PaLI: A Jointly-Scaled Multilingual Language-Image Model

Xi Chen[*] **Xiao Wang** **Soravit Changpinyo** **AJ Piergiovanni** **Piotr Padlewski**
Daniel Salz **Sebastian 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** **Radu Soricut**

Google Research

Abstract

Effective scaling and a flexible task interface enable large language models to excel at many tasks. **PaLI** (**Pa**thways **L**anguage and **I**mage model) extends this approach to the joint modeling of language and vision. PaLI generates text based on visual and textual inputs, and with this interface performs many vision, language, and multimodal tasks, in many languages. To train PaLI, we make use of large pretrained encoder-decoder language models and Vision Transformers (ViTs). This allows us to capitalize on their existing capabilities and leverage the substantial cost of training them. We find that joint scaling of the vision and language components is important. Since existing Transformers for language are much larger than their vision counterparts, we train the largest ViT to date (ViT-e) to quantify the benefits from even larger-capacity vision models. To train PaLI, we create a large multilingual mix of pretraining tasks, based on a new image-text training set containing 10B images and texts in over 100 languages. PaLI achieves state-of-the-art in multiple vision and language tasks (such as captioning, visual question-answering, scene-text understanding), while retaining a simple, modular, and scalable design.

1 Introduction

Increasing neural network capacity has been a successful trend in the modeling of language and vision tasks. On the language side, models such as T5 (Raffel et al., 2020), GPT-3 (Brown et al., 2020), Megatron-Turing (Shoeybi et al., 2019), GLAM (Du et al., 2022), Chinchilla (Hoffmann et al., 2022), and PaLM (Chowdhery et al., 2022) have shown significant advantages from training large Transformers on large amounts text data. On the vision side, CNNs (Mahajan et al., 2018; Huang et al., 2019; Kolesnikov et al., 2020), Vision Transformers (Dosovitskiy et al., 2021), and other models (Tolstikhin et al., 2021; Riquelme et al., 2021) have seen similar benefits from scale (Zhai et al., 2022a), albeit to a lesser extent than in language. Language-and-vision modeling has followed a similar trend, examples include SimVLM (Wang et al., 2021), Florence (Yuan et al., 2021), CoCa (Yu et al., 2022), GIT (Wang et al., 2022a), BEiT (Wang et al., 2022c), and Flamingo (Alayrac et al., 2022).

[*] Correspondence: chillxichen@google.com

	Image Captioning			Visual Question Answering			
	COCO	NoCaps*	TextCaps*	VQAv2*	TextVQA*	VizWiz-QA*	OKVQA
GIT2	145.0	124.8	145.0	81.9	67.3	70.1	-
Flamingo	138.1	-	-	82.1	54.1	65.4	57.8
BEiT-3	147.6	-	-	84.0	-	-	-
CoCa	143.6	120.6	-	82.3	-	-	-
PaLI (Ours)	**149.1**	**124.4**	**160.4**	**84.3**	**73.1**	**73.3**	**64.5**

Table 1: PaLI model results on image-language tasks. Test set results are reported where possible. COCO result is on the Karpathy-test (Karpathy and Fei-Fei, 2015) Benchmarks labeled with "*" are evaluated on public server. VQA tasks are evaluated in the open-vocabulary generation setting, which is more challenging than the closed-vocabulary classification setting (numbers shown in gray). See Section 4 for all results. CIDEr scores (Vedantam et al., 2015) are reported for the image captioning tasks and VQA accuracy (Antol et al., 2015) for the VQA tasks.

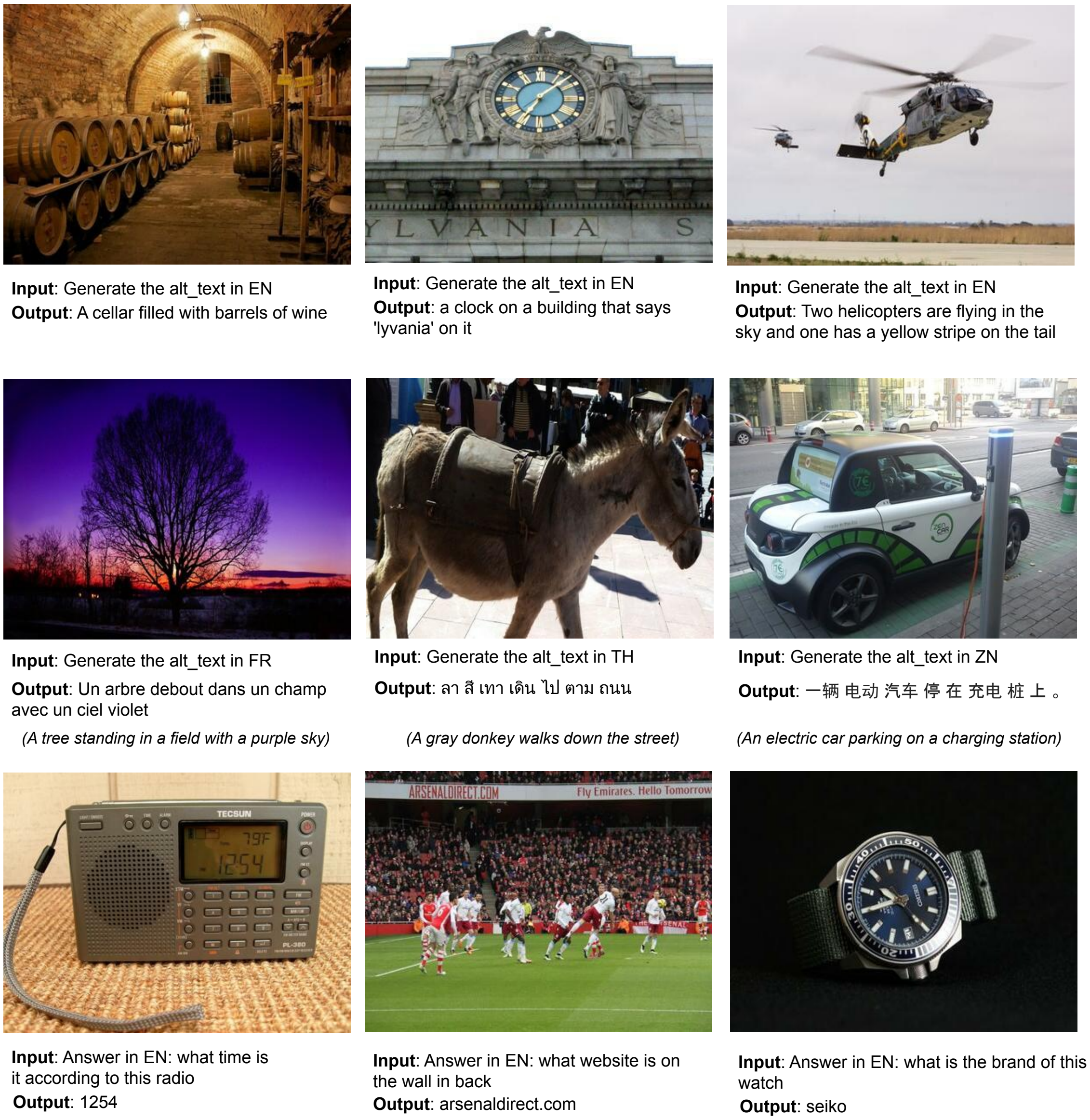
We continue this line of work with **PaLI** (**Pa**thways **L**anguage and **I**mage). PaLI performs many image-only, language-only, and image+language tasks, across many languages, using a single "image-and-text to text" interface. A key ingredient to PaLI is the reuse of large unimodal backbones for language and vision modeling, in order to transfer existing capabilities and reduce training cost. On the language side, we reuse the 13B parameter mT5-XXL (Xue et al., 2021). mT5-XXL already packages language understanding and generation capabilities. We show that these capabilities can be maintained and extended into a multimodal setting. On the vision side, in addition to reusing the 2B-parameter ViT-G model (Zhai et al., 2022a), we train a 4B-parameter model, which we call ViT-e ("enormous"). ViT-e achieves good performance on image-only tasks, such as 90.9% ImageNet finetuning, and 84.9% on ObjectNet (Barbu et al., 2019).

We find benefits from jointly scaling both the vision and the language components, with vision providing a better return on investment (accuracy improvement per parameter/FLOP). As a result, the capacity of our largest PaLI model, PaLI-17B, is distributed relatively equitably between the two modalities, with the ViT-e component accounting for about 25% of the total parameter count. This is not always the case for prior work in large-capacity vision and language modeling (Wang et al., 2022a; Alayrac et al., 2022), due to the prior scale mismatch between vision and language backbones.

We enable knowledge-sharing between multiple image and/or language tasks by casting them into a generalized VQA-like task. We frame all tasks using an "image+query to answer" modeling interface, in which both the query and answer are expressed as text tokens. This allows PaLI to capitalize on transfer learning across tasks, and enhance language-and-image understanding capabilities in a wide range of vision and language problems: image captioning, visual question-answering, scene-text understanding, and others (Figure 1).

To train PaLI-17B, we build a new high-volume image-and-language dataset, WebLI, which consists of 10B pairs of image-text examples. Importantly, the WebLI dataset contains text in over 100 languages. By training the model to perform multimodal tasks in many languages, we greatly increase the task diversity, and test the model's ability to effectively scale both across tasks and across languages. As a reference for future usage, we provide a data card to report information about the WebLI and its construction.

PaLI-17B achieves SOTA results on multiple benchmarks, outperforming some strong models (Table 1). Specifically, PaLI outperforms recent and concurrent models on the long-standing COCO Captioning benchmark (Chen et al., 2015), with **149.1** CIDEr score on the Karpathy split (Karpathy and Fei-Fei, 2015). PaLI also achieves a new SOTA of **84.3%** on VQAv2 (Goyal et al., 2017) while using an open-vocabulary text generative setting that is similar to Flamingo (Alayrac et al., 2022). This result outperforms even models evaluated in a fixed-vocabulary classification setting, e.g. CoCa (Yu et al., 2022), SimVLM (Wang et al., 2021), BEiT-3 (Wang et al., 2022c). Last but not least, our work provides a scaling roadmap for future multimodal models. Model scaling is particularly important for language-image understanding in a multilingual setting. Our results support



Input: Generate the alt_text in EN
Output: A cellar filled with barrels of wine

Input: Generate the alt_text in EN
Output: a clock on a building that says 'lyvania' on it

Input: Generate the alt_text in EN
Output: Two helicopters are flying in the sky and one has a yellow stripe on the tail

Input: Generate the alt_text in FR
Output: Un arbre debout dans un champ avec un ciel violet
(A tree standing in a field with a purple sky)

Input: Generate the alt_text in TH
Output: ลา สี เทา เดิน ไป ตาม ถนน
(A gray donkey walks down the street)

Input: Generate the alt_text in ZN
Output: 一辆 电动 汽车 停 在 充电 桩 上 。
(An electric car parking on a charging station)

Input: Answer in EN: what time is it according to this radio
Output: 1254

Input: Answer in EN: what website is on the wall in back
Output: arsenaldirect.com

Input: Answer in EN: what is the brand of this watch
Output: seiko

Figure 1: PaLI addresses a variety of vision and language tasks across many languages, for example, image captioning, visual question answering, scene-text understanding, etc. Images from the publicly-available TextVQA (Singh et al., 2019) and TextCaps (Sidorov et al., 2020) datasets are shown, together with PaLI inputs and outputs.

the conclusion that scaling the components of each modality yields better performance compared to more skewed alternatives.

