

Flamingo: a Visual Language Model for Few-Shot Learning

Jean-Baptiste Alayrac^{*,‡}, Jeff Donahue^{*}, Pauline Luc^{*}, Antoine Miech^{*}, Iain Barr[†], Yana Hasson[†], Karel Lenc[†], Arthur Mensch[†], Katie Millican[†], Malcolm Reynolds[†], Roman Ring[†], Eliza Rutherford[†], Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, Karen Simonyan^{*,‡}

^{*}Equal contributions, ordered alphabetically, [†]Equal contributions, ordered alphabetically, [‡]Equal senior contributions

Building models that can be rapidly adapted to numerous tasks using only a handful of annotated examples is an open challenge for multimodal machine learning research. We introduce Flamingo, a family of Visual Language Models (VLM) with this ability. Flamingo models include key architectural innovations to: (i) bridge powerful pretrained vision-only and language-only models, (ii) handle sequences of arbitrarily interleaved visual and textual data, and (iii) seamlessly ingest images or videos as inputs. Thanks to their flexibility, Flamingo models can be trained on large-scale multimodal web corpora containing arbitrarily interleaved text and images, which is key to endow them with in-context few-shot learning capabilities. We perform a thorough evaluation of the proposed Flamingo models, exploring and measuring their ability to rapidly adapt to a variety of image and video understanding benchmarks. These include open-ended tasks such as visual question-answering, where the model is prompted with a question which it has to answer, captioning tasks, which evaluate the ability to describe a scene or an event, and close-ended tasks such as multiple choice visual question-answering. For tasks lying anywhere on this spectrum, we demonstrate that a *single* Flamingo model can achieve a new state of the art for few-shot learning, simply by prompting the model with task-specific examples. On many of these benchmarks, *Flamingo* actually surpasses the performance of models that are fine-tuned on thousands of times more task-specific data.

1	Introduction	3	4	Experiments	18
1.1	Challenges of multimodal generative modelling	4	4.1	Training and evaluation setting	18
1.2	Contributions	5	4.2	Few-shot learning with Flamingo models	22
2	Related work	5	4.3	Fine-tuning <i>Flamingo</i> as a pretrained vision-language model	25
2.1	Language modelling	6	4.4	Ablation studies	25
2.2	Joint vision and language modelling .	6	5	Qualitative results	30
2.3	Web-scale vision and language training datasets	8	6	Discussion	33
2.4	Few-shot learning in vision	8	6.1	Limitations, failure cases and opportunities	33
3	Approach	9	6.2	Benefits, risks and mitigation strategies	36
3.1	Flamingo models architecture	10	7	Conclusion	38
3.2	Training on a mixture of vision and language datasets	14	8	Appendix	52
3.3	Task adaptation with few-shot in-context learning	16			

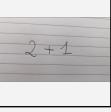
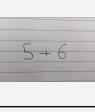
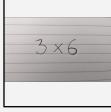
Input Prompt	Completion	
 <p>This is a chinchilla. They are mainly found in Chile.</p>  <p>This is a shiba. They are very popular in Japan.</p>  <p>This is</p>	→ a flamingo. They are found in the Caribbean and South America.	
 <p>What is the title of this painting? Answer: The Hallucinogenic Toreador.</p>  <p>Where is this painting displayed? Answer: Louvre Museum, Paris.</p>  <p>What is the name of the city where this was painted? Answer:</p>	→ Arles.	
 <p>Output: "Underground"</p>  <p>Output: "Congress"</p>  <p>Output:</p>	→ "Soulomes"	
 <p>2+1=3</p>  <p>5+6=11</p>  <p>3x6=18</p>	→ 3x6=18	
 <p>Output: A propaganda poster depicting a cat dressed as French emperor Napoleon holding a piece of cheese.</p>  <p>Output: A pink room with a flamingo pool float.</p>  <p>Output:</p>	→ A portrait of Salvador Dali with a robot head.	
 <p>Les sanglots longs des violons de l'automne blessent mon cœur d'une langueur monotone.</p>  <p>Pour qui sont ces serpents qui sifflent sur vos têtes?</p> 	→ Je suis un cœur qui bat pour vous.	
 <p>pandas: 3</p>  <p>dogs: 2</p> 	→ giraffes: 4	
<p>I like reading</p>  <p>, my favourite play is Hamlet. I also like</p>  <p>, my favorite book is</p>	→ Dreams from my Father.	
 <p>What happens to the man after hitting the ball? Answer:</p>	→ he falls down.	
 <p>This is a picture of two teddy bears on the moon.</p> <p>What are they doing?</p> <p>They are having a conversation.</p> <p>What object are they using?</p> <p>It looks like a computer.</p> <p>Is this surprising?</p> <p>Yes, it is surprising.</p> <p>Why is this picture surprising to you?</p> <p>I think it is surprising because teddy bears are not usually found on the moon.</p>	 <p>What is the common thing about these three images?</p> <p>They are all flamingos.</p> <p>What is the difference between these three images?</p> <p>The first one is a cartoon, the second one is a real flamingo, and the third one is a 3D model of a flamingo.</p>	 <p>This is an apple with a sticker on it.</p> <p>What does the sticker say?</p> <p>The sticker says "iPod".</p> <p>Where is the photo taken?</p> <p>It looks like it's taken in a backyard.</p> <p>Do you think it is printed or handwritten?</p> <p>It looks like it's handwritten.</p> <p>What color is the sticker?</p> <p>It's white.</p>

Figure 1 | Examples of inputs and outputs obtained from our 80B parameter model, *Flamingo*. Like large-scale language models, *Flamingo* can be rapidly adapted to various image and video understanding tasks by simply prompting it with a few examples (top). Out of the box, *Flamingo* is also capable of rich visual dialogue (bottom). More qualitative examples can be found in Section 5.

