

C.1. Datasets & Benchmarks

MSR-VTT [50]: This dataset consists of 10K open domain video clips for video captioning, with 20 captions each. The duration of each video clip is between 10 and 30 seconds. We follow the standard splits proposed by [50] and report results on the test set.

VATEX [51]: VATEX includes captions for 41K videos sampled from the Kinetics-600 dataset, with 10 English captions each. We report results on the English public test set.

ActivityNet Captions [52]: This dataset consists of 100K temporally localized sentences for 20k videos. We follow the standard split containing 50/25/25% of the dataset for training, validation and testing, and use ground truth temporal proposals at evaluation following [52]. Note that following other works [58], we use the `val_1` split for validation and `val_2` split for testing.

Spoken Moments in Time (SMIT) [53]: This dataset consists of long captions obtained via audio recordings for 500k short video clips. While this dataset has been traditionally only used for text to video retrieval, we find that it is a strong benchmark for captioning as it is the largest manually annotated set of videos with text captions.

ActivityNet-QA [56]: The dataset contains 58,000 question-answer pairs for videos in the ActivityNet dataset [85]. We report accuracy (using exact string match)

Model	VQA v2	OK VQA	Text VQA	VizWiz VQA	ST VQA	OCR VQA	Info VQA	Doc VQA	Chart QA	Avg.
Split	test-dev	val	val	test-dev	val	test	test	test	test	-
Previous Multi-task SOTA	84.3	64.5	68.4	71.6	75.1	71.3	40.0	76.6	70.5	-
Single-task FT	**86.0**	**66.1**	**71.9**	**72.6**	**80.2**	**75.9**	49.2	80.0	**70.9**	-
Multi-task FT	84.3	63.5	71.4	71.4	79.0	73.4	**50.7**	**80.9**	70.6	-
Multi-task (+/-)	-1.7	-2.6	-0.5	-1.2	-1.2	-2.4	+1.5	+0.9	-0.3	-0.8

Table 19. Scores from multi-task finetuning compared with those from single-task finetuning for VQA. Validation or test-dev set numbers are reported for some tasks.

Model	OCR-task Indicator	ST-VQA	TextVQA	VQAv2	3-task avg.
mT5-base + Vanilla ViT-22B	No	42.6	36.1	68.9	49.2
mT5-base + ViT-22B-OCR	No	**47.0**	38.9	69.8	**51.9**
mT5-base + ViT-22B-OCR	Yes	46.2	**39.4**	**70.2**	**51.9**

Table 20. Advantage of the OCR co-training stage of ViT-22B. Pretraining is performed with resolution 224×224 and finetuning is with 448×448. Numbers reported are on validation split.

Mixture	COCO	VQAv2
without ViT-VQGAN data	139.3	77.3
with 10% ViT-VQGAN data	139.7	77.1

Table 21. Ablation experiment showing adding ViT-VQGAN tokens does not harm understanding performance (captioning and VQA tasks).

Model	VQAv2$_{@224}$	COCO$_{@224}$	TextVQA$_{@490}$	TallyQA$_{@224}$
PaLI-3B [5]	76.0	141.4	41.6	65.9
PaLI-17B [5]	77.8	142.5	51.8	71.9
PaLI-X-55B	**80.8**	**144.2**	**65.2**	**77.2**

Table 22. Performance vs scale for VQAv2, COCO captions, TextVQA and TallyQA by evaluating stage 1 224-res checkpoint.

on the test split. Note that we do open-ended generation for all VideoQA datasets.

MSR-VTT-QA [55]: This dataset was created using a semi-automatic pipeline on top of the MSR-VTT dataset. We report accuracy (using exact string match) on the test split.

NExT-QA [54]: We focus on the Open-Ended QA task, which consists of 52,044 question-answer pairs for a total of 5,440 videos (sampled from the VidOr dataset[86]). Exactly following Next-QA [54] and Flamingo [10], we report the Wu-Palmer Similarity (WUPS) on the test set.

D. Additional results: Image Classification

Setup for zero-shot and finetuning evaluation The setup used for the experiments here uses the PaLI-X model to generate directly the (English) class name using the captioning prompt. The output is considered correct if it matches exactly the class name (apart from ImageNet-REAL, where we check if the class corresponding to the output is in the set of correct labels).

Zero-shot Evaluation results We use the same scoring technique as in PaLI [5] to evaluate PaLI-X in zero-shot setting (without training on any Imagenet data). We use the PaLI-X model obtained after the first stage of training (using the base 224 image resolution).