2 Related Work

Pretrained models have proven effective in both vision (Dosovitskiy et al., 2021; Zhai et al., 2022a) and language (Raffel et al., 2020; Brown et al., 2020) tasks. Image-text pretraining has also become the default approach to tackle V&L tasks (Tan and Bansal, 2019; Chen et al., 2020; Zhang et al., 2021; Cho et al., 2021; Hu et al., 2022). While benefiting from the text representation and generation capabilities of the Transformer architecture, some of these vision-language models rely on external systems (such as Fast(er) R-CNN (Ren et al., 2015)) to provide detected object names and the related precomputed dense features. Such reliance limited the capability to scale up the model and performance. With the introduction of Vision Transformers (Dosovitskiy et al., 2021), vision and language modalities can be jointly modeled by transformers in a more scalable fashion (Yuan et al., 2021; Yu et al., 2022; Wang et al., 2022a; Alayrac et al., 2022).

Contrastive learning techniques are recently used in image-text pretraining (Radford et al., 2021; Jia et al., 2021), where the model is trained on image-text pairs dataset collected from public web. The high-level idea is to learn a shared embedding space for both image and text, such that paired image and text stays close to each other, while unpaired image and text are distant from each other. The follow up work (Pham et al., 2021; Zhai et al., 2022b) studies the impact of the training data and batch size in contrastive learning. They observed that additional high quality data (Pham et al., 2021) or a pretrained vision model (Zhai et al., 2022b) can lead to better vision-language models, and a large batch size is generally beneficial to contrastive learning. Furthermore, Zhai et al. (2022b) show that with a pretrained and locked vision model, one needs to train only a paired text encoder model to get good language embeddings. Yuan et al. (2021) extend contrastively pretrained models to more downstream tasks, including object detection and video recognition tasks with task-specific adaptations.

Another approach is to train vision-language models to generate text autoregressively, which has found success in image captioning problems (Donahue et al., 2015; Vinyals et al., 2015). This approach has the advantage of a unified formulation of vision-language tasks as a text generation problem (Cho et al., 2021; Wang et al., 2022b; Piergiovanni et al., 2022b). In (Cho et al., 2021), the vision-language model is trained to recover masked text. SimVLM (Wang et al., 2021) proposed an image-language preraining approach leveraging a prefix language modeling objective. The unified framework in (Wang et al., 2022b) extends the generation capability to include text to image generation.

Recently, there has been several other works exploring along these directions of joint vision and language modeling, while increasing the model capacity. CoCa (Yu et al., 2022) pretrained a 2.1B image-text encoder-decoder model jointly with contrastive loss and generative loss. GIT (Wang et al., 2022a) proposed a model consisting of a single image encoder and a text decoder. They adapted a Swin-like structure pretrained by contrastive learning as the image encoder, while training their GIT model with a captioning (generative) loss. In their latest version, GIT2 (Wang et al., 2022a), the model size is scaled up to 5.1B, with the majority of parameters on the vision side (4.8B). BEiT-3 (Wang et al., 2022c) designed an architecture with vision, language, and vision-language experts, operating with a shared multi-head self-attention followed by a switch for "expert" modules, resulting in a 1.9B model trained from scratch on a variety of public image, text and image-text datasets. Flamingo (Alayrac et al., 2022) is built upon a 70B language model (Hoffmann et al., 2022) as a decoder-only model whose majority of parameters are frozen in order to preserve language-generation capabilities.

Besides model capacity, scaling up the dataset for vision-language pretraining has been demonstrated to be beneficial as well. LXMERT (Tan and Bansal, 2019) performed their cross-modality LM based on human annotated datasets including COCO (Chen et al., 2015) and Visual Genome (Krishna et al., 2017). Vision-language pretraining can also benefit from automatically mined and curated larger datasets such as Conceptual Captions (CC3M) and CC12M (Sharma et al., 2018; Changpinyo et al., 2021), with 3 million and 12 million examples, respectively. LEMON (Hu et al., 2022) further pushes the dataset size created in a similar way to 200M examples. For better scaling the model, larger, noisier datasets such as the ALIGN dataset (1.8B) (Jia et al., 2021) have been constructed, and their benefit has been observed in works like SimVLM (Wang et al., 2021) and CoCa (Yu et al., 2022).

3 The PaLI Model

In this section we detail the model architecture, training data, and protocol used to train PaLI.

3.1 Architecture

With PaLI, we aim to perform both unimodal (language, vision) and multimodal (language and vision) tasks. Typically, many of these tasks are best handled by different models. For instance, image classification, and many formulations of VQA, require predicting elements from a fixed set, while

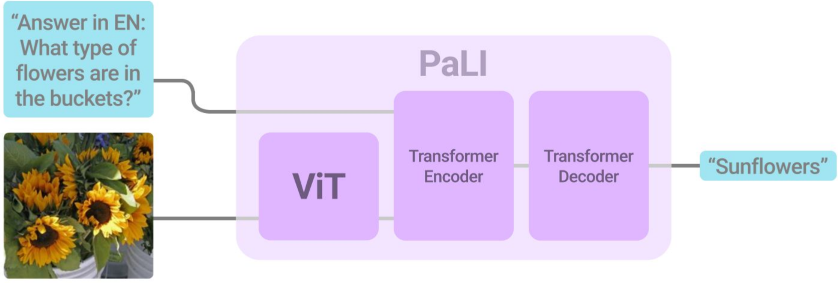


Figure 2: The PaLI main architecture is simple and scalable. It uses an encoder-decoder Transformer model, with a large-capacity ViT component for image processing.

language-only tasks and image captioning require open-vocabulary text generation. We resolve this by using the most sufficiently general interface needed for all tasks considered: the model accepts as input an image and text string, and generates text as output. The same interface is used both during pretraining and fine-tuning. Since all tasks are performed with the same model, i.e. we have no tasks-specific parameters or "heads", we use text-based prompts to indicate to the model which task to perform.

Figure 2 shows a high-level schematic of the model architecture. At its core, we have a text encoder-decoder Transformer (Vaswani et al., 2017). To include vision as input, the text encoder is fed with a sequence of visual "tokens": output features of a Vision Transformer which takes as input an image. No pooling is applied to the output of the Vision Transformer before passing the visual tokens to the encoder-decoder model via cross-attention.

We reuse previously trained unimodal checkpoints. For the text encoder-decoder, we reuse pre-trained mT5 (Xue et al., 2021) models, while for the image encoder, we reuse large vanilla ViT models (Dosovitskiy et al., 2021; Zhai et al., 2022a).

The visual component We introduce and train the largest vanilla ViT architecture to date, named **ViT-e**. ViT-e has the same architecture and uses the same training recipe as the 1.8B parameter ViT-G model (Zhai et al., 2022a), and it is scaled to 4B parameters. The only other difference is that we apply learning rate cool-down twice, once with and once without inception crop augmentation, and average ("soup") the weights of the two models as in Wortsman et al. (2022). While the scaling laws have been studied in both the vision domain and the language domain, scaling behaviour is less explored in combined vision and language models. Scaling up vision backbones leads to saturating gains on classification tasks such as ImageNet (Zhai et al., 2022a). We further confirm this, observing that ViT-e is only marginally better than ViT-G on ImageNet (Table 10). However, we observe substantial performance improvements from ViT-e on vision-language tasks in PaLI, see Section 4. For example, ViT-e yields almost three additional CIDEr points over ViT-G on the COCO captioning task. This hints towards future headroom for vision-language tasks with even larger ViT backbones.

The language component We adopt the mT5 (Xue et al., 2021) backbone as our language modeling component. We experiment using the pretrained mT5-Large (1B parameters) and the mT5-XXL (13B parameters), from which we initialize the language encoder-decoder of PaLI. We train on a mix of many tasks, including pure language understanding tasks (see Section 3.3). This helps avoid catastrophic forgetting of the mT5's language understanding and generation abilities. As a result, PaLI-17B continues to achieve similar levels of language-understanding accuracy on both the English benchmarks (Wang et al., 2019a) and across the languages measured by the XTREME (Hu et al., 2020) benchmark (Section 4).

The overall model Three model sizes are considered (Table 2): 1) A version with 3B parameters, PaLI-3B, where the language component is initialized from mT5-Large (Xue et al., 2021) (1B parameters), and the vision component is ViT-G (Zhai et al., 2022a) (1.8B parameters). 2) A version with 15B parameters, PaLI-15B, where the language component is initialized from mT5-XXL (Xue et al., 2021) (13B parameters), and the vision component is ViT-G (1.8B parameters). 3) The main

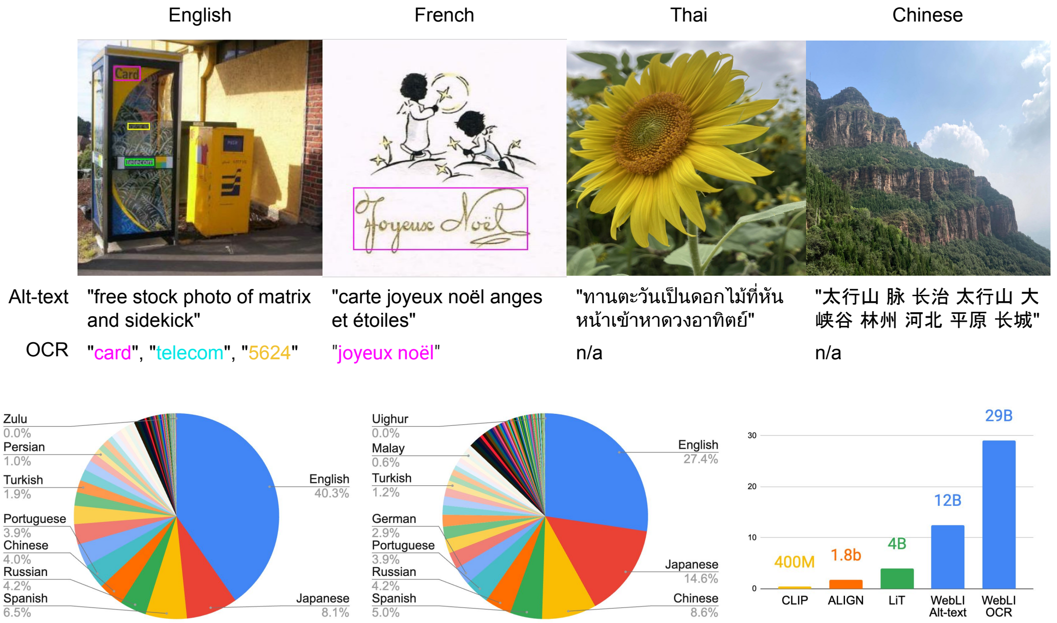


Figure 3: The WebLI dataset. Top: Sampled images[4] associated with multilingual alt-text (available) and OCR (computed using GCP Vision API[5]). Bottom left/middle: Statistics of recognized languages from alt-text/OCR. Bottom right: Image-text pair counts, compared against other large-scale vision-language datasets.

version with 17B parameters, PaLI-17B, where the language model is initialized from mT5-XXL, and the vision component is the newly-trained ViT-e model (4B parameters).