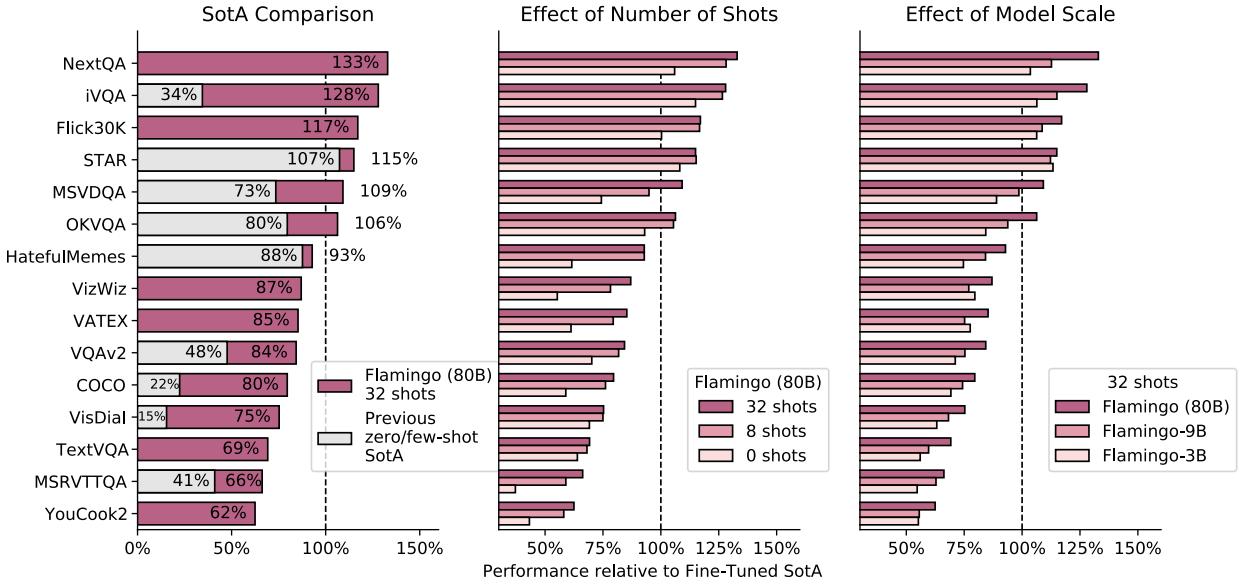


Figure 2 | Overview of the results of the Flamingo models. *Left:* Our largest model, dubbed *Flamingo*, outperforms state-of-the-art fine-tuned models on six out of the 16 tasks we consider despite not using any fine-tuning at all. For all 16 tasks where published few-shot results are available, *Flamingo* outperforms them by a large margin and sets the new few-shot state of the art. *Center:* *Flamingo* performance improves with the number of shots. *Right:* The performance of the *Flamingo* models increases with the model scale. *Note:* We omit RareAct, our 16th benchmark, as it is a zero-shot benchmark with no available fine-tuning results.

1. Introduction

One key aspect of intelligence is the ability to quickly learn how to perform a new task given a short instruction (Griffiths et al., 2019; Markman, 1989). While initial progress has been made towards a similar capability in computer vision, the most used paradigm still consists of first pretraining on a large amount of supervised multimodal data collected from the web, before fine-tuning the model on the task of interest (Lu et al., 2019; Wang et al., 2021; Zellers et al., 2022). However, fine-tuning often requires many thousands of annotated data points in order to succeed. On top of the annotation burden, this approach often requires careful per-task hyperparameter tuning and is also resource intensive. Recently, multimodal vision-language models trained with a contrastive objective (Jia et al., 2021; Radford et al., 2021) have enabled zero-shot adaptation to novel tasks, without the need for fine-tuning. However, because these models simply provide a similarity score between a text and an image, they can only tackle limited use cases such as classification, where a finite set of outcomes is provided beforehand. They crucially lack the ability to generate language, which makes them less suitable to more open-ended tasks such as captioning or visual question-answering. Others have explored visually conditioned language generation (Cho et al., 2021; Tsimpoukelli et al., 2021; Wang et al., 2022, 2021; Xu et al., 2021) but have not yet shown good performance in low data regimes. In this paper, we introduce *Flamingo*, a Visual Language Model (VLM) that sets a new state of the art in few-shot learning on a wide range of open-ended vision and language tasks, simply by being prompted with a few input/output examples, as illustrated in Figure 1. Of the 16 tasks we consider, *Flamingo* also surpasses the fine-tuned state of the art in 6 of the cases, despite using orders of magnitude less task-specific training data (see Figure 2).

To achieve this, *Flamingo* takes inspiration from recent work in large-scale generative language models (LMs) which are good few-shot learners (Brown et al., 2020; Chowdhery et al., 2022; Hoffmann et al., 2022; Rae et al., 2021). A single large LM can indeed achieve strong performance on many

tasks using only its text interface: a few examples of a task are provided to the model as a prompt, along with a query input, and the model generates a continuation to produce a predicted output for the task on that query. In principle, the same can be done for many image and video understanding tasks such as classification, captioning, or question-answering: these can be cast as text prediction problems with visual input conditioning. The difference from a LM is that the model must be able to ingest a multimodal prompt containing both image and/or videos interleaved with text. The Flamingo model has this capability – it is a visually-conditioned autoregressive text generation model able to ingest a sequence of text tokens interleaved with images and/or videos, and produce text as output. Flamingo models fuse large LMs with powerful visual embeddings – each separately pretrained and frozen – by adding novel architecture components in between.

A crucial aspect for the performance of large LMs is that they are trained on a large amount of text data. This training provides general-purpose generation capabilities that allows these LMs to perform well when prompted with task examples. Similarly, we demonstrate here that the way we train the Flamingo models matters greatly for their final performance. They are trained on a carefully chosen mixture of complementary large-scale multimodal data coming only from the web, *without using any data annotated for machine learning purposes*. After this training is done, a Flamingo model can be directly adapted to vision tasks via simple few-shot learning without any additional task-specific tuning.

1.1. Challenges of multimodal generative modelling

Although similar in spirit to large language models, the multimodality of our setting poses a number of challenges not present in the language-only domain. In the following we detail some of these challenges and briefly describe how we overcome them with Flamingo.

