This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

On Scaling up a Multilingual Vision and Language Model

Xi Chen, Josip Djolonga°, Piotr Padlewski°, Basil Mustafa°, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, Hexiang Hu, Mandar Joshi, Bo Pang, Ceslee Montgomery, Paulina Pietrzyk, Marvin Ritter, AJ Piergiovanni, Matthias Minderer, Filip Pavetic, Austin Waters, Gang Li, Ibrahim Alabdulmohsin, Lucas Beyer, Julien Amelot, Kenton Lee, Andreas Peter Steiner, Yang Li, Daniel Keysers, Anurag Arnab, Yuanzhong Xu, Keran Rong, Alexander Kolesnikov, Mojtaba Seyedhosseini, Anelia Angelova†, Xiaohua Zhai†, Neil Houlsby†, Radu Soricut†
Google

pali-communications@google.com °:Significant technical contribution †: Project lead

Abstract

We explore the boundaries of scaling up a multilingual vision and language model, both in terms of size of the components and the breadth of its training task mixture. Our model achieves new levels of performance on a wide-range of varied and complex tasks, including multiple image-based captioning and question-answering tasks, image-based document understanding and few-shot (in-context) learning, as well as object detection, video question answering, and video captioning. Our model advances the state-of-the-art on most vision-and-language benchmarks considered (20+ of them). Finally, we observe emerging capabilities, such as complex counting and multilingual object detection, tasks that are not explicitly in the training mix.

1. Introduction

The success of scaling language models [1–4] makes it appealing to similarly scale Vision-Language (V&L) models, and investigate the improvements, capabilities, and emergent properties of such models. Inspired by the work in [5], we present PaLI-X, a multilingual vision and language model with reusable scaled-up components, consisting of a pretrained large-capacity visual encoder (using [6] as the starting point) and a pretrained language-only encoder-decoder (using [7] as the starting point), further trained at-scale on a vision-and-language data mixture using a combination of self-supervision and full-supervision signals.

One clear pattern that emerges from the combination of results from PaLI [5] and the work we present in this paper is that scaling *both* V&L components together brings increases in performance across a wide range of tasks. We show this by comparing against the same benchmarks used for PaLI (Fig. 1, Left), and also against new benchmarks for which the new capabilities of PaLI-X are evaluated (e.g., ChartQA, AI2D, DocVQA, InfographicVQA, as well as video understanding tasks). We observe that scaling leads to large improvements over the results of the PaLI model, and also over specialized large-scale models that are trained specifically to solve certain tasks, often with the help of (often much larger) text-only LLMs [8]. In particular, we find that increasing both the effective capacity of the vision component (which [9] does more unilaterally) and of the language component (which [10] also does unilaterally) is beneficial; the new PaLI-X model provides more balanced parameter allocation than any other prior work (roughly 40%-60% split of the total capacity).

Aside from demonstrating the consistent impact of scale, the original contribution of PaLI-X consists in leveraging the mixture-of-objectives proposed in [7] for vision-and-language modeling, and showing that it results in a model that improves both state-of-the-art results and the Pareto frontier for fine-tuning and few-shot (Fig. 1, Right).

We also observe emergent properties based on PaLI-X's results compared to previous models with similar architecture but smaller sizes. For instance, we report drastically improved performance on the counting ability (See Table 1 and Appendix B), both for the plain variety (count all instances of a class) and the complex variety (count instances based on a natural language description), that are not attributable to training design[1]. Additionally, we present qualitative insights into the model's performance (Appendix A), with an

[1]Plain counting is usually achievable via good object detection, while complex counting requires a fine-grained understanding of the alignment between language-based specifications and visually-based occurrences.

emphasis on multilingual transfer learning such as the ability to detect objects using non-English labels (Fig. 2), and the ability to switch between the language of text present in the image (e.g., English) and the language of the generated image caption (e.g., Romanian).

Our technical contributions include the following:

1. We scale a Vision-Language model to achieve outstanding performance on a wide variety of benchmarks. We observe that scaling *both* the Vision & Language components is advantageous and report that performance continues to consistently benefit from scale beyond 50B.
2. While larger scales are clearly beneficial, we show that, how to train the model is equally important . Specifically it is key to use a mixture of objectives that combines prefix-completion and masked-token completion, which improves the Pareto frontier for fine-tuning vs few-shot performance at this scale.
3. We show that continuing co-training a high-capacity vision encoder (ViT-22B) with image classification and OCR label classification[2] can gain significant improvements on V&L tasks for which the understanding of text-within-image is crucial.
4. Overall, our PaLI-X model improves SoTA results on 20+ benchmarks, and we show that it is the first of its kind to simultaneously adapt via multitask fine-tuning to a diverse set of benchmarks without significant performance degradation. This, along with our observation of the multimodal emergent property around counting and object detection, demonstrates the generalizability of PaLI-X.

2. Related Work

Similar to large language models such as GPT4 [12] and PaLM [1], the benefit of scale has also been observed in recent vision and vision-language models. Flamingo [10] used a frozen language component and demonstrated the benefit of scaling up this part up to 70B parameters on the few-shot multimodal capabilities, while the vision encoder is fixed with 435M parameters. GIT [9], on the other hand, explored scaling of the vision component up to 4.8B parameter, with a 300M parameter language decoder. PaLI [5] explored jointly scaling the vision and language component, to 4B and 17B, respectively, and showed that scaling both components benefits a wide range of vision-language tasks. All these models took advantage of vision and language unimodal pretrained models as backbones to start multimodal training. Recently, on the vision model side, a vision transformer with 22B parameter has been introduced [6]. In this work, we make use of a ViT-22B model specifically tuned for OCR capability to explore scaling Vision-Language models to even larger parameter regime.