Model	Components	Image Encoder	Multimodal Encoder-Decoder	Total
PaLI-3B	ViT-G, mT5-L	1.8B	1.2B	3.0B
PaLI-15B	ViT-G, mT5-XXL	1.8B	13B	14.8B
PaLI-17B	ViT-e, mT5-XXL	3.9B	13B	16.9B

Table 2: The size in terms of number of parameters for the trained PaLI model versions.

3.2 Data

Scaling studies for deep learning show that larger models require larger datasets to train effectively (Hoffmann et al., 2022; Kaplan et al., 2020; Zhai et al., 2022a). To unlock the potential of image-language pretraining, we introduce WebLI, a multilingual image-language dataset built from images and texts available on the public web. Examples and statistics for the WebLI corpus are shown in Figure 3, and a complete Data Card (Pushkarna et al., 2022) is given in the Appendix.

WebLI scales up the image language data collection from English-only datasets to 109 languages, which enables us to pretrain PaLI multilingually, and perform downstream tasks across many languages. The data collection process is similar to those reported in (Jia et al., 2021; Zhai et al., 2022b). Due to the abundance of multilingual content on the internet, the collection process for the WebLI dataset can be scaled to cover 10 billion images and 12 billion alt-texts. In addition to annotation with web text, we apply the GCP Vision API to extract OCR annotations on all images, resulting in 29 billion image-OCR pairs.

[5] https://cloud.google.com/vision

[5] The second image is by jopradier (original), used under the CC BY-NC-SA 2.0 license. Remaining images are also used with permissions.

Due to the scale of WebLI, to mitigate train-to-test leakage, we perform near de-duplication of the images against the train, validation, and test splits of 68 common vision/vision-language datasets. Eliminating these images from the WebLI dataset does not result in any significant shrinkage (0.36%), and avoids any potential "leakeage" of examples from the pretraining setup to the downstream evaluation tasks.

To improve the data quality in terms of image-text alignment, we score image and alt-text pairs based on their cross-modal similarity. This score is measured with cosine similarity between embedding representations from each modality, computed as follows. The image embeddings are trained with a graph-based, semi-supervised representation learning approach, as described in Juan et al. (2019). Then, the text embeddings are learned using the frozen image embeddings, based on a contrastive approach using a Transformer encoder for the text, which forces both modality representations to the same embedding space. To balance quality and retain scale, we tune a threshold on the image and alt-text pairs' score, and end up retaining only the top 10% scoring of the original WebLI image-text pairs (about 1B examples), which we use to train PaLI.

3.3 The Pretraining Task Mixture

To accommodate diverse tasks in the image-language space, we train PaLI using a mixture of pretraining tasks. This mixture is designed to span a range of general capabilities useful for downstream tasks. Following the task interface described in Section 3, we specify each task using a training data source and a template-based prompt, described below. For each task, the model is trained using a language-model–style teacher forcing (Goodfellow et al., 2016) with a standard softmax cross-entropy loss.

- **Span corruption on text-only data** uses the same technique described by Xue et al. (2021), corrupting 15% of the tokens from a given text-only example and using "sentinels" of the form $\langle \text{extra\_id\_}k \rangle$ for each corrupted span; the text-only examples are using a sample of 100M of the text-only examples used to train GLaM and PaLM (Du et al., 2022; Chowdhery et al., 2022).
- **Split-captioning (SplitCap) on WebLI alt-text data** is inspired by the pretraining objective of Wang et al. (2021), and works by splitting each alt-text string randomly into two parts, $\langle \text{cap}_1 \rangle$ and $\langle \text{cap}_2 \rangle$. It uses the prompt "*Generate the alt_text in* $\langle \text{lang} \rangle$ *at* $\langle \text{pos} \rangle$*:* $\langle \text{cap}_1 \rangle$ $\langle \text{extra\_id\_0} \rangle$" (where $\langle \text{lang} \rangle$ is the language code of the alt-text string, and $\langle \text{pos} \rangle$ is the number of words in $\langle \text{cap}_1 \rangle$), with $\langle \text{cap}_2 \rangle$ as the target.
- **Captioning (Cap) on CC3M-35L on native and translated alt-text data** using the prompt "*Generate the alt_text in* $\langle \text{lang} \rangle$ *at 0:* $\langle \text{extra\_id\_0} \rangle$", with the alt-text string in language $\langle \text{lang} \rangle$ as the target. CC3M-35L is Conceptual Captions (Sharma et al., 2018) training data, translated into an additional 34 languages (the same as the non-English ones covered by Crossmodal-3600 (Thapliyal et al., 2022), except for Cusco-Quechua), for a total of 100M examples.
- **OCR on WebLI OCR-text data** using the prompt "*Generate the ocr_text in* $\langle \text{lang} \rangle$*:* $\langle \text{extra\_id\_0} \rangle$", with $\langle \text{OCR\_text} \rangle$ as the target, where $\langle \text{OCR\_text} \rangle$ is the concatenation of the annotated OCR texts in language $\langle \text{lang} \rangle$ (Kil et al., 2022) produced by the GCP Vision API for the input image.
- **English and Cross-Lingual VQA on native and translated** VQ^2A**-CC3M-35L-100M VQA triplets** using, for a given $\langle image, [question], [answer] \rangle$ VQA triple, the prompt: "*Answer in EN: [question]* $\langle \text{extra\_id\_0} \rangle$", with $[answer]$ for the target. VQ^2A-CC3M-35L-100M is a 100M random subset of VQ^2A-CC3M (Changpinyo et al., 2022a), translated into the same additional 34 languages as mentioned above. Note that we use English answers in all instances here, as the English-native answers for VQA are often short and too prone to errors to perform out-of-context automatic translation.
- **English and Cross-Lingual visual question generation (VQG) on native and translated** VQ^2A**-CC3M-35L-100M VQA triplets** using, for a given $\langle image, [question], [answer] \rangle$ VQA triple, the prompt: "*Generate a question in* $\langle \text{lang} \rangle$ *for [answer]:* $\langle \text{extra\_id\_0} \rangle$", with $[question]$ in language $\langle \text{lang} \rangle$ as the target. Similarly, we use only English answers here.
- **English-only Object-Aware (OA) VQA** is based on VQA triplets derived from automatically-produced, non-exhaustive object labels, inspired by Piergiovanni et al. (2022a).

We automatically generate 4 different prompt types, based on the available object labels, as follows. (1) Prompt: "*Answer in EN: List the objects present:* $\langle \text{extra\_id\_0} \rangle$", with the target: $\langle \text{object}_1 \rangle, \ldots, \langle \text{object}_\text{N} \rangle$. (2) Prompt: "*Answer in EN: Is* $\langle \text{object}_\text{k} \rangle$ *in the image?* $\langle \text{extra\_id\_0} \rangle$", with the target "Yes" or "No". (3) Prompt: "*Answer in EN: Is* $\langle \text{object}_1 \rangle$*,* $\ldots$*,* $\langle \text{object}_\text{N} \rangle$ *in the image?* $\langle \text{extra\_id\_0} \rangle$", with the target "Yes" or "No". (4) Prompt: "*Answer in EN: Which of* $\langle \text{object}_1 \rangle$*,* $\ldots$*,* $\langle \text{object}_\text{N} \rangle$ *are in the image?* $\langle \text{extra\_id\_0} \rangle$", with the target made of the list of object labels present. To create these examples, we require object-level annotations, for which we use Open Images (Kuznetsova et al., 2020), from which we create 50M examples.

- **Object detection** is a generative object-detection task inspired by Chen et al. (2021). The target sequence describes bounding-box coordinates and object labels, e.g. "*10 20 90 100 cat 20 30 100 100 dog*". The coordinates are in the y_{min} x_{min} y_{max} x_{max} order, and range between 0 and 999. Unlike Chen et al. (2021), the prompt used contains a set of positive and negative class labels, i.e. object classes that are present and not present in the image (e.g. "*detect cat and dog and leopard*"). The prompt is prefixed with the word "*detect*". For the datasets that do not have negative class labels explicitly defined, we randomly sample non-positive class labels. Since WebLI does not contain bounding box annotations, we train on a mixture of public datasets, totalling 16M images: Open Images (Kuznetsova et al., 2020), Visual Genome (Krishna et al., 2017), and Object365 (Shao et al., 2019). The datasets are de-duplicated against evaluation tasks. These examples are included to increase object awareness capabilities of the model.

The overall size of the data we use for pretraining is 1.6B examples. This dataset is comparable, but slightly smaller and designed to be cleaner than the datasets used in SimVLM (1.8B), CoCa (1.8B), and Flamingo (2.3B). However, unlike for the aforementioned datasets, WebLI is multilingual, so the 1.6B examples follow a long-tailed distribution over the 100+ languages covered. The coefficients for the training mixture are empirically determined, see the Appendix for additional information.

3.4 Training Details

ViT-e We show ViT-e's configuration in Table 3 alongside ViT-g and ViT-G for reference. Width, depth and MLP dimensions are all further scaled up in ViT-e, resulting in a model with 4B parameters. The model training setup is copied from the ViT-G model (Zhai et al., 2022a), on the JFT-3B dataset (Zhai et al., 2022a), with $16,384$ batch size, 224×224 resolution. We train the model for 1M steps using 0.0008 initial learning rate, with an inverse square-root learning rate decay, and a linear cool-down to zero for the final 100k steps. The only additional technique added is model souping (Wortsman et al., 2022): we run the 900K to 1M cool-down twice, once with inception cropping and once with resizing only. Thus, the final ViT-e model consists of the average weights of these two cool-downs. ViT-e is pretrained using the `big_vision` codebase (Beyer et al., 2022).

Name	**Width**	**Depth**	**MLP**	**Heads**	**Params (M)**	**GFLOPs**	
						224^2	384^2
g/14	1408	40	6144	16	1011	533.1	1596.4
G/14	1664	48	8192	16	1843	965.3	2859.9
e/14	1792	56	15360	16	3926	1980	5777

Table 3: ViT-e architecture details.