Unifying strong single-modal models. Training a multimodal vision and language model, starting from a pretrained LM, has the potential to save the immense computational resources used to train the LM in the first place. However, a LM trained only on text has no built-in means of incorporating inputs from other modalities. Introducing such inputs to a pretrained LM is one challenge in our setting: it is crucial to keep the pretrained model’s language understanding and generation capabilities fully intact, avoiding any destabilising interventions, while still leveraging the model’s full capacity and depth on the inputs from the new modality. To address this difficulty, we propose to interleave cross-attention layers with regular language-only self-attention layers that are kept frozen during training. To minimize the effect of these newly added layers at initialization, we introduce specific gating mechanisms which greatly improve stability and final performance.

Supporting both images and videos. The 2D spatial structure and high dimensionality of images and videos (of even modest resolution) is not immediately amenable to the homogeneous treatment as a 1D sequence commonly used in unimodal text generation. Transformer models in particular, which the latest LMs build on, would suffer from memory limitations with the naive addition of high-dimensional visual data into the sequence – computation scales quadratically with the sequence length (number of tokens), for example. Even when based on Transformer architectures, state-of-the-art vision architectures often rely on local 2D priors to improve efficiency by making use of inductive biases not suitable for text (Liu et al., 2021). A unified treatment of static images and videos poses additional difficulties. To address this challenge we use a Perceiver-based (Jaegle et al., 2021) architecture that can produce a small fixed number of visual tokens (around a hundred) per image/video, given a large varying number of visual input features (up to several thousand). We show that this approach makes it possible to scale to large inputs while still retaining model expressivity.

Obtaining heterogeneous training data to induce good generalist capabilities. Training large models with billions of parameters successfully requires huge datasets. Paired image / caption datasets (Jia et al., 2021; Radford et al., 2021) alone may not be general enough to induce few-shot learning and task induction capabilities à la GPT-3. Several large-scale datasets for language have been collected semi-automatically by scraping billions of web pages at a large scale (Brown et al., 2020; Rae et al., 2021; Raffel et al., 2019), but an equivalent multimodal dataset which induces similar few-shot abilities has yet to be collected. To overcome this challenge, we explore how to obtain a multimodal dataset by scraping text and images from web pages. Examples from the latter consist of text with interleaved images, corresponding to the page layout when viewed in a browser. However, despite the generality of such data, the images and text are often only weakly related. To address this issue, we combine this dataset with standard paired image/text and video/text datasets, where the visual and language are typically more strongly related. We demonstrate that this careful mixture of datasets is essential for Flamingo’s general few-shot task induction capability and to ensure highly relevant generated outputs for visual inputs.

1.2. Contributions

The contributions of this work are the following:

1. We introduce the Flamingo family of Visual Language Models (VLMs) which can perform various multimodal tasks (such as captioning, visual dialogue, classification or visual question answering) from only a few input/output examples. This is enabled by the following contributions:
 - (a) A novel architecture for accepting arbitrarily interleaved visual data and text as input and generating output text in an open-ended manner.
 - (b) Architectural innovations and training strategies that effectively leverage large pretrained vision-only and language-only models, preserving the benefits of these initial models while efficiently fusing the modalities. Starting from Chinchilla, a 70B state-of-the-art LM (Hoffmann et al., 2022), we train *Flamingo*, an 80B parameter VLM.
 - (c) Efficient ways to adapt to visual inputs of varying size, making Flamingo applicable to images and videos.
2. We quantitatively evaluate how Flamingo models can be adapted to various tasks via few-shot learning. We notably reserve a large set of held-out benchmarks which have not been used for validation of any design decisions or hyperparameters of the approach. We use these to estimate unbiased few-shot performance.
3. *Flamingo* sets a new state of the art in few-shot learning on a wide array of 16 multimodal language and image/video understanding tasks. In 6 out of these 16 tasks, *Flamingo* also outperforms the fine-tuned state of the art, despite using only 32 task-specific examples which is around 1000 times less task-specific training data than current state-of-the-art. With a larger annotation budget, *Flamingo* can also be effectively fine-tuned to set a new state of the art on five additional challenging benchmarks: VQAv2, VATEX, VizWiz, MSRVTTQA, and HatefulMemes.

2. Related work

We review relevant related work. First, we give a survey of the field of large-scale language modelling, which has significantly influenced our approach. Next, we give a review on joint vision and language modelling and contrast our work with the recent myriad of vision-language models. As getting the right training data is key, we cover the literature exploring multimodal training data from the web, and how our training data differs. Finally, we give an overview of few-shot learning approaches.

2.1. Language modelling

The field of language modelling has made substantial progress in recent years, following the introduction of Transformers (Vaswani et al., 2017), which improved the modelling of long-range dependencies over RNN-based approaches, while significantly increasing the throughput of models and therefore the amount of data seen during training. The paradigm of first pretraining the model on a vast amount of noisy data, then adapting the model for downstream usage has become standard, following earlier work on language modelling with RNNs from Graves (2013); Howard and Ruder (2018); Jozefowicz et al. (2016); Mikolov et al. (2010). Pretraining may use a masked language modelling loss (BERT and T5, Devlin et al., 2018; Raffel et al., 2019) or a next-token prediction loss (Sutskever et al., 2011). Our approach consists of training a visual language model on a large-scale vision and text dataset scraped from the web, a multimodal equivalent of C4 (Raffel et al., 2019) and The Pile (Gao et al., 2020).

In the last two years, following the success of GPT-3 (Brown et al., 2020), vast improvements have been obtained by increasing the size of language models. This trend has been justified by the findings of Kaplan et al. (2020), who show that language modelling performance is strongly correlated with model size. Recently, Hoffmann et al. (2022) have refined these findings, showing that the number of data tokens should scale at the same rate as the model size to maximise computational efficiency. Based on these findings, they introduced the Chinchilla family of models, which we build upon, using the 70B parameter Chinchilla model as the base LM for our largest Flamingo model.