As first shown in [13], *large* language models are sometimes able to solve new unseen tasks at inference as long as a few examples –or *shots*– are provided as inputs. This is usually referred to as in-context learning [14]. Follow-up work proposed improved ways to split and prompt the shots, such as Chain of Thought [15] or Least-to-Most prompting [16]. So far, the vast majority of this work has been done in the context of language inputs [17]. In this work, we explore multimodal in-context learning with pairs of images and captions. Our work is aligned in spirit to Flamingo [10] that uses interleaved image text pairs in the same web page and in-context tuning [18] during pre-training. We first group the image-text pairs by url and split each group to a "shots" set and a "target" set. Then we use the few examples in the "shots" set as input features to predict the examples in the target set.

Besides solving vision-language tasks in multiple domains, recent VLMs also attempted solving these tasks at once instead of fine-tuning on each individual benchmark. Unified-IO [19] performed multitask fine-tuning and reported solid results across 16 benchmarks. Spotlight [20] reported that inside the UI domain, multitask fine-tuning can achieve a performance close to task-specific fine-tuning. In this work, we show that PaLI-X can be simultaneously fine-tuned with a diverse set of benchmarks in multiple domains without performance degradation.

3. Model

3.1. Architecture

The PaLI-X model architecture follows the encoder-decoder architecture: image(s) are processed by a ViT encoder, with the resulting visual embeddings fed to an encoder-decoder backbone, along with embeddings from additional text input (e.g., question / prefix / prompt). More details are provided in Appendix A.

Visual component Our visual backbone is scaled to 22B parameters, as introduced by [6], the largest dense ViT model to date. To equip the model with a variety of complex vision-language tasks, we specifically focus on its OCR capabilities. To that end, we incorporate an OCR-based pretraining as follows: images from the WebLI dataset [5] are annotated with OCR-text detected by GCP Vision API; the encoder is then further pre-trained with a mixture of the original JFT-based classification task and a new OCR-based classification task (whether or not a given token occurred in the image according to OCR results). See Appendix A for additional details on the visual component. PaLI-X is designed to take $n >= 1$ images as inputs (for few-shot and video understanding), with tasks involving a single image as the $n = 1$ case. For $n > 1$, each image is independently processed by the ViT module, and the patch-level embeddings coming out of ViT are flattened and concatenated to

[2] We use OCR tokens produced by the GCP Vision API over the training images as targets.

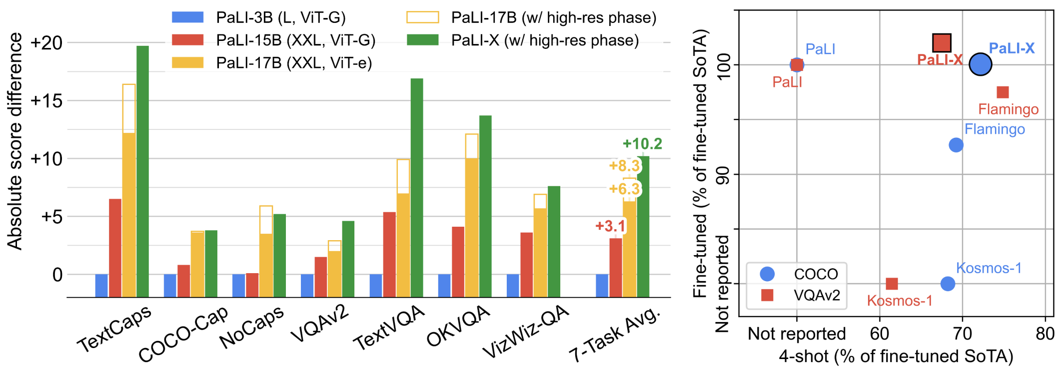

Figure 1. [Left] Comparing PaLI-X against PaLI on image-captioning and VQA benchmarks. [Right] The Pareto frontier between few-shot and fine-tuned performance, comparing PaLI-X with PaLI [5], Flamingo [10], and Kosmos-1 [11].

form the visual input (See Appendix A). Note that similar to the single-image case, there is no pooling over the spatial dimension before visual embeddings are aggregated over the temporal dimension. That is, for an n-frame input with k-patches per frame, the resulting visual input has $n * k$ tokens.

Overall model The encoder-decoder backbone is initialized from a variant of the UL2 [7] encoder-decoder model that uses 32B parameters. The architecture of this variant has 50 layers in both encoder and decoder (up from 32 layers in [7]), and is pretrained on a mixture of text data similar to [7]. The visual embeddings, after going through a projection layer, are concatenated with the token embeddings of the text input, and fed to the encoder-decoder backbone. Most of the pretraining tasks (with the exception of the masked image token task) predict text-only output from this multimodal input. The text input to the model typically consists of a prompt that marks what type of task it is (e.g., "*Generate caption in* ⟨lang⟩" for captioning tasks) and encode necessary textual input for the task (e.g., "*Answer in* ⟨lang⟩*: {question}*" for VQA tasks). For tasks that need OCR capabilities, we experiment with either relying solely on the text-encoding capabilities of the vision encoder, or optionally including tokens extracted by an upstream OCR system fed as additional text inputs.

Few-shot formulation In the few-shot setting, for a given *target example* the model receives a number of "labeled" examples (in the form of additional ⟨image, text⟩ pairs) that we refer to as *shots/exemplars*. The hypothesis is that information contained in these exemplars provides the model with useful context to generate predictions for the target example. Formally, the input with N shots is a sequence $(t_1, \ldots, t_N, t_T, i_1, \ldots, i_N, i_T)$, where $t_1 : t_N$ and $i_1 : i_N$ are texts and images for the N shots, and t_T and i_T are the text (prompt) and image for the target example. PaLI-X processes this input as follows: all images, including the target one, are first independently processed by the visual encoder, and the resulting patch-level embeddings are flattened and concatenated to form the visual input sequence. After going through a projection layer, they are concatenated with the text embeddings to form the multimodal input sequence used by the encoder. We implement additional optimizations including distributing the exemplars between the encoder and the decoder, and an attention re-weighting mechanism (see Appendix B).