The overall model The overall PaLI models are implemented in `JAX/Flax` (Bradbury et al., 2018) using the open-source `T5X` (Roberts et al., 2022) and `Flaxformer` (Heek et al., 2020) frameworks. For the learning rate, we use a 1k-step linear warmup, followed by inverse square-root decay. For PaLI-3B, we use a peak learning rate of 1e-2. For larger models, PaLI-15B and PaLI-17B, we use a peak learning rate of 5e-3. We use the Adafactor (Shazeer and Stern, 2018) optimizer with $\beta_1 = 0$ and second-moment exponential decay set to 0.8.

The largest model, PaLI-17B, is pretrained using 1,024 GCP-TPUv4 chips for 7 days. It uses a four-way model partitioning (Roberts et al., 2022) and a batch size of 4,096. This is slightly less TPU resources than used to train other large vision and language models on TPUs. SimVLM used 2,048 GCP-TPUv3 for 5 days (Wang et al., 2021), while CoCa used 2,048 GCP-TPUv4 chips for 5 days (Yu et al., 2022). Flamingo used 1,536 GCP-TPUv4 chips for 15 days (Alayrac et al., 2022).

During training, the model passes over 1.6B images, one epoch over the entire pretraining dataset. The image resolution for this pass is 224×224. During training, only the parameters of the language component are updated and the vision component is frozen, which provides a boost in performance (Sec. 4.7).

For the largest model, PaLI-17B, we perform a further high-resolution (588×588) pre-finetuning for the multilingual tasks, similar to previous works (Radford et al., 2021; Jia et al., 2021; Yuan et al., 2021; Yu et al., 2022). This second stage of training is only for 10k steps at batch size 1024 (10M examples in total) and is performed on a subset of the full training mix. Appendix C contains details. In this high-resolution finetuning phase, all of the parameters of PaLI are updated.

4 Experiments

We evaluate on multiple downstream tasks that include a number of vision and language benchmarks, and additionally language-only and vision-only benchmarks. Out of the seven English-only V&L benchmarks we consider, PaLI-17B establishes new SOTA numbers for five of them (including the well established COCO captioning and VQAv2 benchmarks), while treating them as open-vocabulary tasks with a 100 language vocabulary containing 250k tokens. We also establish a new SOTA results on multilingual image captioning and multilingual VQA tasks.

Unless specified otherwise, for the multimodal tasks, we take the model checkpoint with additional high-resolution pre-finetuning (Section 3), and finetune at the same resolution on the downstream task.

4.1 Image Captioning

We finetune and evaluate the PaLI model variants on three English-only image captioning benchmarks (Table 4): COCO Captions (Chen et al., 2015), NoCaps (Agrawal et al., 2019), and TextCaps (Sidorov et al., 2020). We then quantify the multilingual captioning capability of PaLI on the 35-language benchmark Crossmodal-3600 (Thapliyal et al., 2022). For all the benchmarks, cross-entropy loss is used for finetuning.

COCO Captions We finetune on COCO Captions (Chen et al., 2015) based on the widely adopted Karpathy split (Karpathy and Fei-Fei, 2015). PaLI outperforms the latest SOTA trained with cross-entropy loss on COCO (Wang et al., 2022c), and establishes a new high at 149.1 CIDEr (Vedantam et al., 2015) points for models trained without CIDEr-optimization training.

NoCaps This dataset (Agrawal et al., 2019) is an evaluation benchmark for image captioning that has similar style to COCO, but targets many more visual concepts than those included in the COCO. We follow previous works by evaluating NoCaps using a model finetuned on COCO. PaLI-17B achieves a 124.4 CIDEr score on test, comparable to the recent result of 124.8 from GIT2 (Wang et al., 2022a). GIT2 achieves 124.2, 125.5, 123.3 on in-domain, near-domain, and out-of-domain splits of the NoCaps test set, respectively. PaLI-17B achieves 121.1, 124.4 and 126.7, respectively. Therefore, it seems that PaLI is more performant out-of-domain, while GIT2 performs better on in-domain examples. This observation suggests that for PaLI-17B, the domain transfer from COCO to NoCaps is slightly suboptimal, compared with models pretrained with English only. Nevertheless, PaLI-17B outperforms all prior models on recognizing and describing long-tail objects outside of COCO's domain.

TextCaps This dataset (Sidorov et al., 2020) focuses on caption generation for images containing text. We finetune on TextCaps using OCR strings generated by the GCP Vision API, similar to the protocol used in (Yang et al., 2021). Following Kil et al. (2022), we order the OCR items based on their locations in the image, from top left to bottom right. We only include the OCR strings themselves, without the OCR-item locations provided by the API. GIT2 (Wang et al., 2022a) has demonstrated strong performance without the OCR input, while PaLI-17B shows the superiority of levaraging a specialized OCR system for a better recipe to solve these tasks. Results on evaluating PaLI-17B on TextCaps without OCR as input is provided in Appendix D.

	COCO	NoCaps		TextCaps	
Model	Karpathy-test	val	test	val	test
LEMON	139.1	117.3	114.3	-	-
SimVLM	143.3	112.2	110.3	-	-
CoCa	143.6	122.4	120.6	-	-
GIT	144.8	125.5	123.4	143.7	138.2
GIT2	145.0	**126.9**	**124.8**	148.6	145.0
OFA	145.3	-	-	-	-
Flamingo	138.1	-	-	-	-
BEiT-3	147.6	-	-	-	-
PaLI-3B	145.4	121.1	-	143.6	-
PaLI-15B	146.2	121.2	-	150.1	-
PaLI-17B	**149.1**	**127.0**	**124.4**	**160.0**	**160.4**

Table 4: CIDEr results for image captioning over the English-only benchmarks COCO Captions (Karpathy split), NoCaps, and TextCaps.

Multilingual captioning on XM-3600 Following Thapliyal et al. (2022), we normalize the unicode, tokenize, and remove all punctuation before calculating CIDEr scores. For languages without word boundaries such as Chinese, Japanese, Korean and Thai, a neural model is used for segmenting the text. The results on both PaLI-3B and PaLI-17B are based on models pretrained at 224×224 resolution and fine-tuned on COCO-35L at the same resolution. Table 5 contains the results. To illustrate the range of improvements over a variety of language families with different scripts and different resources, we use seven languages to show their exact CIDEr scores, in addition to the 35-language average score.

Model	en	fr	hi	iw	ro	th	zh	35-lang avg.
Thapliyal et al. (2022)	57.6	40.9	20.6	16.1	13.9	35.5	19.8	28.9
PaLI-3B	92.8	68.6	30.3	39.2	30.3	65.9	32.2	47.0
PaLI-17B	**98.1**	**75.5**	**31.3**	**46.8**	**35.8**	**72.1**	**36.5**	**53.4**

Table 5: CIDEr scores on image captioning for the Crossmodal-3600 benchmark, covering seven diverse languages (English, French, Hindi, Hebrew, Romanian, Thai, and Chinese), as well as the average of the 35 languages covered by the benchmark.

4.2 Visual Question Answering

We finetune and evaluate on four English-only visual question-answering (VQA) benchmarks (Table 6): VQAv2 (Goyal et al., 2017), TextVQA (Singh et al., 2019), VizWiz-QA (Gurari et al., 2018), and OKVQA (Marino et al., 2019).

Note that all of the VQA results reported in this paper are performed in the open-vocabulary setting using the 250k mT5 (Xue et al., 2021) vocabulary. Most prior works, e.g. SimVLM (Wang et al., 2021), CoCa (Yu et al., 2022), BEiT-3 (Wang et al., 2022c), use the VQA-as-classification setting, where a best answer among a predefined set (usually of size 3k) needs to be selected. Note that the VQA-as-open-generation setting is challenging because: (1) The generated text is directly compared to the desired answer and only an exact match is counted as accurate. (2) The PaLI vocabulary covers 100+ languages and is significantly larger than both those used in the classification setting, and those used by previous single-language open-generation models (Alayrac et al., 2022; Wang et al., 2022a).

VQAv2 PaLI-17B achieves 84.3 accuracy on this benchmark, and outperforms previous SOTA as follows: (1) By +2.2 accuracy points on the open-vocabulary generation setting, compared to Flamingo (Alayrac et al., 2022) at 82.1 accuracy. (2) By +0.3 accuracy points when compared against the best result on the closed-vocabulary classification setting, BEiT-3 (Wang et al., 2022c), at 84.0 accuracy.

	VQAv2		OKVQA	TextVQA		VizWiz-QA	
Method	test-dev	test-std	val	val	test	test-dev	test
SimVLM	80.03	80.34	-	-	-	-	-
CoCa	82.3	82.3	-	-	-	-	-
GIT	78.56	78.81	-	59.93	59.75	68.0	67.5
GIT2	81.74	81.92	-	68.38	67.27	70.97	70.1
OFA	82.0	82.0	-	-	-	-	-
Flamingo	82.0	82.1	57.8*	57.1	54.1	65.7	65.4
BEiT-3	84.2	84.0	-	-	-	-	-
KAT	-	-	54.4	-	-	-	-
Mia	-	-	-	-	**73.67**†	-	-
PaLI-3B	79.3	-	52.4	58.75	-	66.4	-
PaLI-15B	80.8	-	56.5	64.12	-	70.0	-
PaLI-17B	**84.3**	**84.3**	**64.5**	**70.45**	**73.06**	**74.4**	**73.3**

Table 6: VQA Accuracy results on VQAv2, OKVQA, TextVQA, and VizWiz-QA. PaLI models are evaluated in the open-vocabulary generation setting, and still outperform previous models that use closed-vocabulary classification evaluations (SimVLM, CoCa, BEiT3, OFA (Wang et al., 2022b)). The result on OKVQA by Flamingo (with "*") is obtained in a 32-shot learning setup. Mia (Qiao et al., 2021) (with "†") is the winning model of TextVQA Challenge 2021, based on fine-tuning T5-XL (Raffel et al., 2020). Numbers shown in gray are from models using closed-vocabulary classification.

OKVQA This vision and language benchmark requires external knowledge to answer its questions, that is, knowledge that is not directly present in the image input, and instead needs to be indirectly inferred by the model. PaLI-17B achieves 64.5 accuracy, pushing SOTA for the pretrain-finetune setup higher by 10.1 accuracy points, compared to KAT (Gui et al., 2021) at 54.4 accuracy. The best result for the 32-shot learning setup is from Flamingo (Alayrac et al., 2022), at 57.8 accuracy. The results from Flamingo and PaLI-17B suggest that leveraging external knowledge does not necessarily require specific training, and instead can be achieved with generic large-capacity models trained on large amounts of data.

TextVQA & VizWiz-QA Both TextVQA (Singh et al., 2019) and VizWiz-QA (Gurari et al., 2018) require the ability to perform question answering in the presence of text in the input image. TextVQA shares the same set of images with TextCaps, and are both based on images from Open Images (Kuznetsova et al., 2020). We finetune using OCR strings generated by the GCP Vision API, similar to the protocol in TAP (Yang et al., 2021) and Mia (Qiao et al., 2021). Evaluating PaLI-17B on TextVQA and VizWiz-QA without OCR as input is provided in Appendix D.