Large-scale pretrained language models may be adapted in different ways to specific downstream tasks. GPT-3 popularized the use of in-context learning, which entails prompting an autoregressive language model with a few pairs of (*input*, *expected output*), followed by a query *input*. This approach is appealing in that no further training is required for downstream usage. Other downstream adaptation techniques typically involve adding to or modifying a limited number of parameters of the model based on data from a downstream task—limiting the dimension of the optimisation problem can prevent catastrophic forgetting (McCloskey and Cohen, 1989) of the pretraining task. One notable approach involves directly learning a prefix or prompt in embedding space (Lester et al., 2021; Li and Liang, 2021). Alternatively, Zaken et al. (2021) explores fine-tuning a fraction of the original language model, or the biases only. Finally, Houlsby et al. (2019) adds a few adapter layers on top and in the middle of the model and trains only these adapters, similar to the way we modify the computational graph of the original language model in our approach. Adapter techniques come in many different forms in the multimodal setting; we detail them in the next section.

2.2. Joint vision and language modelling

These LM breakthroughs have been influential for vision and language modelling. We cover three families of related models. For simplicity, we omit the field of language conditioned visual generation.

Multimodal BERT-based approaches. BERT (Devlin et al., 2018) inspired a large number of follow-up multimodal works. Numerous works such as Chen et al. (2020); Fu et al. (2021); Gan et al. (2020); Hendricks et al. (2021); Li et al. (2020,); Lu et al. (2019); Singh et al. (2021); Su et al. (2019); Tan and Bansal (2019); Wang et al. (2021,); Zellers et al. (2021, 2022); Zhu and Yang (2020) often apply a pretrained object detector on images or videos to obtain visual region proposals, treated as *visual words*. VideoBERT (Sun et al., 2019) instead tokenizes video frames as visual words using *k*-means. These *visual tokens* are later embedded and fed, together with text tokens, into a bi-directional transformer. In addition to the MLM loss (Devlin et al., 2018), a masked region modelling loss (MRM) (Chen et al., 2020; Lu et al., 2019) is often applied by masking visual tokens. A cross-modal

matching loss (Chen et al., 2020; Lu et al., 2019; Singh et al., 2021; Tan and Bansal, 2019; Zhu and Yang, 2020) can also be used to predict whether pairs of vision and text inputs are matching. We differ from this family of models as Flamingo models do not require fine-tuning on new tasks. Moreover, Flamingo models can generate text which is not natively the case for BERT-style models.

Contrastive dual encoder approaches. More recently, a large family of vision-language models, based on contrastive learning (Alayrac et al., 2020; Bain et al., 2021; Jain et al., 2021; Jia et al., 2021; Li et al., 2021; Miech et al., 2020; Pham et al., 2021; Radford et al., 2021; Yao et al., 2021; Yuan et al., 2021; Zhai et al., 2021) have emerged. They often encode vision and text inputs with separate encoders, producing individual vision and language vectors embedded into a joint space using a contrastive loss. The strength of contrastive approaches is their capability to learn a highly generic visual representation, and to do so efficiently at scale. Moreover, when pretrained on a large and diverse dataset (Jia et al., 2021; Miech et al., 2020; Pham et al., 2021; Radford et al., 2021; Zhai et al., 2021), strong zero-shot vision-text retrieval and classification performance can be obtained. Unfortunately, as they are only trained to match visual data to text description, these models can only be adapted to close-ended tasks. Finally, it is challenging to adapt contrastive models using a handful of examples. In fact, Radford et al. (2021) indicated that using as few as two training examples per-class actually *decreased* the CLIP zero-shot performance. In contrast, Flamingo models can significantly improve with as few as four examples. In our work, we leverage contrastive learning as a technique to pretrain our vision encoder on billions of web images with text descriptions.

Visual language models (VLM). Most similar to our work are the visual language models able to generate text in an autoregressive manner (Donahue et al., 2015; Hu et al., 2021; Luo et al., 2020; Vinyals et al., 2015). The first application (Donahue et al., 2015; Vinyals et al., 2015) was visual captioning where the goal is to describe an image with text. VirTex (Desai and Johnson, 2021) proposed captioning as a pretext task to learn a visual representation from text descriptions. Related to our approach, CM3 (Aghajanyan et al., 2022) also proposes to go beyond the generation of a single caption for images but instead, generate the content of HTML web pages containing images. Recently, several concurrent works (Cho et al., 2021; Li et al., 2022; Wang et al., 2022, 2021; Zhu et al., 2021) also propose to formulate numerous vision tasks as text generation problems, including classification, visual question answering, visual entailment, visual captioning, visual grounding and even object detection (Chen et al., 2021). They simplified prior work relying on cumbersome task-specific heads.

Training large-scale language models is data-hungry and highly computationally demanding (Hoffmann et al., 2022). As a consequence, building on top of a powerful pretrained text-only language model has been explored in concurrent work. VisualGPT (Chen et al., 2021) showed the benefit of initializing the weights of a VLM with a pretrained language-only model for data-efficient training. Tsimpoukelli et al. (2021) pushed this idea even further by *freezing* the pretrained LM weights. The conditioning of the model on vision is similar to the idea of prefix tuning (Lester et al., 2021; Li and Liang, 2021; Zhou et al., 2021; Zhu et al., 2021) (also called *prompt tuning*) where the prefix is encoded by a trainable visual encoder. The benefit of building on top of a strong frozen language model is that it may enable the visual language model to retain similar powerful language-only abilities such as few-shot language adaptation, external knowledge retrieval, or dialogue capabilities. Several works followed this idea with architectural differences in the conditioning of the frozen language model. For instance, MAGMA (Eichenberg et al., 2021) adds bottleneck adapters (Houlsby et al., 2019; Sung et al., 2021) within the frozen language model; ClipCap (Mokady et al., 2021) proposes to use a vision-to-prefix transformer to map the vision features into a prefix instead of using a simple linear layer mapping. VC-GPT (Luo et al., 2022) moves away from the visual prefix tuning approach. They instead explore conditioning the frozen language model by grafting new learnt layers

to the frozen language model. Finally, PICA (Yang et al., 2021) and Socratic Models (Zeng et al., 2022) propose to use off-the-shelf vision-language models (Radford et al., 2021; Zhang et al., 2021) to communicate the content of images using language descriptions to GPT-3 (Brown et al., 2020).