3.2. Pretraining Data and Mixture

The main pretraining data for our model is based on WebLI [5], consisting of roughly one billion images with alt-texts from the web and OCR annotations (using the GCP Vision API), covering over 100 languages. In addition to WebLI ⟨image, text⟩ pairs, we introduce here *Episodic WebLI* data, where each episode corresponds to a set of such pairs. We aim to have each episode contain loosely related images (i.e., they are clustered according to their URL field), so as to encourage attention among examples in an "episode". In training, we sample 5 images and the alt_text from each episodic example; the first 4 images are used as context, and the alt_text of the 5th image as the target. We find this new dataset (with 75M episodes and around 400M images in total) important for developing the few-shot capabilities of the model.

The pretraining mixture consists of the following data and objectives: (i) span corruption on text-only data (15% of tokens); (ii) split-captioning on WebLI alt-text data [5, 21]; (iii) captioning on CC3M [22] on native and translated alt-text data (over the same 35 languages covered by XM3600 [23]); (iv) split-ocr [24] on WebLI OCR annotations; (v) visual-question-answering objective over ⟨image, question, answer⟩ pairs generated using the VQ^{2}A method [25] over the CoCo-Captions training data, over native and translated text (same 35 language pairs); (vi) visual-question-generation objective, using the same pairs as above; (vii) visual-question-answering objective over ⟨image, question, answer⟩ pairs using the Object-Aware method [26] (English only); (viii) captioning on Episodic WebLI examples (target alt-text predicted from the remaining alt-text and images); (ix) visual-question-answering on 4-pair examples (resembling Episodic WebLI and using VQ^2A-CC3M pairs), with the answer target conditioned on the other pairs of ⟨image, question, answer⟩ data. (x) pix2struct objective, introduced in [27], targeting page layout and structure using screenshot images paired with DOM-tree representations of html pages. (xi) split-captioning on short video data, using the VTP data [10] (using four frames per video). (xii) object-detection objective on WebLI data, whereby an OWL-ViT model [28] (L/14) is used to annotate WebLI images, resulting in hundreds of pseudo object labels and bounding boxes per image. (xiii) image-token prediction objective, whereby we tokenize WebLI images (256×256 resolution) using a ViT-VQGAN [29] model with patch size 16×16 (256 tokens per image); this objective is framed as a 2D masked-token task (i.e., fill-in the missing grid pieces, with the corresponding image pixels also masked). Note that the image-token prediction objective is added mainly as a condition to check whether it adversarially impacts the performance on language-output tasks; our ablation experiments show that is does not. When assembling the mixture, our rule of thumb was to avoid training on a huge chunk of data for two times. Thus, for the larger datasets, we mix them together with weight proportional to the number of examples in the corresponding dataset. For the smaller datasets, we mix them in with up to two epochs based on empirical evidence or heuristics. We note here that other mixing ratios are also possible in order to achieve similar performance. We performed similarity-based deduplications to remove image from the pretraining mix that are identical or similar to those in the evaluation benchmarks combined, following [5].

3.3. Training Stages

Our model is trained in two stages. In stage 1, the visual encoder (after mixed-objective training) is kept frozen, while the rest of the parameters are trained on a total of 2.2B examples at the base resolution 224×224 (native to ViT-22B), using the entire mixture. In stage 2, it continues training using only the OCR-related objectives (pix2struct and split-ocr) plus the object detection objective; this is done in several sub-stages, during which image resolution is gradually increased to 448×448, 672×672 and finally 756×756.

4. Experiments

4.1. Image Captioning and VQA

Our results demonstrate that the larger capacity in PaLI-X scales well in both its vision and language components, and it is particularly beneficial for more challenging scene-text and document understanding tasks. Our model outperforms the SOTA on diverse vision-language tasks, with significant margins in some cases. The Image Captioning and VQA benchmarks used for evaluation are summarized in Appendix B, including 6 Image Captioning benchmarks (COCO (Karpathy split [30]), NoCaps [31], TextCaps [32], VizWizCap [33], Screen2Words [34], Widget-Cap [35]) and 13 VQA benchmarks. These tasks span a wide range of visual domains, from natural images, illustrations to documents and user interfaces (UIs). We also include results of multilingual captioning on XM3600 in Appendix B.

4.1.1 Per-task fine-tuning results

Experimental setup We fine-tune PaLI-X with frozen ViT-22B; the learning rate follows a linear decay from initial value 1e-4 for all fine-tuning experiments. See Appendix B for more details.

First, we present benchmarks results for the condition where external OCR systems are not used (Table 1, see Appendix B for an extended table.). The trend is that PaLI-X matches or improves SoTA results on these benchmarks, with a particularly significant improvement on the TallyQA benchmark over MoVie [49] (specialized counting model), at +11.1 for simple counting questions (e.g., "how many giraffes") and +18.8 for complex counting questions (e.g., "how many giraffes are drinking water"); there are significant improvements over PaLI [5] as well, indicating that scale plays an important role in the ability of such models to perform counting tasks. We additionally note the state-of-the-art result on VQAv2 at 86.1 accuracy, achieved with an open-vocabulary generative approach, and the performance on OKVQA at 66.1 accuracy, matching the much-larger PaLM-E [37] model performance.