Cross-lingual and Multilingual VQA on xGQA and MaXM Both xGQA (Pfeiffer et al., 2022) and MaXM (Changpinyo et al., 2022b) are test-only VQA benchmarks that require multilingual understanding of visual questions. The setting in xGQA is cross-lingual (English-answers only), whereas the one in MaXM is multilingual (answer in the same language as the question). We evaluate the PaLI-17B model pretrained at 224×224 image resolution and fine-tuned on the native+translated VQAv2 (Goyal et al., 2017) (the Karpathy train split) in the 13 languages covered by xGQA and MaXM (VQAv2-13L) at 378×378 resolution. Table 7 shows significant gains on both benchmarks across all languages.

4.3 Language-understanding Capabilities

Since PaLI is pretrained with a diverse mixture of multimodal tasks with image and text data, it raises the question on whether it would "forget" its language modeling capability, and therefore exhibit inferior performance on language-understanding tasks compared to its unimodal starting checkpoint (mT5-XXL in the case of PaLI-17B).

Therefore, we compare mT5-XXL and PaLI-17B on a range of language understanding benchmarks, including the English-only SuperGLUE benchmark (Wang et al., 2019a), as well as three multilingual

Model	xGQA							
	en	bn	de	id	ko	pt	ru	zh
MPT-2en (Changpinyo et al., 2022b)	41.5	38.6	40.5	39.5	38.7	39.8	39.5	39.5
PaLI-17B	**54.2**	**50.0**	**52.2**	**50.6**	**50.4**	**51.3**	**50.3**	**50.6**

Model	MaXM						
	en	fr	hi	iw	ro	th	zh
MPT-2⟨lang⟩ (Changpinyo et al., 2022b)	22.7	25.3	31.3	24.0	26.3	32.3	22.0
PaLI-17B	**33.0**	**34.0**	**37.3**	**32.3**	**32.7**	**40.3**	**31.0**

Table 7: Cross-lingual VQA results on xGQA (Pfeiffer et al., 2022) (top) and multilingual VQA results on MaXM (Changpinyo et al., 2022b) (bottom). All models are fine-tuned on translated VQAv2 in 13 languages. Exact-match accuracy is reported.

benchmarks from the XTREME (Hu et al., 2020): XNLI (Conneau et al., 2018), which is a textual entailment task covering 14 languages, XQuAD (Artetxe et al., 2020) and TyDiQA-GoldP (Clark et al., 2020), which are both question-answering tasks covering 10 and 11 languages, respectively. For the three XTREME benchmarks, we evaluate in the zero-shot (ZS) transfer setting, whereas for SuperGLUE the models are finetuned (FT).

Table 8 summarizes the results. Despite the pretraining mixture heavily favoring the V&L tasks, PaLI-17B is able to maintain a high-level of language-understanding capabilities for English, and it is on-par with the state-of-the-art mT5-XXL checkpoint on the XTREME benchmarks.

Model Method	SuperGLUE FT	XNLI ZS	XQuAD ZS	TyDiQA-GoldP ZS
Metric	Avg. Score	Accuracy	F1/EM	F1/EM
mT5-XXL (Xue et al., 2021)	89.2	85.0	82.5 / 66.8	80.8 / 65.9
mT5-XXL (our setting)	89.3	84.5	82.6 / 66.6	81.6 / 66.3
PaLI-17B	88.2	84.9	81.8 / 66.0	81.2 / 66.5

Table 8: Results on SuperGLUE and three XTREME tasks. The first row is the result reported by mT5 (Xue et al., 2021) and ByT5 (Xue et al., 2022) paper. The second row is our repetition using the publicly available mT5-XXL checkpoint, which is also the starting point for PaLI-17B. The third row results are using the trained PaLI-17B model.

4.4 Zero-shot Image Classification

We evaluate the PaLI models at 224x224 resolution (before high-resolution pre-finetuning) on ImageNet and ImageNet OOD evaluation sets: original ImageNet (Deng et al., 2009), ImageNet-R (Hendrycks et al., 2021a), ImageNet-A (Hendrycks et al., 2021b), ImageNet-Sketch (Wang et al., 2019b), ImageNet-v2 (Recht et al., 2019) and ObjectNet (Barbu et al., 2019).

We use the same interface as for all other tasks. That is, we do not train a classifier on top of PaLI, but conditioned on the image, use PaLI's decoder to score strings corresponding to each class directly. For each image, each class is scored using the prompt "*Generate alt_text in EN at 2: Photo of* ⟨extra_id_0⟩", scoring against all 1,000 classes with a target "⟨*en_class_name*⟩", where ⟨en_class_name⟩ stands for a classification label in English, such as "goldfish", "great white shark", etc. Flamingo (Alayrac et al., 2022) and GIT2 (Wang et al., 2022a) also evaluate on ImageNet in this "generative style", however, they both perform some adaptation to the dataset (either few-shot prompting or full fine-tuning), whereas we evaluate directly "zero shot". Nonetheless, we include their performances for context.

The results of evaluation are presented in Table 9, which shows improved performance with scale. Additional Top 5 results are in the Appendix (Table 21). There is no precedent for large scale zero-

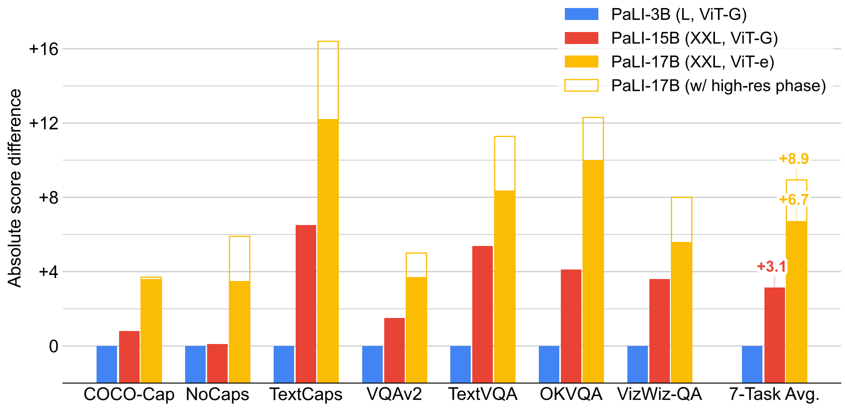


Figure 4: PaLI scaling for a number of V&L tasks; we report CIDEr scores for captioning tasks, and accuracy scores for VQA tasks, for each model size. Both scaling the language-side of the model (from 1B to 13B parameters) and the vision-side of the model (from 2B to 4B parameters) yield improvements across all tasks. The results represented by solid bars are from the standard 224×224 resolution pretraining. The empty orange bars correspond to PaLI-17B results using high-resolution training during the pretraining phase.

shot evaluation on ImageNet with a generative model. However, PaLI outperforms 1-shot learning with Flamingo, and approaches the 5-shot performance. The performance gap with full-finetuning to ImageNet with a generative model (GIT2) is still large, and it is an open question whether a sufficiently large multimodal model can match this performance without seeing any ImageNet examples. Our results indicate that these models can close the gap further, but likely improved training techniques are required.

Model (ImageNet data)	INet	INet-R	INet-A	INet-Sketch	INet-v2	ObjNet
GIT2 (full dataset)	89.22	-	-	-	-	-
Flamingo-80B (1-shot)	71.9	-	-	-	-	-
Flamingo-80B (5-shot)	77.3	-	-	-	-	-
PaLI-3B (0-shot)	70.06	80.15	37.92	61.11	62.55	38.87
PaLI-15B (0-shot)	70.27	81.21	41.16	61.03	62.81	39.51
PaLI-17B (0-shot)	72.11	81.97	44.70	63.83	64.46	42.62

Table 9: Top 1 accuracy results of 0-shot image classification on ImageNet (Deng et al., 2009), ImageNet-R (Hendrycks et al., 2021a), ImageNet-A (Hendrycks et al., 2021b), ImageNet-Sketch (Wang et al., 2019b), Imagenet-v2 (Recht et al., 2019) and ObjectNet (Barbu et al., 2019).

4.5 Model Scaling

Due to the simple modular architecture, the image and language components of PaLI can be scaled independently. In this section, we quantify how scaling affects performance for the V&L benchmarks we consider.

First, we demonstrate that jointly scaling both the capacity both components leads to accuracy improvements. Figure 4 quantifies this increase in accuracy across all 7 V&L benchmarks we consider, and shows that these improvements are noticeable both when scaling the language-model capacity (from L to XXL), and the vision-model capacity (from ViT-G to ViT-e).

Second, we focus on the performance of our vision component, ViT-e. For context, in prior work, V&L scaling is conducted at lower model capacity: for instance, LEMON (Hu et al., 2022) explores

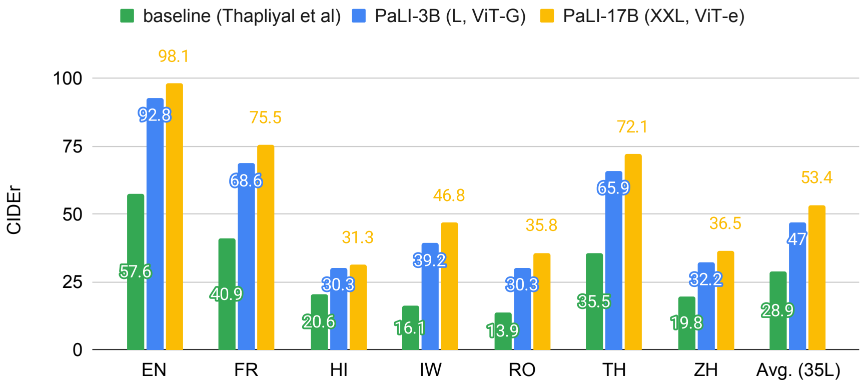


Figure 5: PaLI Scaling performance across multiple languages (See Table 5), using the Crossmodal-3600 benchmark. Larger scale models are important for better performance in these languages, especially low resource ones.

models up to 675M parameters for captioning tasks; CoCa (Yu et al., 2022) scales up both the vision encoder over three model sizes, at 86M, 303M, and 1B, while the corresponding language sizes are 297M, 484M and 1.1B, for a total largest-capacity model of 2.1B parameters; in the latest version of GIT, GIT2 (Wang et al., 2022a), the encoder part is scaled up from 637M parameters to 4.8B parameters. Alternatively, in the larger-capacity–case of Flamingo (Alayrac et al., 2022), a small-capacity image backbone is used, while the scaling to large sizes is done via the language-modeling backbone.

Figure 4 shows that scaling the visual component is important: when scaling from a ViT-G to a ViT-e model, although the overall model size is increased by only about 2B parameters (12% capacity increase), the average performance improvement over all seven benchmarks (additional +3.6) is larger than the one obtained with much larger increases in the capacity of the language model (+3.1). The high-resolution pre-finetuning phase at 588×588 resolution brings an additional +2.2 points, which also indicates the potential of scaling up the vision component of the model.