The Flamingo models share numerous ideas with some of the aforementioned VLMs: (i) we rely on a frozen pretrained language model, (ii) we also make use of a transformer-based mapper between the vision encoder and the frozen language model, (iii) we train cross-attention layers interleaved with the frozen language model layers. We differ from existing work as:

- Flamingo models can be rapidly adapted, *without fine-tuning*, to new tasks using few-shot examples, outperforming several fully supervised task-specific state-of-the-art models.
- The architecture of Flamingo models is versatile enough to ingest sequences of arbitrarily interleaved text, videos, and images.

2.3. Web-scale vision and language training datasets

Manually annotated vision and language datasets are costly to obtain and thus relatively small (10k-100k) in scale (Antol et al., 2015; Chen et al., 2015; Marino et al., 2019; Wang et al., 2019; Xiao et al., 2021; Young et al., 2014). Several recent works instead leveraged web scraping to automatically generate aligned visuals and text. For example, some approaches pair images (Changpinyo et al., 2021; Jia et al., 2021; Sharma et al., 2018; Thomee et al., 2016) or videos (Bain et al., 2021) with available alt-text or video descriptions. Miech et al. (2019); Zellers et al. (2021, 2022) explored the use of video with speech transcribed in text as supervision. These works only considered pairs of single images/videos and text descriptions as training data. In addition to this type of paired data, we show the importance of also training on entire multimodal webpages containing interleaved images and text as a single sequence. CM3 (Aghajanyan et al., 2022) follows a similar approach, additionally generating the full HTML markup from pages, while we simplify the text prediction task for the base LM by scraping only the natural language text from the title and main body of the webpage. Finally, we emphasize that *Flamingo* achieves state-of-the-art performance across a wide range of benchmarks without training on commonly used and curated datasets such as VQAv2, COCO or ImageNet. Instead, *Flamingo* is trained solely on *task-agnostic web scraped data*.

2.4. Few-shot learning in vision

Few-shot learning has been extensively studied in computer vision (Fei-Fei et al., 2006; Hariharan and Girshick, 2017). With just a few examples, models can rapidly adapt to new tasks. Several benchmarks have been created to measure progress in this area such as miniImagenet (Vinyals et al., 2016), Omniglot (Lake et al., 2011), Meta-Dataset (Triantafillou et al., 2019), Dax-Blicket Fast Binding dataset (Tsimploukelli et al., 2021) or ORBIT (Massiceti et al., 2021). One category of approaches to few-shot learning is based on analysing the similarities between query and support examples (Doersch et al., 2020; Snell et al., 2017; Tian et al., 2020; Vinyals et al., 2016). Another, based on optimisations using gradient updates, aims at finding a good model initialization that is well-suited for adaptation (Bertinetto et al., 2018; Finn et al., 2017; Zintgraf et al., 2019). In contrast, other work (Bertinetto et al., 2016; Gordon et al., 2018; Requeima et al., 2019) use support examples to directly generate model weights adapted to the novel task. A final category is based on the idea of “in-context” few-shot prompts for language models (Brown et al., 2020) and it has recently been extended to images and language (Tsimploukelli et al., 2021; Yang et al., 2021). This category is the basis of our work. We follow the in-context learning paradigm and also propose the first model for in-context few-shot video learning.

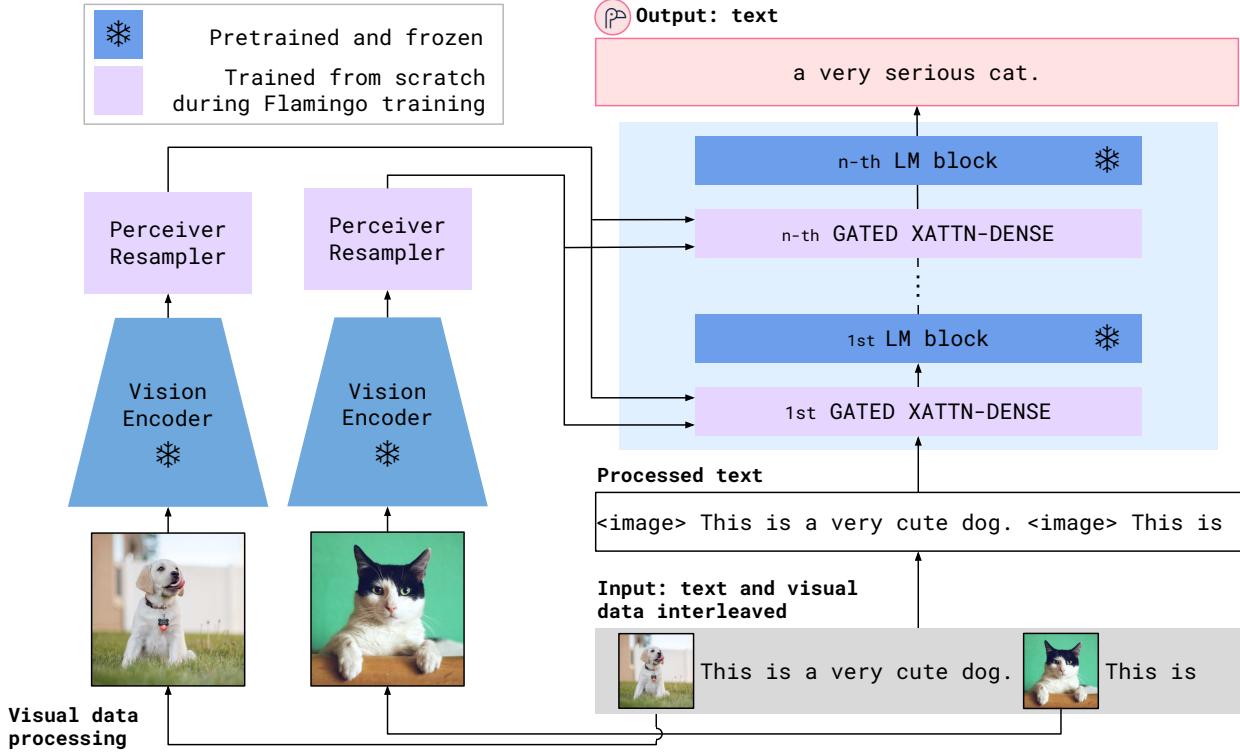


Figure 3 | Overview of the Flamingo model. The Flamingo models are a family of visual language model (VLM) that can take as input visual data interleaved with text and can produce free-form text as output. Key to its performance are novel architectural components and pretraining strategies described in Section 3.