Next, we examine text-heavy V&L benchmarks, for which upstream OCR systems can be used to improve performance. As shown in Table 2, PaLI-X improves SoTA for all Captioning and VQA benchmarks across the board, either without or with additional OCR input (using GCP Vision API). For instance, a significant jump of +42.9 points is observed on AI2D[3], a multiple-choice benchmark where choices are provided along with each question. Being able to have the text choices as input benefits PaLI-X compared with the previous SoTA Pix2Struct [27] which has to render the text on the image, but this does not explain all the improvements. In a question-only configuration (no answer choice present), PaLI-X achieves 46.3 on AI2D, more than 4 points higher than Pix2Struct's result.

In general, having access to OCR texts extracted by an external OCR pipeline boosts performance. Still, for several benchmarks (e.g., AI2D, ChartQA, OCRVQA and Widget-

[3]As with all the other benchmarks, our training examples are carefully deduped to exclude images occurring in these benchmarks, including AI2D. Such results, therefore, are *not* attributable to train-test data leakage.

	COCO	NoCaps		VQAv2		OKVQA	TallyQA	
Model	Karp.-test	val	test	test-dev	test-std	val	simple	complex
GIT2 [9] (5.1B)	145.0	126.9	**124.8**	81.74	81.92	-	-	-
Flamingo [10] (80B)	138.1	-	-	82.0	82.1	57.8*	-	-
BEiT-3 [36] (1.9B)	147.6	-	-	84.2	84.0	-	-	-
PaLM-E [37] (562B)	138.7	-	-	80.0	-	**66.1** -	-	
MoVie	-	-	-	69.26	-	-	74.9	56.8
PaLI [5](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 1. Results on COCO Captions (Karpathy split), NoCaps, VQAv2 [38], OKVQA [39], and TallyQA [40] with end-to-end modeling without OCR pipeline input ("simple" and "complex" are test subsplits).

Model	Text Caps	VizWiz Cap	Text VQA	VizWiz VQA	ST VQA	OCR VQA	Info VQA	Doc VQA	AI2D	Chart QA	Screen2 Words	Widget Cap	OVEN	Info Seek
***with** OCR pipeline input*														
SoTA	160.4 [5]	124.7 [5]	73.67 [41]	73.3 [5]	79.9 [5]	67.5 [42]	47.4 [43]	**87.8** [44]	38.5 [45]	45.5 [46]	- -	- -	- -	- -
PaLI-X	**163.7**	**125.7**	**80.78**	**74.6**	**84.5**	**77.3**	**54.8**	86.8	**81.4**	**72.3**	-	-	-	-
***without** OCR pipeline input*														
SoTA	145.0 [9]	120.8 [9]	67.27 [9]	70.7 [5]	75.8 [9]	71.3 [27]	40.0 [27]	76.6 [27]	42.1 [27]	70.5 [8]	109.4 [27]	141.8 [20]	31.6 [47]	8.2 [48]
PaLI-X	**147.0**	**122.7**	**71.44**	**70.9**	**79.9**	**75.0**	**49.2**	**80.0**	**81.2**	**70.9**	**127.9**	**153.0**	**38.3**	**10.8**

Table 2. Results on benchmarks more focused on text understanding capabilities. For OVEN [47] & InfoSeek [48], we employ the 224×224 resolution setup for fair comparison (on human split).

Cap), PaLI-X's end-to-end performance when using its intrinsic OCR capability is close to that leveraging additional OCR input. A common feature for these benchmarks is that they have well-oriented text – diagrams, charts, book covers or user interfaces, with reasonably large font size at 756×756 resolution. For tasks involving scene text in natural images (TextCaps, TextVQA, STVQA) or very high density of small texts (DocVQA, InfoVQA), results still highlight clear benefits when utilizing an external OCR model.

4.1.2 Multitask Fine-tuning

We simultaneously fine-tune and evaluate the pretrained checkpoints on multiple benchmarks belonging to the same category. We deduplicated every training set over the test sets of every task in the mixture to prevent the leakage of any test-set examples into the mixed training set. This is useful as it leads to a single fine-tuned model that performs all the tasks, rather than having to fine-tune each task separately. We performed such multitask fine-tuning on all Image Captioning benchmarks and all VQA benchmarks, respectively.

Table 3 shows the multitask fine-tuning result for captioning tasks. The performance on COCO is slightly decreased in the multitask setting, which is likely a result of this task needing longer training to converge. For Screen2Words, having the smallest train and dev/test sets could be responsible for the performance fluctuation. Notably, VizWiz-Cap and Widget-Cap shows improved performance from multitask fine-tuning. Overall, the average performance decreases by 1.4 points (0.2 excluding Screen2Words) with multitask fine-tuning, while offering the clear advantage of having a single checkpoint to perform all these tasks. Appendix B shows similar results for VQA tasks. We consider this outcome a positive result that establishes the on-par performance between multitask fine-tuning and single-task fine-tuning for diverse benchmarks, in contrast with previous work which argued a gap between single-task and multitask fine-tuning [19], or demonstrated little gap over benchmarks from the same domain [20].

4.1.3 Few-shot Evaluation

We fine-tuned the PaLI-X model on a mixture of few-shot tasks. The few-shot mixture contains Episodic mixtures, (Non-Episodic) Webli and (Non-Episodic) CC3M data. Note that all of these datasets were already used in previous stages of training, but with lower mixture proportions. During pre-training, we only use up to 4 shots, with both encoder and decoder shots (see Appendix B). For fine-tuning, we use up to 8 encoder shots and do not use decoder shots.