The results are presented in Table 24. We compare the results to PaLI [5] - previous zero-shot generative SOTA, and Flamingo [10] - another generative model of similar architecture with comparable 1-shot and 5-shot results. Overall, we report that the results between PaLI and PaLI-X for 0-shot are similar.

Finetuning To test image classification capabilities, we finetune PaLI-X on ImageNet [62] and evaluate the resulting model on ImageNet-REAL [63] and out-of-distribution datasets: ImageNet-R [64], ImageNet-A [65], ImageNet-Sketch [66], ImageNet-v2 [67].

We use the model from the first training stage (at resolution 224) and the one from the last training stage (at resolution 756). We use the same training hyperparameters for all of runs (selected without any hyperparameter tuning).

The results can be seen in Table 25. We compare the results to generative model with open vocab – GiT2 [9] (using 384 image resolution), which is the current SOTA for full-finetuning on ImageNet. PaLI-X achieves close to SOTA results for generative models on Imagenet, and other datasets.

		MSR-VTT	VATEX	ANet-Cap	SMIT	M-V-QA	ANet-QA	NExT-QA
Original size	train	6513	25991	37421	481094	158581	32000	37523
	valid.	497	3000	17505	14604	12278	18000	5343
	test	2990	6000	17031	3513	72821	8000	9178
Dataset size	train	4768	22902	30982	481094	116943	28020	37523
	valid.	327	2657	14604	8096	8215	15890	5343
	test	2144	5276	14234	3513	53014	7050	9178
% Remaining	train	73.21	88.12	82.79	100.00	73.74	87.56	100.00
	valid.	65.79	88.57	83.43	100.00	66.91	88.28	100.00
	test	71.71	87.93	83.58	100.00	72.80	88.13	100.00

Table 23. We freshly collect the data sets from the respective data sources. In cases where there are multiple question-answer pairs per video we report the number of question-answer pairs. Similarly, for ActivityNet Captions we report the number of captions. Due to missing videos which were removed after the original data sets were defined, most of our data sets are missing 10% of the videos or more.

Model (ImageNet data)	INet	REAL	INet-R	INet-A	INet-Sketch	INet-v2	ObjNet
Flamingo-80B (1-shot)	71.9	-	-	-	-	-	-
Flamingo-80B (5-shot)	77.3	-	-	-	-	-	-
PaLI (17B) (0-shot)	**72.11**	**76.43**	81.97	44.70	**63.83**	**64.46**	42.62
PaLI-X (0-shot)	71.16	75.75	**82.96**	**46.13**	61.58	63.91	**44.58**

Table 24. Top 1 accuracy results of 0-shot image classification on ImageNet [62], ImageNet-REAL [63], ImageNet-R [64], ImageNet-A [65], ImageNet-Sketch [66], Imagenet-v2 [67] and ObjectNet [87].

Model (resolution)	INet	REAL	INet-R	INet-A	INet-Sketch	INet-v2
GIT2 (384)	**89.22**	-	-	-	-	-
PaLI 3B (224)	85.11	88.71	**81.11**	45.71	70.00	78.23
PaLI 17B (224)	86.13	88.84	78.21	50.00	71.21	78.91
PaLI-X (224)	88.22	90.36	77.66	55.97	72.56	81.42
PaLI-X (756)	88.82	90.80	79.97	**73.47**	**73.39**	83.48
PaLI-X † (756)	89.19	**90.98**	80.06	72.57	73.37	**83.66**

Table 25. Classification (top-1) accuracy with Imagenet [62] fine-tuning on: ImageNet, ImageNet-REAL [63], ImageNet-R [64], ImageNet-A [65], ImageNet-Sketch [66], Imagenet-v2 [67] (resolution in parentheses). PaLI-X † fine-tuned for 2.2x more steps.

E. Object Detection

E.1. Object detection as a VLM task

Object detection is framed similarly to Pix2seq [70], with two key differences: the use of a natural language vocabulary, and class-conditioning. Prompt classes are fed to PaLI-X's text encoder, in the format `detect class1 and class2 and class3`. The model is trained to only output bounding boxes corresponding to classes in this prompt. We represent bounding boxes as coordinates in the same style as pix2seq [70]; that is, 4 integers $\texttt{y}_{\texttt{min}}$ $\texttt{x}_{\texttt{min}}$ $\texttt{y}_{\texttt{max}}$ $\texttt{x}_{\texttt{max}}$ ranging from 0 to 999. Figure 8 shows an example input.