3. Approach

This section describes our approach to building the Flamingo model: a visual language model that accepts text interleaved with images/videos as input and outputs free-form text. Despite its apparent simplicity, this API is sufficiently expressive to tackle a diverse range of tasks. In particular, it handles both open-ended tasks such as visual question-answering or captioning, which require generating text, and close-ended tasks such as classification, which require choosing the best category or answer amongst a set. Most importantly, it is amenable to few-shot in-context learning where examples of annotated visual and text pairs are provided in an interleaved prompt to steer the model to a desired task behavior, without having to change or adapt the model weights (see Section 3.3).

The architectural components shown in Figure 3 are key to the performance of Flamingo models, and these are chosen according to two objectives. The first objective is to leverage pretrained models without having to spend compute training them from scratch. On the vision side, we pretrain a vision encoder with a contrastive text-image approach, à la CLIP (Radford et al., 2021). The role of this model is to extract semantic spatial features that describe attributes that one would want to query about a visual datum: color, shape, nature, positions of objects, etc. On the language side, we start from an existing autoregressive language model (LM) trained on a large and diverse text corpus (Hoffmann et al., 2022). By doing so, Flamingo models gain strong generative language abilities and access to a large amount of knowledge stored in the LM weights. The second objective is to bridge these pretrained models harmoniously. To do so, we freeze the weights of these models so that their initial capacity remains unchanged. We then link them via two learnable architecture components. First, the Perceiver Resampler (Section 3.1.1) receives spatio-temporal features from the Vision Encoder (obtained from either image or video) and outputs a fixed-size set of visual tokens.

Second, those visual tokens are used to condition the frozen LM using freshly initialised cross attention layers (Section 3.1.2) that are interleaved between the pretrained LM layers. These new layers offer an expressive way for the LM to incorporate visual information for the next-token prediction task.

An important aspect of the Flamingo models is that they can model the likelihood of text y interleaved with a sequence of images/videos x as illustrated in Figure 3. More formally, we model the visually conditioned text likelihood as follows

$$p(y|x) = \prod_{\ell=1}^L p(y_\ell|y_{<\ell}, x_{\leq \ell}), \quad (1)$$

where y_ℓ is the ℓ -th language token composing the input text, $y_{<\ell}$ is the set of preceding tokens, $x_{\leq \ell}$ is the set of images/videos preceding token y_ℓ in the interleaved sequence and p is parametrized by a Flamingo model. This setting is formalized in more details in Section 3.1.3. Such interleaved modeling is enabled in the Flamingo architecture by cross attention causal masks (see Figure 6) that specify the conditional dependencies within the multimodal sequence accordingly. The ability to handle interleaved text and visual sequences makes it natural to use Flamingo models for in-context few-shot learning, analogously to GPT-3 with few-shot text prompting. To condition a Flamingo model for a new multimodal task, one simply composes a few-shot prompt by alternating visual inputs and expected text responses, followed by a final “test” image or video (Figure 8). Once this prompt is provided to the model, one can either sample output text, or evaluate the probability of a fixed set of completions (see Section 3.3). The model is trained by maximizing the likelihood of Equation (1) on a diverse mixture of datasets described in Section 3.2.

3.1. Flamingo models architecture

This section goes into the details of our architecture. We first describe the visual stack of our method, that handles the input x , and the text generative decoder producing the sequence y . We then formalize how the visual data is interleaved with the text and how this influences the language generation process.

3.1.1. Visual processing and the Perceiver Resampler

Vision encoder: from pixels to features. Our model’s vision encoder is a pretrained Normalizer-Free ResNet (NFNet) (Brock et al., 2021) – we use the F6 model. We chose this family of visual backbones thanks to their excellent trade-off between performance and efficiency given our hardware. We pretrain the vision encoder using a contrastive objective on our datasets of image and text pairs, using the two-term contrastive loss from Radford et al. (2021). Contrastive similarities are computed as the dot product of the mean pooling of the image encoder output, and the mean pooled output of a BERT model (Devlin et al., 2018). The performance of our contrastive pretrained model is given in Section 4.2.3 and more details about the contrastive training are given in Appendix C.2.2. During the main training phase of the Flamingo models, we leave the vision encoder frozen as it performs favorably when compared to training the vision model directly from our text generation objective (see ablations in Section 4.4). The output of the final stage is a 2D spatial grid of features X_f , which is then flattened to 1D as shown in Figure 4. Our model also handles video inputs. In this setting, frames are sampled at 1 FPS, encoded independently to obtain a sequence of T feature maps X_f which are then concatenated before being fed to downstream components (Figure 4).

Perceiver Resampler: from varying-size large feature maps to few visual tokens. The Perceiver Resampler module connects the vision encoder to the frozen language model as shown in the overall