We evaluate the few-shot performance on COCO caption (Karpathy test split [30]), and XM3600 [23] datasets. For each task, we first create a "shots pool" with 256 examples that are randomly selected from the task's training set. As the XM3600 benchmark does not come with a training set, we use Google Translate API to enhance the COCO Karpathy

Method	COCO	NoCaps	Text Caps	VizWiz Cap	Screen2 Words	Widget Cap	Avg.
Split	Karp.-test	val	val	test-dev	test	test	-
SOTA (Single-task FT)	149.1	**127.0**	148.6	119.4	109.4	136.7	
PaLI-X Single-task FT	**149.2**	126.3	150.8	123.1	**127.9**	153.2	-
PaLI-X Multitask FT	147.3	125.6	**154.6**	**124.2**	120.6	**153.7**	-
Multitask (+/-)	-1.9	-0.7	+3.8	+1.1	-7.3*	+0.5	-1.4 (-0.2 w/o "*")

Table 3. Scores from multitask fine-tuning compared with those from single-task fine-tuning for Image Captioning. Validation or test-dev set numbers are reported for some tasks.

training set with captions in the 35 languages represented in XM3600. Then, for each test data point, we randomly pick N shots from the pool as the actual few-shot examples. Following [10], we also evaluate on 2 text-only shots settings where only the textual part of 2 randomly sampled few-shot examples are used.

Table 4 reports the few-shot captioning performance on English and multilingual captioning, as well as few-shot VQA performance on VQAv2. PaLI-X achieves SOTA few-shot results on COCO with both 4 shots and 32 shots; it outperforms previous SOTA by +4.4 CIDEr points for 4-shot, suggesting a strong ability to efficiently gather hints from few examples. We also report few-shot CIDEr scores averaged over 35 languages using XM3600, demonstrating PaLI-X's multilingual capabilities. Meanwhile, although PaLI-X also performs decently on VQAv2, the gap behind the SoTA Flamingo model [10] (which freezes the language backbone) may be the result of losing some of the few-shot text-only QA capability by fine-tuning the language backbone, which supports the hypothesis regarding the tension between few-shot and fine-tuning abilities.

4.2. Video Captioning and Question Answering

We fine-tune and evaluate the PaLI-X model on 4 video captioning (MSR-VTT [50], VATEX [51], ActivityNet Captions [52], Spoken Moments in Time [53]) and 3 video question answering benchmarks (NExT-QA [54], MSR-VTT-QA [55], ActivityNet-QA [56]). A brief description of each benchmark and clarifications on their usage are provided in Appendix C.

Experimental setup We fine-tune our model (with base resolution 224×224) for each task separately, use the validation split for early stopping, and report performance on the test split. We use a learning rate of 10^{-4} for all tasks, and do not adapt any hyperparameters for specific tasks. Frames are sampled using a fixed temporal stride for each dataset (determined based on the video length distribution in that dataset such that the product of the number of frames and stride is larger than the total number of frames for half of the videos), and we experimented with including up to 8 or 16 frames per video. We did not include pooling over the spatial dimension; embeddings for 16×16 patches per frame are provided as visual input to the multimodal encoder.

Results We report CIDEr score for the video captioning tasks. Video QA tasks are treated as open-ended generation tasks; we report full-string accuracy (for MSR-VTT-QA and ActivityNet-QA) and WUPS metrics (NExT-QA) in [54, 61]. As shown in Table 5, the 16-frames version has an edge over the 8-frame version, sometimes with a significant margin (e.g., close to a 6 point increase in CIDEr score for ActivityNet-Captions). More importantly, while PaLI-X pretraining was dominated by image-text tasks, we were able to achieve new SOTA performance for 4 tasks[4], and performed close to SOTA on MSR-VTT Captions and QA.

4.3. Image classification

To test image classification capabilities we fine-tuned PaLI-X and models from [5] on ImageNet [62] and evaluated 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 used the model from the first training stage (at resolution 224) and the one from the last training stage (at resolution 756). We used the same training hyperparameters for all of runs (selected without any hyperparameter tuning; mode details in Appendix D).

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 fine-tuning on ImageNet. PaLI-X achieves SOTA results for generative models on Imagenet, and other datasets. We also performed zero-shot evaluation for PaLI-X and the results can be found in Appendix D.

4.4. Object Detection

Object detection can be easily formulated in our model as shown in pix2seq [70], The dataset mix used for pre-training

[4]As noted in Table 5, current SOTA on NExT-QA for the open-ended QA task was achieved by Flamingo 32-shot, which had outperformed prior fine-tuning SOTA. To the best of our knowledge, PaLI-X performance on this task does outperform existing published fine-tuning performances, with the caveat that we do not have information on what Flamingo fine-tuning would have achieved on this task.

Method	COCO Captions 4 shots	COCO Captions 32 shots	XM3600 Cap. (35-lang avg.) 4 shots	XM3600 Cap. (35-lang avg.) 32 shots	VQAv2 4 shots	VQAv2 32 shots
Prev. SoTA [10]	103.2	113.8	N/A (53.6 w/ fine-tune [5])		**63.1**	**67.6**
PaLI-X	**107.6**	**114.5**	45.1	47.1	56.9	57.1

Table 4. Few-shot performance of the PaLI-X model (multilingual captioning for XM3600).

Method	MSR-VTT Cap. [50]	MSR-VTT QA [55]	Activity-Net Cap. [52]	Activity-Net QA [56]	VATEX Cap. [51]	SMIT Cap. [53]	NExT-QA QA [54]
Prior SOTA	**80.3**	**48.0**	52.5	44.7	94.0†	28.1‡	33.5§
	mPLUG2 [57]	mPLUG2 [57]	PDVC [58]	VINDLU [59]	GIT2 [9]	MV-GPT [60]	Flamingo 32shot [10]
PaLI-X (8fr)	74.6	46.9	49.0	48.4	66.0	42.5	37.0
PaLI-X (16fr)	76.8	47.1	**54.9**	**49.4**	69.3	**43.5**	**38.3**

Table 5. Results for Video Captioning and Video-QA using 8 frames (8fr) or 16 frames (16fr). †GIT2 uses Self-Critical Sequence Training to directly optimize the CIDEr metric for VATEX. ‡SMIT has not been used for video captioning before, we apply MV-GPT [60] and report results on the test set. §Numbers were obtained using 32-shot; since Flamingo 32-shot outperforms fine-tuning SOTA on this open-ended QA task, they did not conduct further fine-tuning experiments for this task.