Finally, the scale of the model impacts performance for multiple languages, especially the scaling of the language-model component, see Figure 5. PaLI has a particularly large relative improvement over the baseline on the languages that the models find harder, and PaLI-17B improves substantially over PaLI-3B (+6.4 CIDEr on average).

4.6 Evaluation of PaLI's Visual Component: ViT-e

Since ViT-e is new and has not been evaluated in the prior work, we evaluate its standalone performance. For this, we perform supervised finetuning on standard classification tasks. Additionally, we perform LiT transfer (Zhai et al., 2022b) to evaluate the frozen representation quality in a zero-shot setup.

	INet-10	INet-25	COCO	VQAv2
ViT-g	84.5	85.4	-	-
ViT-G	84.9	85.6	146.2	80.8
ViT-e	85.2	85.8	149	83.0

Table 10: Impact of scaling ViT. For vision-only tasks, we report 10-shot and 25-shot accuracy on ImageNet. For vision-language tasks, ViT models are paired with the mT5-XXL model in PaLI and we report captioning (COCO) and VQA (VQAv2). For direct comparison, results with ViT-e on COCO and VQAv2 do not include the high resolution phase of pretraining.

Table 10 compares the ViT-e architecture with the smaller ViT-G and ViT-g architectures on vision only and vision-language tasks. The results suggest that V&L tasks could benefit more from scaling up the vision backbone, even on the high end. In Table 11, we finetune the pretrained ViT-e model on the ImageNet dataset, and then report the evaluation scores on several out-of-distribution test variants: ImageNet-v2, ObjectNet, and ReaL (Beyer et al., 2020). We follow the finetuning protocol of Zhai et al. (2022a), but use a 560 × 560 resolution. We evaluate the

Model	INet	INet-v2	INet-R	INet-A	ObjNet	ReaL	VTAB-N
CLIP (Radford et al., 2021)	76.2	70.1	88.9	77.2	72.3	-	73.9
ALIGN (Jia et al., 2021)	76.4	70.1	92.2	75.8	72.2	-	-
BASIC (Pham et al., 2021)	85.7	80.6	95.7	85.6	78.9	-	-
CoCa (Yu et al., 2022)	86.3	80.7	96.5	90.2	82.7	-	-
LiT ViT-g (Zhai et al., 2022b)	85.2	79.8	94.9	81.8	82.5	88.6	74.7
LiT ViT-e (ours)	85.4	80.6	96.1	88.0	84.9	88.4	76.9

Table 12: Zero-shot transfer results of ViT-e on ImageNet, OOD test sets and VTAB-Natural datasets.

finetuned model at 644×644 (Touvron et al., 2019) (chosen according to a held-out 2% of the training set), results are reported in Table 11. ViT-e achieves 90.9% top-1 accuracy on ImageNet and shows clear benefits on the OOD benchmarks.

We follow LiT (Zhai et al., 2022b) to add zero-shot transfer capabilities to the (frozen) ViT-e model, the visual component of PaLI. More specifically, we tune a text encoder, when the ViT image encoder is frozen. We use the English subset of the WebLI dataset for the text encoder training, since all evaluation tasks in Table 12 are in English. These results highlight that going from ViT-g to ViT-e provides consistently better results. Notably, LiT with ViT-e achieves 84.9% zero-shot accuracy on the challenging out-of-distribution ObjectNet test set, setting the new state-of-the-art. The VTAB-Natural benchmark (Zhai et al., 2019) consists of seven diverse natural image datasets, for which LiT also benefits from ViT-e over ViT-g. Detailed results on each VTAB-Natural task are in Appendix E.

Model	INet	INet-v2	ObjNet	ReaL
ViT-G	90.5	83.3	70.5	90.8
CoCa	91.0	-	-	-
ViT-e	90.9	84.3	72.0	91.1

Table 11: ViT-e on ImageNet and OOD test sets.

We also test multilingual performance using WebLI in this setting. We further perform LiT transfer using the same multilingual WebLI dataset as used to train PaLI, and use Crossmodal-3600 to evaluate the cross-lingual image-text retrieval performance. Figure 6 shows that LiT ViT-e pretrained on the English subset substantially outperforms the same model pretrained on the multilingual dataset. The same observation applies to a few languages that are similar to English, e.g. Spanish (es), French (fr), Italian (it). However, the multilingual model performs much better on most other languages, especially those with a non-latin script such as Chinese (zh), Japanese (ja), Korean (ko), and Hebrew (iw). On average (avg), the multilingual LiT ViT-e outperforms the English-only model by a large margin. More results could be found from Appendix Table 22. These results highlight the importance of having good multilingual benchmarks to measure the benefits of training models on diverse datasets such as WebLI.

4.7 Ablations

We ablate the following aspects of our pretraining strategy: (i) the composition of the task mixture; (ii) whether to freeze or fine-tune ViT during pretraining. Table 13 shows different pretraining mixtures for both PaLI-3B and PaLI-15B. We did not evaluate all task combinations due to computational constraints, however, we found that adding more tasks consistently improved performance. Table 14 shows that freezing ViT during pretraining leads to an improvement in downstream finetuning on COCO.

4.8 Limitations

Despite good performance, our model has a number of limitations. For example, the model might not describe very thoroughly a complex scene with many objects because most of the source data does not have complex annotations. We have tried to mitigate this with the object-aware and localization aware queries, added to the data.

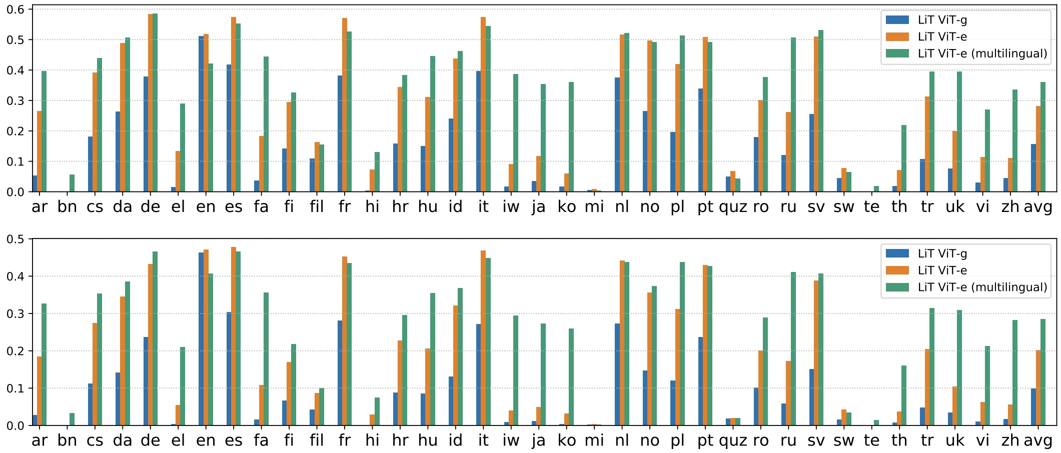


Figure 6: Zero-shot image-text retrieval results on all 36 languages of Crossmodal-3600. Top: image-to-text retrieval accuracy; bottom: text-to-image retrieval accuracy.

Model	Component					Result		
	WebLI	Text dataset	CC3M	VQA	Other	TextVQA	VQAv2 (val)	XM-3600$_{@224}$
PaLI-3B	✓	✓				57.2$_{@378}$	78.4$_{@378}$	91.7 (EN) / 33.6 (6L)
	✓	✓	✓	✓	✓	58.8$_{@378}$	79.3$_{@378}$	92.8 (EN) / 44.4 (6L)
PaLI-15B	✓		✓			-	74.8$_{@224}$	87.1 (EN) / 47.1 (6L)
	✓		✓	✓		-	76.3$_{@224}$	-
	✓	✓	✓	✓	✓	-	77.8$_{@224}$	96.8 (EN) / 49.0 (6L)

Table 13: Advantage of our comprehensive pretraining mixture over mixtures with subsets of components on both VQA and captioning tasks. Results labeled with "@378" are obtained with image resolution 378×378. Other results are all from resolution 224×224. “Other” refers to VQG, Object-Aware (OA) and detection components. The text-only data is a subset of 100M examples from the dataset used in PaLM (Chowdhery et al., 2022).

We also noticed that some of the multilingual capabilities are lost when fine-tuned on English-only data, which is consistent with other model fine-tuning behavior. Ideally these models should be fine-tuned on a mix of multiple datasets including multilingual ones.

There are limitations related to the evaluation procedures of the benchmarks. Since we are evaluating in the open-vocabulary generative setting, for example in VQA, the model might generate a correct response which is a synonym or a paraphrase of the target response and does not match the target exactly. In these cases the answer is counted as incorrect. Fixed-vocabulary approaches do not suffer from these issues, but are limited in generalization beyond the answers of a specific dataset. Further, in terms of evaluation, some benchmarks might need more comprehensive strategies to avoid evaluations with Western-centric bias. Multilingual models and benchmarks are a first step in that direction.

5 Model Fairness, Biases, and Other Potential Issues

Models trained on web data are at risk of being biased or unfair due to biases in that data. A first step towards addressing those risks is being transparent about their existence, and then measuring them. To this end, we add a data card (Pushkarna et al., 2022) for WebLI and model card (Mitchell et al., 2019) for PaLI in the Appendix.

To understand the demographic properties of the data, we sample 112,782 (0.001% of the full data set, randomly sampled due to the limitations of the labeling tool, described next) examples and analyze both images and texts of the sampled data with the Know Your Data (KYD) tool. We use KYD to analyze the perceived gender presentation of image subjects (Schumann et al., 2021) along with gender expressed through pronouns in text. In the sampled images, 54% of people appear feminine

Model	ViT during finetuning	ViT during pretraining	COCO (Karp. test)$_{@224}$
PaLI-3B	Fine-tuned	Frozen Fine-tuned	139.3 138.8
PaLI-15B	Fine-tuned	Frozen Fine-tuned	141.4 140.1

Table 14: Frozen versus fine-tuned ViT during pretraining. In this comparison, COCO is performed at resolution 224×224 rather than higher resolution in the main results.

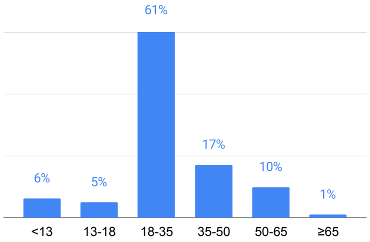


Figure 7: The distribution of ages recognized from the sampled images of WebLI.

presenting with 46% masculine presenting. In the sampled text, female pronouns (e.g., she, her) are used 30% of the time, male pronouns (e.g., he, him) 38% of the time, and they or them (either singular or plural) 31% of the time. We also analyze the perceived age of individuals appearing in the sampled images, resulting in the distribution displayed in Figure 7.