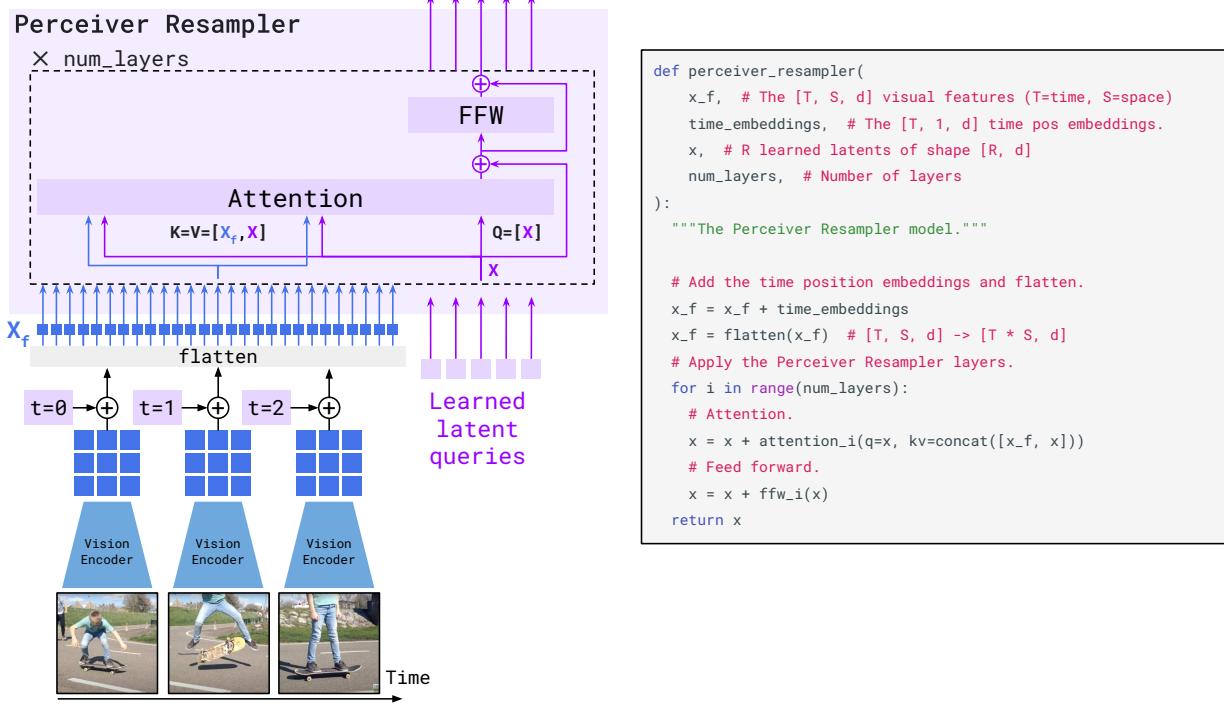


Figure 4 | **The Perceiver Resampler** module maps a *variable size* grid of spatio-temporal visual features coming out of the Vision Encoder to a *fixed number* of output tokens (five in the figure), independently of the input image resolution or the number of input video frames. This transformer has a set of learned latent vectors as queries, and the keys and values are a concatenation of the spatio-temporal visual features with the learned latent vectors. More details can be found in Section 3.1.1.

architecture in Figure 3. It takes as input a *variable number* of image or video features from the vision encoder and produces a *fixed number* of visual outputs as illustrated in Figure 4 (hence the name *Resampler*). The motivation for re-sampling the visual input to a fixed and small number (in practice 64) of outputs is to significantly reduce the computational complexity of vision-text cross attention, particularly important when dealing with multiple long videos. In similar spirit to Perceiver (Jaegle et al., 2021) and DETR (Carion et al., 2020), we learn a predefined number of latent input queries. These latent queries are fed to a transformer stack and cross attend to the flattened visual features X_f . These visual features are obtained by first adding a learnt temporal position encoding to each spatial grid of features corresponding to a given frame of the video (an image being considered as a single-frame video). Note that we only use temporal encodings and no spatial grid position encodings; we did not observe improvements from the latter, potentially because CNNs implicitly encode space information channel-wise (Islam et al., 2021). The visual features are then flattened and concatenated as illustrated in Figure 4. The number of output tokens of the Resampler is thus equal to the number of learnt latent queries. Unlike in DETR and Perceiver, the keys and values computed from the learnt latents are concatenated to the keys and values obtained from X_f , which we found to perform slightly better. We show later in the ablation studies (Section 4.4), that using such a vision-language resampler module outperforms a plain transformer and an MLP. More architectural details are provided in Table 13.

3.1.2. Conditioning a frozen language model on visual representations

As illustrated in Figure 5, text generation is performed by a Transformer decoder, conditioned on the visual representations X produced by the Perceiver Resampler. We build this model by interleaving

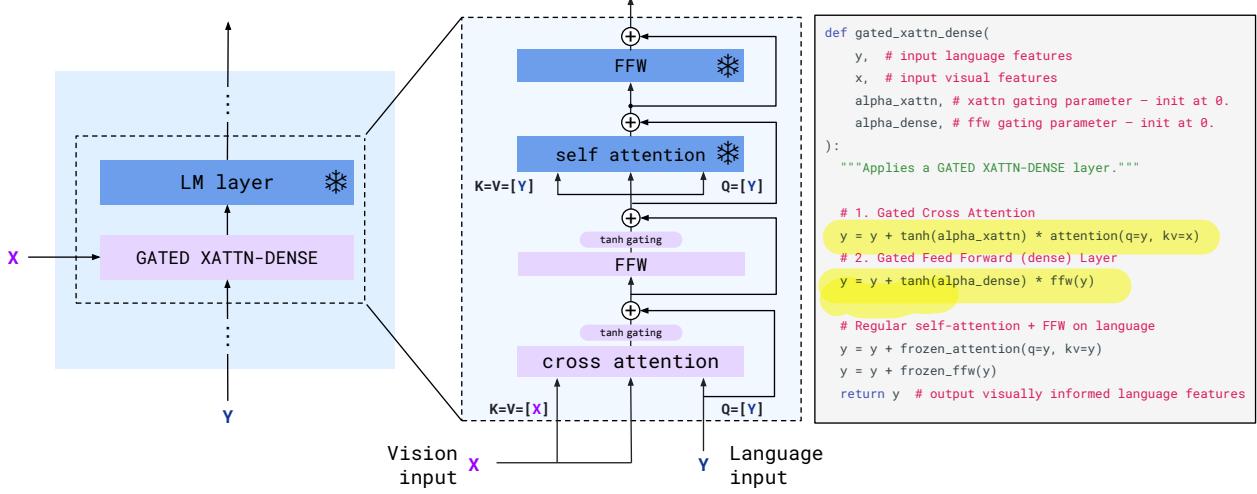


Figure 5 | **GATED XATTN-DENSE** layers. We insert new cross-attention layers, whose keys and values are obtained from the vision features while using language queries, followed by dense feed forward layers in between existing pretrained and frozen LM layers in order to condition the LM on visual inputs. These layers are *gated* so that the LM is kept intact at initialization for improved stability and performance.

pretrained blocks obtained from a text-only language model, and blocks trained from scratch that use the output of the Perceiver Resampler as one input. The pretrained text-only model is a decoder-only model trained on MassiveText (Rae et al., 2021), an English-only mixture of datasets obtained by scraping various Internet sources. Our largest Flamingo model relies on the 70B Chinchilla model trained by Hoffmann et al. (2022).