Model (resolution)	INet [62]	REAL [63]	INet-R [64]	INet-A [65]	INet-Sketch [66]	INet-v2 [67]
GIT2 [9] (384)	**89.22**	-	-	-	-	-
PaLI-17B [5] (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)	**89.19**	**90.98**	**80.06**	**72.57**	**73.37**	**83.66**

Table 6. Classification accuracy (top-1) fine-tuned on Imagenet [62].

Credits: Watermelon/Cat; Sarah Pflug (burst), Bowls; ariesandrea (flickr), Wall; Matthew Henry (burst)

Figure 2. Examples demonstrating multilingual, OCR and other capabilities transferred to detection.

	LVIS AP	LVIS AP_{Rare}
ViLD [68]†	29.3	26.3
Region-CLIP [69]†	32.3	22.0
OwLViT-L/16 [28]†	34.7	25.6
OwLViT-L/16 [28]‡	34.6	31.2
PaLI-X (Zeroshot)	12.36	12.16
PaLI-X (Detection-tuned)	30.64	31.42

Table 7. PaLI-X object detection results on LVIS. The diverse pre-training mix enables parity performance between LVIS rare and common classes. Other related approaches are shown for context, but are not directly comparable. †: tuned on non-rare LVIS. ‡ training set further includes Object365 and Visual Genome

is presented in Sec. 3; detection data was included up to and including the stage using resolution 672, after which a separate detection-specific model was fine-tuned on detection data. Before detection-specific tuning, LVIS [71] & COCO labels were removed from all detection training datasets, allowing zero-shot evaluation on LVIS.

Bounding box mean AP on LVIS is shown in Table 7, including zero-shot performance; the detection-tuned model reaches an AP of 31 in general, and 31.4 on rare classes, and about 12 for both in zero-shot. Performance on rare classes was on par with performance on common classes, a difficult feat traditionally accomplished by complicated sampling schedules and augmentations. In our set up, it is directly enabled by PaLI-X's diverse training mix. This could likely be further improved with investment in fine-tuning e.g. using noise-augmentation methods from pix2seq [70], or a further stage of high-resolution, LVIS only training. Qualitatively, we observe emergence of many interesting phenomena enabled by co-training with non-detection tasks; for example, multilingual detection, OCR bounding boxes and longer descriptions, none of which are included in detection training, are often handled well by PaLI-X. Additional results and information can be found in Appendix E.3.

	Gender		Ethnicity			Age			
	Lowest	Highest	Lowest	Median	Highest	Lowest	Median	Highest	**Overall**
Toxicity	0.14%	0.19%	0.00%	0.13%	0.39%	0.00%	0.17%	0.31%	**0.01%**
Profanity	0.00%	0.02%	0.00%	0.00%	0.05%	0.00%	0.00%	0.03%	**0.00%**

Table 8. Average toxicity/profanity in the captions generated by PaLI-X on FairFace dataset.

Ethnicity	*White*	*Hispanic*	*Southeast Asian*	*East Asian*	*Indian*	*Middle Eastern*	*Black*
Precision	0.956	0.827	0.907	0.943	0.952	0.957	0.859
Recall	0.836	0.786	0.753	0.827	0.909	0.943	0.792
Age	*0-19*	*20-29*	*30-39*	*40-49*	*50-59*	*60-69*	*>70*
Precision	-	0.887	0.940	0.938	1.000	1.000	1.000
Recall	-	0.880	0.840	0.792	0.868	0.761	1.000

Table 9. Precision and recall of PaLI-X on the task of predicting the gender presentation attribute for the FairFace dataset. Results are disaggregated by ethnicity and age (gender of minors not included).

5. Model Fairness, Biases, and Other Potential Issues

Large models, if left unchecked, have the potential to inflict harm on society – such as amplifying biases [72–75], causing disparities [74, 76, 77], or encoding narrow cultural perspectives [78, 79]. Hence, evaluating PaLI-X for such potential issues is important.

Toxicity/profanity. We estimate the level of toxicity and profanity in the generated captions, including when disaggregated across subgroups. We use the FairFace dataset [80] that comprises of images of people with ground-truth attributes: gender presentation, age and ethnicity. We generate captions and use the Perspective API [81] (threshold > 0.8) to measure toxicity and profanity. Table 8 summarizes the results; we observe a low level of toxicity/profanity across all slices. Appendix F provides a more detailed breakdown.

Bias. We estimate the level of demographic parity (DP) [82] in PaLI-X with respect to gender and occupation. Overall, 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. See Appendix F for a visualization and further details.

Performance Disparity. We compare how well PaLI-X performs across different subgroups in a VQA task. Since an analysis of this sort requires ground-truth annotations of protected attributes, we use FairFace dataset where the task is to predict the gender presentation attribute provided in the dataset given the image. Table 9 reports the disaggregated precision & recall. We observe a lower performance for Hispanics and Blacks compared to others, possibly because they are under-represented in the data.

Limitations. See Appendix F for a discussion about some of the limitations of this analysis.