We consider all the effort above a first step, and know that it will be important to continue to measure and mitigate bias as we apply our model to new tasks. Deeper analysis will include the study of the model's recognition capabilities and potential biases observed towards specific attributes, e.g. related to gender, age, etc. and how scaling affects these observations.

6 Conclusions and broader impacts

We present PaLI, a jointly scaled language-and-image model targeting a variety of language-image, language and image tasks. PaLI reuses pretrained unimodal models and capitalizes on their abilities, while also offsetting the substantial cost of their large-scale training efforts. We scale PaLI across both the language and the vision components, and leverage a large language-image training dataset, WebLI, covering over 100 languages. PaLI establishes new state-of-the-art results on multiple image and language tasks, outperforming existing strong models. Our model shows particularly important improvements on multilingual tasks.

Large models may have broader societal impact. While such models have demonstrated strong performance on public benchmarks, they might contain unknown biases or stereotypes, or propagate inaccurate or otherwise distorted information. While we have made efforts to measure some of these issues, such models need to be re-assessed carefully before being used for specific purposes. The dataset used for pretraining is automatically harvested, and filtering of the data is automatic. That process may leave undesirable images or text annotations, descriptions or concepts to be incorporated into the model. We have also attempted to train the model to operate in more than 100 languages, which we believe is an important step forward for image-language models. However, languages have various levels of data presence and coverage, so the language-generated text varies in quality depending on the language, and might contain inaccurate or undesirable outputs.

Acknowledgements

We would like to thank Erica Moreira, Victor Gomes, Tom Small, Sarah Laszlo, Kathy Meier-Hellstern, Susanna Ricco, Emily Denton, Bo Pang, Wei Li, Jihyung Kil, Tomer Levinboim, Julien Amelot, Zhenhai Zhu, Xiangning Chen, Liang Chen, Filip Pavetic, Daniel Keysers, Matthias Minderer, Josip Djolonga, Ibrahim Alabdulmohsin, Mostafa Dehghani, Yi Tay, Rich Lee, Austin Tarango, Elizabeth Adkison, James Cockerille, Eric Ni, Anna Davies, Maysam Moussalem, Jeremiah Harmsen, Claire Cui, Slav Petrov, Tania Bedrax-Weiss, Joelle Barral, Tom Duerig, Paul Natsev, Fernando Pereira, Jeff Dean, and Zoubin Ghahramani.

References

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

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Joshua Tenenbaum, and Boris Katz. 2019. 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.

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

Lucas Beyer, Xiaohua Zhai, and Alexander Kolesnikov. 2022. Big Vision. https://github.com/google-research/big_vision.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. 2018. JAX: composable transformations of Python+NumPy programs.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 1877–1901.

Soravit Changpinyo, Doron Kukliansky, Idan Szpektor, Xi Chen, Nan Ding, and Radu Soricut. 2022a. All you may need for VQA are image captions. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1947–1963.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568.

Soravit Changpinyo, Linting Xue, Idan Szpektor, Ashish V. Thapliyal, Julien Amelot, Xi Chen, and Radu Soricut. 2022b. Towards multi-lingual visual question answering. *arXiv preprint arXiv:2209.05401*.

Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. 2021. Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO Captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942.

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

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

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

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021,*.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. 2022. GLaM: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pages 5547–5569.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. 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.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2021. KAT: A knowledge augmented transformer for vision-and-language. *arXiv preprint arXiv:2112.08614*.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. VizWiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Jonathan Heek, Anselm Levskaya, Avital Oliver, Marvin Ritter, Bertrand Rondepierre, Andreas Steiner, and Marc van Zee. 2020. Flax: A neural network library and ecosystem for JAX.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. 2021a. 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.

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

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421.

Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17980–17989.

Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Mia Xu Chen, Dehao Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V Le, Yonghui Wu, et al. 2019. GPipe: efficient training of giant neural networks using pipeline parallelism. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 103–112.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916.

Norman P Jouppi, Doe Hyun Yoon, George Kurian, Sheng Li, Nishant Patil, James Laudon, Cliff Young, and David Patterson. 2020. A domain-specific supercomputer for training deep neural networks. *Communications of the ACM*, 63(7):67–78.

Da-Cheng Juan, Chun-Ta Lu, Zhen Li, Futang Peng, Aleksei Timofeev, Yi-Ting Chen, Yaxi Gao, Tom Duerig, Andrew Tomkins, and Sujith Ravi. 2019. Graph-rise: Graph-regularized image semantic embedding. *arXiv preprint arXiv:1902.10814*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

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

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. 2020. Big Transfer (BiT): General visual representation learning. *Lecture Notes in Computer Science*, pages 491–507.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2020. The Open Images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981.

Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. 2018. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. 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.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2022. xGQA: Cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511.

Hieu Pham, Zihang Dai, Golnaz Ghiasi, Hanxiao Liu, Adams Wei Yu, Minh-Thang Luong, Mingxing Tan, and Quoc V Le. 2021. Combined scaling for zero-shot transfer learning. *arXiv preprint arXiv:2111.10050*.

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

AJ Piergiovanni, Wei Li, Weicheng Kuo, Mohammad Saffar, Fred Bertsch, and Anelia Angelova. 2022b. Answer-Me: Multi-task learning for generalization to many question-answering tasks. *arXiv preprint arXiv:2205.00949*.

Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data Cards: Purposeful and transparent dataset documentation for responsible AI. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1776–1826.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. 2021. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595.

Adam Roberts, Hyung Won Chung, Anselm Levskaya, Gaurav Mishra, James Bradbury, Daniel Andor, Sharan Narang, Brian Lester, Colin Gaffney, Afroz Mohiuddin, et al. 2022. Scaling up models and data with `t5x` and `seqio`. *arXiv preprint arXiv:2203.17189*.

Candice Schumann, Susanna Ricco, Utsav Prabhu, Vittorio Ferrari, and Caroline Pantofaru. 2021. A step toward more inclusive people annotations for fairness. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, pages 916–925.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. 2019. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2556–2565.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-LM: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

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

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Ashish V. Thapliyal, Jordi Pont-Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. *arXiv preprint arXiv:2205.12522*.

Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. 2021. MLP-Mixer: An all-MLP architecture for vision. *Advances in Neural Information Processing Systems*, 34:24261–24272.

Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Herve Jegou. 2019. Fixing the train-test resolution discrepancy. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 8252–8262.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019a. SuperGLUE: a stickier benchmark for general-purpose language understanding systems. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 3266–3280.

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

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022a. GIT: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022b. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*.

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

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. SimVLM: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, pages 23965–23998.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Zhengyuan Yang, Yijuan Lu, Jianfeng Wang, Xi Yin, Dinei Florencio, Lijuan Wang, Cha Zhang, Lei Zhang, and Jiebo Luo. 2021. TAP: Text-aware pre-training for Text-VQA and Text-Caption. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8751–8761.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. CoCa: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. 2022a. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12104–12113.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. 2019. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022b. LiT: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. VinVL: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588.

A PaLI Model Card

Following (Mitchell et al., 2019), we present the PaLI model card in Table 15. Detailed model card can be found in the link[6].

Model Summary	
Model Architecture	PaLI is a multimodal sequence-to-sequence Transformer (Vaswani et al., 2017) model derived from the T5 (Raffel et al., 2020) encoder-decoder architecture. It takes text tokens and ViT (Dosovitskiy et al., 2021) dense image embeddings as inputs to an encoder and autoregressively predicts discrete text tokens with a decoder.
Input(s)	A pair of image and text.
Output(s)	Generated text.
Usage	
Application	The model is for research prototype and the current version is not available for the public.
Known Caveats	No.
System Type	
System Description	This is a standalone model.
Upstream Dependencies	No.
Downstream Dependencies	No.
Implementation Frameworks	
Hardware & Software	Hardware: TPU v4 (Jouppi et al., 2020). Software: T5X (Roberts et al., 2022), JAX (Bradbury et al., 2018), Flaxformer (Heek et al., 2020) Details are reported in Section 3.4.
Compute Requirements	Reported in Section 3.4.
Model Characteristics	
Model Initialization	The model is initialized from pre-trained language (mT5) (Xue et al., 2021) and Vision Transformer (ViT) (Zhai et al., 2022a; Dosovitskiy et al., 2021) checkpoints.
Model Status	This is a static model trained on an offline dataset.
Model Stats	The largest PaLI model has 17B parameters, which consists of a 13B parameter mT5-XXL model and a 4B parameter ViT-e model. We have also trained 3B and 15B parameter models.
Data Overview	
Training dataset	The model is pre-trained on the following mixture of datasets: WebLI (Table 16), a subset of PaLM/GLaM Dataset (Du et al., 2022; Chowdhery et al., 2022), CC3M-35L (Sharma et al., 2018), VQ2A-CC3M-35L (Changpinyo et al., 2022a), Open Images (Kuznetsova et al., 2020), Visual Genome (Krishna et al., 2017) and Object365 (Shao et al., 2019). Details are reported in Section 3.3.

[6] https://github.com/google-research/google-research/tree/master/pali/pali_model_card.pdf

Evaluation and Fine-tuning Dataset	• **Vision + language tasks** – **Image captioning (English)**: COCO (Chen et al., 2015), NoCaps (Agrawal et al., 2019), TextCaps (Sidorov et al., 2020) – **Image captioning (multilingual)**: Crossmodal-3600 (Thapliyal et al., 2022) – **Visual question answering (English)**: VQAv2 (Goyal et al., 2017), OKVQA (Gui et al., 2021), TextVQA (Singh et al., 2019), VizWiz-QA (Gurari et al., 2018) – **Visual question answering (multilingual)**: xGQA (Pfeiffer et al., 2022), MaXM (Changpinyo et al., 2022b) • **Vision-only tasks** – **Image classification (fine-tuning)**: ImageNet (Deng et al., 2009), ImageNet-V2 (Recht et al., 2019), ObjectNet (Barbu et al., 2019), ReaL (Beyer et al., 2020) – **Image classification (zero-shot)**: ImageNet (Deng et al., 2009), ImageNet-V2 (Recht et al., 2019), ImageNet-R (Hendrycks et al., 2021a), ImageNet-A (Hendrycks et al., 2021b), ImageNet-Sketch (Wang et al., 2019b), ObjectNet (Barbu et al., 2019), ReaL (Beyer et al., 2020), VTAB (Zhai et al., 2019) • **Language-only tasks** – **Natural language inference (English)**: SuperGLUE (Wang et al., 2019a) – Natural language inference (multilingual): XNLI (Conneau et al., 2018) – **Question Answering (multilingual)**: XQuAD (Artetxe et al., 2020), TyDiQA (Clark et al., 2020)
Evaluation Results	
Evaluation Results	Reported in Section 4.
Model Usage & Limitations	
Sensitive Use	The model is capable of open-ended text generations. This model should not be used for any of the unacceptable language model use cases, e.g., generation of toxic speech.
Known Limitations	Reported in Section 4.8.
Ethical Considerations & Risks	Reported in Section 5.