Prompt sampling hyperparameters During training, a prompt for each example. We construct prompts from three pieces of information:

- *Positives*: These are the bounding boxes for objects definitely present in the image. During training, per example we sample $p^+ \sim \mathcal{U}(0, P^+_{\max})$, and keep that proportion of positives.
- *Negatives*: These are the known instance negatives i.e. bounding boxes for objects definitely not present. For exhaustively labelled datasets like COCO, this is simply classes not labelled as positives. For non-exhaustively labelled datasets like LVIS, these are the classes not labelled as positives, which were presented to raters. During training sample $f^- \sim \mathcal{U}(0, 5.0)$, and use up to $f^- \times n^+$, where n^+ is the number of positives after sampling p^+.
- *Global negatives*: These are negatives which are not explicitly labelled as negatives. They are taken from a wider label space combining multiple detection datasets. For a given example, valid global negatives consist of

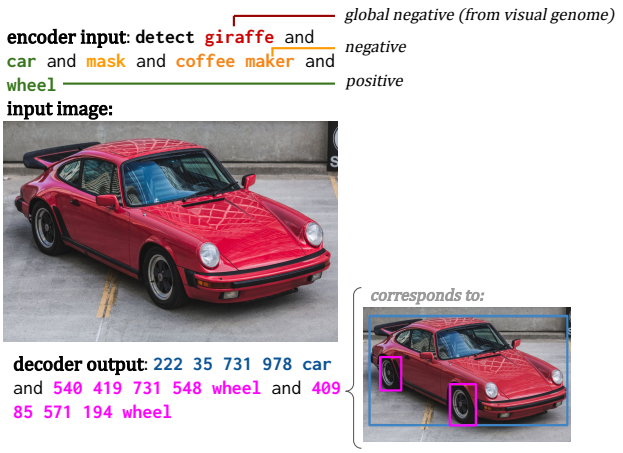

Image credits: Matthew Henry, burst, https://burst.shopify.com/photos/vintage-red-porsche

Figure 8. An example training pair, consisting of the text prompt, the image and the expected output. The prompt consists of multiple classes; we show a hypothetical Open Images V4 example, with positives 'car' and 'wheel', negative 'giraffe' and global negatives 'mask' and 'coffee maker' (sampled from the visual genome label space).

classes from the wider label space not explicitly labelled as positives or negatives. During training, we sample $f^{GN} \sim \mathcal{U}(0, 5.0)$ and append $f \times n^+$ global negatives, where n_+ is the number of positives after sampling p^+. By default, the combined label spaces of Visual Genome, Objects365 and OpenImagesV4 was used as the global label space, with the exception of detection finetuning, where LVIS and COCO label spaces were also added.

We truncate the number of total classes to n_{max}. n_{max} and P^+_{max}) are tuned per dataset to meet sequence lengths. Afer truncatation, we shuffle classes in the prompt.

E.2. Preprocessing

During pre-training, data is preprocessed to remove all LVIS-rare labels, following the protocol of OwlViT [28]. This is not done for detection finetuning. Images are randomly flipped horizontally, and randomly resized to between 0.3 and 2.0 $\times$ their original sized, followed by selecting a random square crop of the current training resolution. If the image is resized to be smaller than the current resolution, it is left as is. Images are finally padded to a square.

E.3. Licenses and attribution for images used in Main Text Figure 2

- **Watermelon** Credit: Sarah Pflug [6]
- **Bowls** Credit: andrea [7]
- **Business cat** Credit: Sarah Pflug [8]
- **Wall** Credit: Matthew Henry [9]

[6] https://burst.shopify.com/photos/cutting-watermelon

[7] https://www.flickr.com/photos/ariesandrea/502826051/ CC-BY-NC-ND 2.0

[8] https://burst.shopify.com/photos/business-cat-in-office

[9] https://burst.shopify.com/photos/man-walking-in-front-of-this-is-paradise-wall?c=urban-life

F. Model Fairness Supplementary Materials

We focus our RAI evaluation on three parts: (1) harmful associations, such as toxicity and profanity, (2) demographic parity in the model's output, such as encoding societal stereotypes/biases, and (3) performance disparity across subgroups. This breakdown follows earlier works in the literature, such as [88]. We provide detailed results and discuss some of the limitations of this analysis in this section.