6. Conclusions

In this work we draw more insights from further scaling vision and language models. We show that the scaling and the improved training recipe results in a model that substantially outperforms previous state-of-the-art models, leads to emergent behaviors and identifies further margins for improvements. In particular, we report that the model achieves significant improvements at document, chart, and infographic understanding, captioning, visual question answering, counting, and performs well on few-shot (in-context) captioning, video captioning and question-answering, and object detection.

7. Acknowledgements

We would like to thank Sarah Laszlo, Kathy Meier-Hellstern, Caroline Pantofaru, Susanna Ricco, Candice Schumann, Ken Burke, Simon Wang, Rachel Hornung, Yichang Chen, Utsav Prabhu, Abhijit Ogale, Kristina Toutanova, Weicheng Kuo, Jihyung Kil, Xiangning Chen, Liang Chen, Rich Lee, Elizabeth Adkison, James Cockerille, Eric Ni, Erica Moreira, Victor Gomes, Jeremiah Harmsen, Claire Cui, Slav Petrov, Tania Bedrax-Weiss, Joelle Barral, Tom Duerig, Paul Natsev, Fernando Pereira, Jeff Dean, and Zoubin Ghahramani for helpful discussions, feedback, and support.

References

[1] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 1, 2

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Jared Kaplan Melanie Subbiah, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Christopher Hesse Clemens Winter, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *NeurIPS*, 2020.

[3] Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten P Bosma, Zongwei Zhou, Tao Wang, Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathleen Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. GLaM: Efficient scaling of language models with mixture-of-experts. In *ICML*, 2022.

[4] PaLM 2 tech report. https://ai.google/static/documents/palm2techreport.pdf. (Accessed on 05/15/2023). 1

[5] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Alexander Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. PaLI: A jointly-scaled multilingual language-image model. In *ICLR*, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[6] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd van Steenkiste, Gamaleldin F. Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Patrick Collier, Alexey Gritsenko, Vighnesh Birodkar, Cristina Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetić, Dustin Tran, Thomas Kipf, Mario Lučić, Xiaohua Zhai, Daniel Keysers, Jeremiah Harmsen, and Neil Houlsby. Scaling vision transformers to 22 billion parameters. In *ICML*, 2023. 1, 2

[7] Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. UL2: Unifying language learning paradigms. In *ICLR*, 2023. 1, 3

[8] Fangyu Liu, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. DePlot: One-shot visual language reasoning by plot-to-table translation. *arXiv preprint arXiv:2212.10505*, 2022. 1, 5

[9] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. GIT: A generative image-to-text transformer for vision and language. *TMLR*, 2022. 1, 2, 5, 6, 7, 8

[10] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[11] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023. 3

[12] https://openai.com/product/gpt-4. 2023. 2

[13] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2

[14] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*, 2022. 2

[15] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022. 2

[16] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-

to-most prompting enables complex reasoning in large language models. In *ICLR*, 2023. 2

[17] Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023. 2

[18] Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. Meta-learning via language model in-context tuning. In *ACL*, 2022. 2

[19] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-IO: A unified model for vision, language, and multi-modal tasks. In *ICLR*, 2023. 2, 5

[20] Gang Li and Yang Li. Spotlight: Mobile UI understanding using vision-language models with a focus. In *ICLR*, 2023. 2, 5

[21] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. SimVLM: Simple visual language model pretraining with weak supervision. In *ICLR*, 2022. 3

[22] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 3

[23] Ashish V. Thapliyal, Jordi Pont-Tuset, Xi Chen, and Radu Soricut. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In *EMNLP*, 2022. 3, 5, 1

[24] Jihyung Kil, Soravit Changpinyo, Xi Chen, Hexiang Hu, Sebastian Goodman, Wei-Lun Chao, and Radu Soricut. PreSTU: Pre-training for scene-text understanding. *arXiv preprint arXiv:2209.05534*, 2022. 3

[25] Soravit Changpinyo, Doron Kukliansky, Idan Szpektor, Xi Chen, Nan Ding, and Radu Soricut. All you may need for VQA are image captions. In *NAACL*, 2022. 3

[26] AJ Piergiovanni, Weicheng Kuo, and Anelia Angelova. Pre-training image-language transformers for open-vocabulary tasks. In *T4V: Transformers for Vision Workshop, Conference on Computer Vision and Pattern Recognition*, 2022. 4

[27] Kenton Lee, Mandar Joshi, Iulia Turc, Hexiang Hu, Fangyu Liu, Julian Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2Struct: Screenshot parsing as pretraining for visual language understanding. In *ICML*, 2023. 4, 5

[28] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection with vision transformers. In *ECCV*, 2022. 4, 7, 10

[29] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved VQGAN. In *ICLR*, 2022. 4

[30] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 4, 5

[31] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8948–8957, 2019. 4

[32] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. TextCaps: a dataset for image captioning with reading comprehension. In *European conference on computer vision*, pages 742–758, 2020. 4

[33] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 417–434. Springer, 2020. 4

[34] Bryan Wang, Gang Li, Xin Zhou, Zhourong Chen, Tovi Grossman, and Yang Li. Screen2Words: Automatic mobile ui summarization with multimodal learning. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, UIST '21, page 498–510, New York, NY, USA, 2021. Association for Computing Machinery. 4

[35] Yang Li, Gang Li, Luheng He, Jingjie Zheng, Hong Li, and Zhiwei Guan. Widget Captioning: Generating natural language description for mobile user interface elements. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5495–5510, Online, November 2020. Association for Computational Linguistics. 4

[36] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: BEiT pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 5

[37] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: An embodied multimodal language model. 2023. 4, 5

[38] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA

matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 5

[39] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 5

[40] Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8076–8084, 2019. 5

[41] Yixuan Qiao, Hao Chen, Jun Wang, Yihao Chen, Xianbin Ye, Ziliang Li, Xianbiao Qi, Peng Gao, and Guotong Xie. Winner team Mia at TextVQA challenge 2021: Vision-and-language representation learning with pre-trained sequence-to-sequence model. *arXiv preprint arXiv:2106.15332*, 2021. 5