Table 15: PaLI model card.

B WebLI Datasheet

Following (Gebru et al., 2021), we present the WebLI datasheet in Table 16. Detailed data card can be found in the link[7].

Motivation	
For what purpose was the dataset created? Who created the dataset? Who funded the creation of the dataset?	The dataset was created to support Google's vision-language research, such as the large-scale pre-training for image understanding, image captioning, visual question answering, object detection etc.
Any other comments?	No user data is included in the data source. Personally identifiable and privileged data are filtered out during the dataset construction.
Composition	
What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?	Each instance is presented as an image and associated texts (alt-text, page title and OCR) collected from the web.
How many instances are there in total (of each type, if appropriate)?	There are 9,624,017,440 instances in total (about 260 TB in size).
Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?	The dataset is built from the public web pages. It is not a complete set but rather a subset of the publicly available image-text pairs.
What data does each instance consist of?	Each instance consists of 20+ features. Most features are from public web pages; a few are from GCP API. The primary features are image pixels and the associated texts, including alt-text, page title and OCR. Other features include rich image and page meta information (e.g. URL, MIME type) and filter signals (attached to alt-text only).
Is there a label or target associated with each instance?	No.
Is any information missing from individual instances?	No.
Are relationships between individual instances made explicit?	There are no relationships between individual instances.
Are there recommended data splits?	There is only one split containing all the instances of the dataset.
Are there any errors, sources of noise, or redundancies in the dataset?	The dataset is built from the web and only applied a few filters. The data is noisy and redundant images or texts may exist.
Is the dataset self-contained, or does it link to or otherwise rely on external resources?	The dataset is self-contained.
Does the dataset contain data that might be considered confidential?	No.

[7] https://github.com/google-research/google-research/tree/master/pali/webli_data_card.pdf

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?	The dataset likely contains data that might be considered offensive, insulting or threatening as the data is collected from the web. We use algorithmic methods and classifiers to remove sensitive / personal identifiable information (PII) / pornographic images.
Collection Process	
How was the data associated with each instance acquired?	Images, alt-text and meta information are from the public web. Text language is identified via GCP Translation API[8]. OCR is annoated via GCP Vision API[9].
What mechanisms or procedures were used to collect the data?	The data was collected using a variety of pipelines, software programs and GCP APIs to extract and filter images and texts.
If the dataset is a sample from a larger set, what was the sampling strategy?	The dataset is built from a subset of public web pages.
Who was involved in the data collection process?	A team of researchers at Google.
Over what timeframe was the data collected?	2021-2022
Were any ethical review processes conducted?	No.
Preprocessing, cleaning, and labeling	
Was any preprocessing, cleaning, or labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?	The dataset is not annotated. Images which are identified as having adult content are excluded. Empty texts and texts (alt-text, page title and OCR) which are identified as PII are excluded. Images identified as having adult content, with improper shape, or with too many paired-texts are excluded.
Is the software used to preprocess, clean, or label the instances available?	No.
Uses	
Has the dataset been used for any tasks already?	Yes, we use the dataset for pre-training PaLI models.
Is there a repository that links to any or all papers or systems that use the dataset?	No.
What (other) tasks could the dataset be used for?	Vision-only tasks (image classification, object detection etc.), language-only tasks (question answering, natural language inference etc.) and vision+Language tasks (image captioning, visual question answering, image-text retrieval etc.).
Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?	The dataset is in a stable version and will be refreshed in the future to follow data policies.
Are there tasks for which the dataset should not be used?	The dataset should not be used for training any of the unacceptable vision, language or vision-language model use cases, e.g., generation of toxic captions or inappropriate images.

[8] https://cloud.google.com/translate

[9] https://cloud.google.com/vision

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?	No.

Table 16: WebLI datasheet.

C More Details of Pretraining and Finetuning

Dataset mixing ratio for pretraining Table 17 provides the data mixing ratio for pretraining all PaLI variants. See Section 3.3 for the description of each dataset.

	Text-only	WebLI alt-text	OCR	CC3M-35L	VQA	VQG	OA	Detection	Total
Amount (M)	100	1000	100	100	100	100	50	16	1566

Table 17: Mixing ratio of each task for pretraining

Continuation of pretraining at higher image resolution The second stage of pretraining at 588×588 image resolution for PaLI-17B was performed using 512 GCP-TPUv4 chips for an additional 3 days. We simplify the mixture of data in this stage to focus on VQA, captioning and OCR capabilities, by including only the OCR, CC3M-35L and VQ^{2}A in the training mixture and making them equally weighted.

Hyperparameters for finetuning the V&L tasks We performed limited hyperparameter search for finetuning. The train steps is mostly selected based on dataset size. The batch size is selected among {128, 256, 512}, and the initial learning rate among {1e-5, 3e-5, 1e-4}. The optimizer setting for finetuning is the same as the setting for pretraining. Note that we did not perform the hyperparameter sweep over all possible combinations. Table 18 summarizes the hyperparameters corresponding to the main results.

Hyper-parameter	COCO and NoCaps	TextCaps	VQAv2	TextVQA	VizWiz-QA	OKVQA
Dropout			0.1			
LR decay schedule			linear decay to zero			
Train steps	20k	10k	20k	5k	5k	5k
Batch size	512	512	256	256	256	256
Initial (peak) LR	3e-5	1e-4	1e-4	1e-4	1e-4	3e-5

Table 18: Hyper-parameters used in fine-tuning experiments.

D Results on TextCaps, TextVQA and VizWiz-QA without Detected OCR as Input

Table 19 shows the results on TextCaps, TextVQA and VizWiz-QA without the detected OCR strings as input. PaLI slightly suffers without OCR input, while its performance remains close to the first version of GIT. This result may suggest that the significantly larger vocab of PaLI adds further difficulty to OCR string generation.

However, for VizWiz-QA, PaLI establishes SOTA performance without OCR input.

E Detailed VTAB Results

For the VTAB benchmark (Zhai et al., 2019), we follow the methodology outlined in (Zhai et al., 2022b). PaLI sets a new state-of-the-art zero-shot performance for the "natural" subset (see Table 20).

		TextCaps	TextVQA	VizWiz-QA	
Method	OCR input?	test	test	test-dev	test-std
TAP (Yang et al., 2021)	Yes	103.2	53.97	-	-
GIT	No	138.2	59.75	68.0	67.5
GIT2	No	145.0	67.27	71.0	70.1
PaLI	No	135.4	58.80	71.6	70.7
PaLI	Yes	160.4	73.06	74.4	73.3

Table 19: Results on TextCaps, TextVQA and VizWiz-QA with and without detected OCR as input for PaLI

	Caltech101	CIFAR-100	DTD	Flowers102	Pets	Sun397	SVHN	Mean
CLIP	**92.8**	77.5	55.7	78.3	93.5	68.4	**51.0**	73.9
LiT *ViT-g*	79.2	83.6	66.6	**92.3**	97.7	76.0	27.5	74.7
LiT *ViT-e*	79.8	**90.4**	**68.8**	91.2	**98.1**	**76.3**	33.8	**76.9**

Table 20: Accuracies for zero-shot evaluation of different VTAB "natural" tasks, and the average over these tasks. Note that CLIP is using OCR for the SVHN task (as opposed to LiT and PaLI, which do not use OCR).

F Top 5 Accuracy on Zero-shot ImageNet Datasets

Model	INet	INet-R	INet-A	INet-sketch	INet-v2	ObjNet
PaLI-3B	84.31	90.05	55.04	76.47	78.49	53.71
PaLI-15B	84.78	90.91	59.00	76.81	79.54	55.29
PaLI	86.18	91.51	62.72	79.30	80.71	58.35

Table 21: Top 5 accuracy results of Zero-shot image classification on ImageNet (Deng et al., 2009), ImageNet-R (Hendrycks et al., 2021a), ImageNet-A (Hendrycks et al., 2021b), ImageNet-Sketch (Wang et al., 2019b), ImageNet-v2 (Recht et al., 2019) and ObjectNet (Barbu et al., 2019).

G Zero-shot Image-text Retrieval Results on Crossmodal-3600

Language	Image-to-text			Text-to-image		
	LiT ViT-g	LiT ViT-e	LiT ViT-e (multilingual)	LiT ViT-g	LiT ViT-e	LiT ViT-e (multilingual)
ar	5.28	26.58	39.69	2.80	18.46	32.60
bn	0.00	0.11	5.67	0.00	0.06	3.31
cs	18.19	39.25	44.03	11.24	27.35	35.24
da	26.44	48.92	50.75	14.07	34.43	38.48
de	37.83	58.42	58.53	23.61	43.25	46.50
el	1.56	13.47	29.03	0.39	5.46	20.92
en	51.22	51.78	42.11	46.24	47.07	40.63
es	41.81	57.50	55.22	30.29	47.71	46.55
fa	3.78	18.39	44.50	1.57	10.74	35.58
fi	14.14	29.42	32.64	6.59	16.91	21.80
fil	10.94	16.39	15.53	4.18	8.66	10.04
fr	38.28	57.06	52.61	28.02	45.20	43.47
hi	0.47	7.33	13.14	0.08	2.90	7.42
hr	15.86	34.47	38.31	8.80	22.72	29.55
hu	15.11	31.17	44.67	8.45	20.52	35.49
id	24.11	43.72	46.33	12.99	32.08	36.75
it	39.69	57.47	54.53	27.07	46.79	44.76
iw	1.75	9.11	38.67	0.86	3.99	29.39
ja	3.61	11.67	35.47	1.20	4.91	27.24
ko	1.78	6.00	36.11	0.35	3.14	25.95
mi	0.58	0.92	0.33	0.19	0.30	0.22
nl	37.47	51.67	52.14	27.26	44.08	43.79
no	26.53	49.69	49.17	14.61	35.59	37.35
pl	19.67	42.03	51.42	12.00	31.13	43.72
pt	33.92	50.81	49.19	23.58	42.97	42.73
quz	5.08	6.83	4.31	1.85	1.89	1.90
ro	17.94	30.08	37.75	10.15	20.06	28.82
ru	12.00	26.22	50.64	5.76	17.19	41.11
sv	25.50	51.00	53.22	15.11	38.80	40.66
sw	4.47	7.75	6.42	1.58	4.17	3.41
te	0.06	0.03	1.92	0.03	0.03	1.42
th	1.89	7.22	22.00	0.79	3.71	16.06
tr	10.72	31.28	39.50	4.73	20.42	31.47
uk	7.67	19.94	39.53	3.38	10.40	30.81
vi	3.08	11.44	27.08	0.98	6.22	21.28
zh	4.53	11.11	33.61	1.67	5.60	28.24
avg	15.64	28.23	35.99	9.79	20.14	28.46

Table 22: Image-to-text and text-to-image zero-shot retrieval results on all 36 languages of Crossmodal-3600. Models are trained following LiT (Zhai et al., 2022b) method with diverse visual backbones (ViT-g or ViT-e) and datasets (English or multilingual).