Toxicity/Profanity. Tables 26 and 27 provide a detailed breakdown of toxicity/profanity results for all subgroups in FairFace dataset. In Tables 28 and 29, we report similar results in the MIAP [89] dataset, disaggregated by perceived gender and age.

Demographic Parity. To estimate the level of demographic parity (DP) in the model's output, we feed an image into PaLI-X with the chosen occupation title as a prefix and record the average log-perplexity score of the captions generated by the model. To ensure that any observed parity would likely reflect unintended biases in the model itself as opposed to the evaluation dataset, we use CelebA [90] that contains celebrity images with gender presentation annotation. Our assumption is that many occupations reflecting societal stereotypes, such as secretaries and plumbers, are quite rare in the CelebA dataset so disparities in output may reflect what is encoded in the model itself.

Figure 9 (TOP) summarizes the overall results. First, PaLI-X tends to assign a higher log-perplexity score to women than men across most occupations; i.e. men are predicted to be more likely to hold such occupations. Second, PaLI-X assigns a higher likelihood for a woman to be ('secretary' & 'actor') and a higher likelihood for a man to be ('guard' & 'plumber') at the 95% confidence level. Figure 9 (BOTTOM) displays the corresponding correlations between perceived gender presentation and occupations within the WebLI dataset, where we use the Pearson correlation coefficient by treating each label as a binary random variable and noting that for binary random variables, zero correlation implies full independence. All absolute correlation coefficients in the data are < 0.2 with 99% of them being < 0.1. The list of occupations is compiled based on [91] and the US job statistics report in [92].

Performance Disparity. See Section 5 and Table 9 for a comparison of how well PaLI-X performs across different

Table 26. Distribution of the predicted toxicity/profanity for the captions generated by PaLI-X on FairFace dataset disaggregated by ethnicity.

Ethnicity	Toxicity			Profanity		
	< 0.2	$0.2 - 0.8$	> 0.8	< 0.2	$0.2 - 0.8$	> 0.8
Middle Eastern	64.24%	35.76%	0.00%	94.87%	5.13%	0.00%
Black	59.47%	40.40%	0.13%	92.67%	7.33%	0.00%
Indian	63.86%	36.07%	0.07%	94.39%	5.61%	0.00%
Hispanic	61.09%	38.79%	0.12%	94.45%	5.55%	0.00%
White	62.45%	37.16%	0.39%	92.85%	7.10%	0.05%
Southeast Asian	63.18%	36.61%	0.21%	93.57%	6.43%	0.00%
East Asian	63.15%	36.72%	0.13%	91.55%	8.45%	0.00%

Table 27. Distribution of the predicted toxicity/profanity for the captions generated by PaLI-X on FairFace dataset disaggregated by age.

Age	Toxicity			Profanity		
	< 0.2	$0.2 - 0.8$	> 0.8	< 0.2	$0.2 - 0.8$	> 0.8
< 19	58.78%	40.00%	0.22%	89.71%	10.29%	0.00%
20 - 29	63.01%	36.86%	0.12%	93.24%	6.73%	0.03%
30 - 39	63.13%	36.70%	0.17%	95.41%	4.59%	0.00%
40 - 49	63.62%	36.31%	0.07%	95.27%	4.73%	0.00%
50 - 59	65.87%	33.88%	0.25%	96.48%	3.52%	0.00%
60 - 69	65.31%	34.38%	0.31%	95.95%	4.05%	0.00%
> 70	66.10%	33.90%	0.00%	92.37%	7.63%	0.00%

Table 28. Distribution of the predicted toxicity/profanity for the captions generated by PaLI-X on MIAP dataset disaggregated by perceived gender.

Perceived Gender	Toxicity			Profanity		
	< 0.2	$0.2 - 0.8$	> 0.8	< 0.2	$0.2 - 0.8$	> 0.8
Predominantly Feminine	53.98%	45.93%	0.09%	90.55%	9.39%	0.07%
Predominantly Masculine	70.76%	29.17%	0.06%	94.97%	5.01%	0.01%

Table 29. Distribution of the predicted toxicity/profanity for the captions generated by PaLI-X on MIAP dataset disaggregated by age bucket.