[42] Ali Furkan Biten, Ron Litman, Yusheng Xie, Srikar Appalaraju, and R. Manmatha. LaTr: Layout-aware transformer for scene-text VQA. In *CVPR*, 2022. 5

[43] Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. Unifying vision, text, and layout for universal document processing. In *CVPR*, 2023. 5

[44] Srikar Appalaraju, Peng Tang, Qi Dong, Nishant Sankaran, Yichu Zhou, and R Manmatha. Docformerv2: Local features for document understanding. *arXiv preprint arXiv:2306.01733*, 2023. 5

[45] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016. 5

[46] Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of ACL*, 2022. 5

[47] Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. *arXiv preprint arXiv:2302.11154*, 2023. 5

[48] Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions? *arXiv preprint arXiv:2302.11713*, 2023. 5

[49] Duy-Kien Nguyen, Vedanuj Goswami, and Xinlei Chen. Revisiting modulated convolutions for visual counting and beyond. *ICLR*, 2021. 4

[50] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 6, 7

[51] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. VATEX: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, 2019. 6, 7

[52] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 6, 7

[53] Mathew Monfort, SouYoung Jin, Alexander Liu, David Harwath, Rogerio Feris, James Glass, and Aude Oliva. Spoken Moments: Learning joint audio-visual representations from video descriptions. In *CVPR*, 2021. 6, 7

[54] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *CVPR*, 2021. 6, 7, 8

[55] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *MM*, 2017. 6, 7, 8

[56] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. ActivityNet-QA: A dataset for understanding complex web videos via question answering. In *AAAI*, 2019. 6, 7

[57] Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, Guohai Xu, Ji Zhang, Songfang Huang, Fei Huang, and Jingren Zhou. mplug-2: A modularized multi-modal foundation model across text, image and video. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 38728–38748. PMLR, 2023. 7

[58] Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. End-to-end dense video captioning with parallel decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6847–6857, 2021. 7

[59] Feng Cheng, Xizi Wang, Jie Lei, David Crandall, Mohit Bansal, and Gedas Bertasius. VindLU: A recipe for effective video-and-language pretraining. In *CVPR*, 2023. 7

[60] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretrain-

ing for multimodal video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17959–17968, 2022. 7

[61] Mario Fritz Mateusz Malinowski. A multi-world approach to question answering about real-world scenes based on uncertain input. In *NeurIPS*, 2014. 6

[62] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255, 2009. 6, 7, 8, 9

[63] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with ImageNet? *arXiv preprint arXiv:2006.07159*, 2020. 6, 7, 8, 9

[64] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 6, 7, 8, 9

[65] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 6, 7, 8, 9

[66] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019. 6, 7, 8, 9

[67] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning*, pages 5389–5400, 2019. 6, 7, 8, 9

[68] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022. 7

[69] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. Regionclip: Region-based language-image pretraining. In *CVPR*, 2022. 7

[70] Ting Chen, Saurabh Saxena, Lala Li, David J. Fleet, and Geoffrey E. Hinton. Pix2Seq: A language modeling framework for object detection. In *The Tenth International Conference on Learning Representations, ICLR*, 2022. 6, 7, 9

[71] Agrim Gupta, Piotr Dollár, and Ross B. Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5356–5364. Computer Vision Foundation / IEEE, 2019. 7

[72] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, 2018. 8

[73] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 2017.

[74] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*, 2017. 8

[75] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *CVPR*, 2020. 8

[76] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAccT*, 2018. 8

[77] Jessica Deuschel, Bettina Finzel, and Ines Rieger. Uncovering the bias in facial expressions. *arXiv preprint arXiv:2011.11311*, 2020. 8

[78] Ben Hutchinson, Jason Baldridge, and Vinodkumar Prabhakaran. Underspecification in scene description-to-depiction tasks. In *AACL / IJCNLP*, 2022. 8

[79] Terrance De Vries, Ishan Misra, Changhan Wang, and Laurens Van der Maaten. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 52–59, 2019. 8

[80] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021. 8

[81] Alyssa Lees, Vinh Q Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. A new generation of perspective API: Efficient multilingual character-level transformers. In *KDD*, 2022. 8

[82] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science*, 2012. 8

[83] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2020. 1

[84] Yaru Hao, Yutao Sun, Li Dong, Zhixiong Han, Yuxian Gu, and Furu Wei. Structured prompting: Scaling in-

context learning to 1,000 examples. *arXiv preprint arXiv:2212.06713*, 2022. 4

[85] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 7

[86] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. Annotating objects and relations in user-generated videos. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 279–287, 2019. 8

[87] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Joshua Tenenbaum, and Boris Katz. ObjectNet: a large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 9453–9463, 2019. 9

[88] Priya Goyal, Adriana Romero Soriano, Caner Hazirbas, Levent Sagun, and Nicolas Usunier. Fairness indicators for systematic assessments of visual feature extractors. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 70–88, 2022. 10

[89] Candice Schumann, Susanna Ricco, Utsav Prabhu, Vittorio Ferrari, and Caroline Rebecca Pantofaru. A step toward more inclusive people annotations for fairness. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2021. 10, 11, 12

[90] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 10

[91] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021. 10

[92] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. 10

[93] Ellis Monk. Monk skin tone scale, 2019. 12

[94] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. Gender recognition or gender reductionism? the social implications of embedded gender recognition systems. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–13, 2018. 12

[95] Os Keyes. The misgendering machines: Trans/hci implications of automatic gender recognition. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–22, 2018. 12