Age Bucket	Toxicity			Profanity		
	< 0.2	$0.2 - 0.8$	> 0.8	< 0.2	$0.2 - 0.8$	> 0.8
0-2 yrs	28.00%	72.00%	0.00%	69.90%	30.10%	0.00%
3-19 yrs	49.96%	49.96%	0.07%	91.46%	8.54%	0.00%
20-59 yrs	66.27%	33.68%	0.05%	93.42%	6.55%	0.03%
> 60 yrs	65.46%	34.54%	0.00%	96.39%	3.61%	0.00%

subgroups in a VQA task, constructed from the FairFace dataset. We present here a different evaluation using the MIAP [89] dataset. For images containing exactly a single individual, we query PaLI-X with the question: "Is there a person in this image?" and evaluate the accuracy of its response. Note that there are no false positives in this evaluation. Table 30 summarizes the results. We observe that PaLI-X maintains a high accuracy across all subgroups.

Limitations. The analysis carried out in this section is necessarily limited, since fairness is a societal concept that cannot be reduced to statistical metrics. We expect RAI evaluations to evolve over time as new issues are detected and reported in the literature and additional datasets become available. Statistical analysis is only a single step and does not substitute for studying the broad and delayed impact of deployed models.

In addition, we rely in some parts on automated tools for

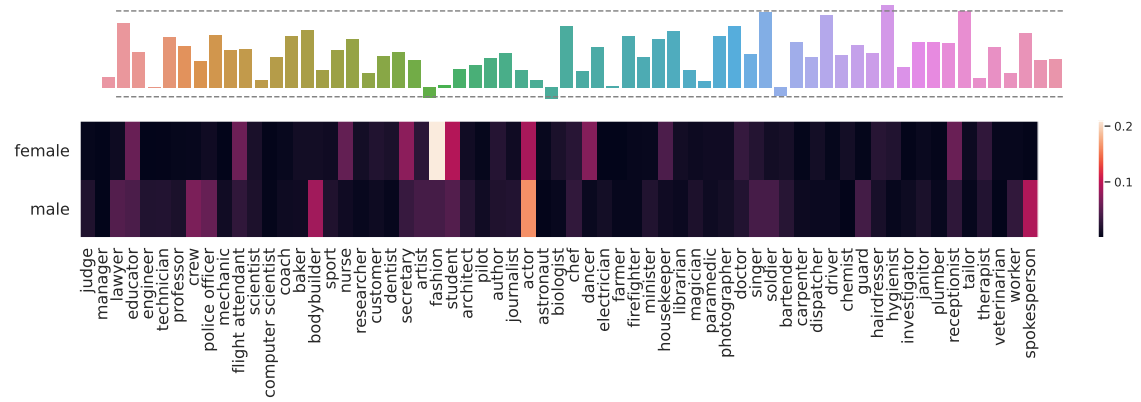

Figure 9. TOP: Level of demographic parity (DP) in PaLI-X's output for CelebA images between women and men. Values close to zero indicate absence of bias. BOTTOM: *Absolute* Pearson correlation coefficients between gender presentation and occupations in WebLI.

Skin Tone	**1** [2]	**2** [871]	**3** [3008]	**4** [522]	**5** [184]	**6** [85]	**7** [54]	**8** [49]	**9** [6]	**10** [1]
	0.00%	0.11%	0.47%	1.53%	0.54%	1.18%	0.00%	0.00%	0.00%	0.00%

Gender	**Predominantly Feminine** [2437]	**Predominantly Masculine** [3544]
	0.53%	0.85%

Age Bucket	**0-2 yrs** [17]	**3-19 yrs** [568]	**20-59 yrs** [4925]	**> 60 yrs** [247]
	0.00%	0.00%	0.77%	0.81%

Table 30. Detection error rate for "person" in PaLI-X using the subset of the MIAP dataset [89] that contain exactly a single individual in the image. PaLI-X maintains a low error rate across all subgroups. Skin tone follows the Monk Skin Tone Scale [93]. Numbers inside square brackets correspond to the size of each bucket.

inferring attributes, which are not perfectly accurate and can lead to a broad categorization of people that misidentifies real identities. We do not support the creation or application of classifiers for sensitive attributes, such as gender or ethnicity, based on visual indicators and encourage readers to delve into the comprehensive work outlining their potential risks, such as [94, 95], for further insight. Also, while we use perceived gender presentation in our analysis that is provided by the data (i.e. in CelebA and FairFace), we acknowledge that people may express their gendered identities in numerous other ways.

In our evaluation, toxicity is predicted based on the generated captions only. However, without knowing the context of the image, this can introduce false positives.