Interleaving new GATED XATTN-DENSE layers within a frozen pretrained LM. We wish to preserve the information contained in the text-only language model weights; we therefore freeze the pretrained blocks during training following Tsimpoukelli et al. (2021). In order to provide sufficient expressivity to the VLM and make it able to condition well on visual inputs, we insert *gated cross-attention dense* (GATED XATTN-DENSE illustrated in Figure 5) blocks in between the original layers, that are trained from scratch. Those blocks are made of a *cross-attention* layer, that attends the visual inputs with specific cross-attention masks (detailed in the next section), followed by an extra *dense feed-forward* (FFW) layer. As in GPT-2-style (Radford et al., 2019) attention layers, we apply layer normalization (Ba et al., 2016) to the keys, values, and queries input to the attention, as well as to the dense FFW inputs. The layer normalization layers include the learnt biases and offsets (with parameters shared between the keys and values). To ensure that at initialization, a forward pass through the conditioned model yields the same results as the original language model, we use a tanh-gating mechanism (Hochreiter and Schmidhuber, 1997). It consists in multiplying the output of a newly added layers by $\tanh(\alpha)$ right before adding it to the input representation from the residual connection, where α is a layer-specific learnable scalar initialized at 0 (similar to Bachlechner et al., 2021). Thus, at initialization, the branch going through the added layers is skipped, and the model output matches the pretrained language model one. During training, the model smoothly transitions from a fully trained text-only model to a visual language model (as displayed in Appendix C.4). This gating mechanism improves both the stability of training and the final performance (see ablations in Section 4.4). These layers are inserted at a certain depth frequency (see Table 1); this controls the ratio of fresh parameters over pretrained frozen parameters, which is key to regulate the model expressivity, memory usage and time complexity. Different frequencies of the added cross attention layers with the compute-performance trade-off is discussed in the ablation studies presented in

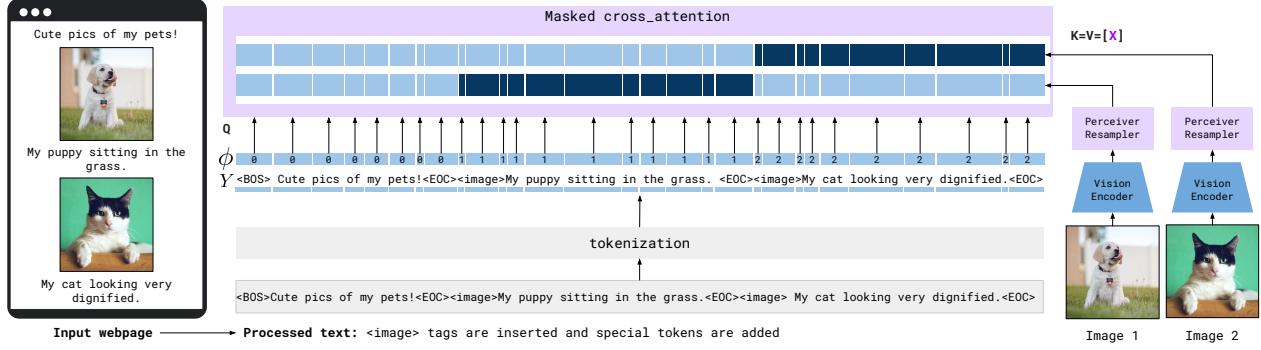


Figure 6 | Interleaved visual data and text support. Given text interleaved with images/videos, e.g. coming from a webpage, we first process the text by inserting `<image>` tags at the location of the visual data in the text as well as special tokens (`<BOS>` for “beginning of sentence” or `<EOC>` for “end of chunk”). The images are processed independently by the Vision Encoder and Perceiver Resampler to extract visual tokens. Following our modeling choice motivated in Section 3.1.3, each text token only cross-attends to the visual tokens corresponding to the last preceding image. The function ϕ illustrated above indicates for each token what is the index of the last preceding image (and 0 if there are no preceding images). In practice, this selective cross-attention is achieved via a masked cross attention mechanism – illustrated here with the dark blue entries (non masked) and light blue entries (masked).

Section 4.4. In these ablation studies, we also compare the proposed solution against other recent conditioning approaches (Desai and Johnson, 2021; Luo et al., 2022) and demonstrate that it offers a better trade off between added expressivity and retention of information acquired during pretraining.

3.1.3. Multi-visual input support: per-image/video attention masking

Interleaved sequence of visual data and text. We consider interleaved image/video and text examples: each example holds a sequence of text y , a sequence of images/videos x , and the sequence of positions of the images in the text. Based on the visual data positions, we define a function $\phi : [1, L] \mapsto [0, N]$ that assigns to each text position the index of the last image/video appearing before this position, (or 0 if no visual data appears before the position). The function ϕ defines which visual inputs we consider usable to predict token ℓ in Equation (1): the set of preceding tokens $y_{<\ell} \triangleq (y_1, \dots, y_{\ell-1})$, and the set of preceding images/videos $x_{\leq \ell} \triangleq \{x_i | i \leq \phi(\ell)\}$.

Multi-image attention. In practice, the multi-image attention is implemented within the GATED XATTN-DENSE layer by first imposing that all text tokens y cross-attend to the full concatenation of all visual tokens coming from the visual sequence x , followed by masking to respect the image-causal modelling introduced in Equation (1). The visual tokens are taken at the output of the Perceiver Resampler and the masking effectively limits the number of visual tokens that a certain text token sees. Typically, we allow each token to attend to the tokens of the image that appeared just before it in the interleaved sequence, i.e. the image x_i such that $\phi(\ell) = i$ as illustrated in Figure 6. We found this scheme to work better than allowing a text token to cross-attend to all previous images directly (as shown in Section 4.4). Although the model can only *directly* attend to a single image at any given point, there is still a causal dependency on all previous images in the sequence via causal self-attention in the text decoder.

An important advantage of this single-image cross-attention scheme is that it allows the model to seamlessly generalise to any number of images, regardless of how many are used during training. Indeed, the proposed attention scheme does not depend on the number of images that one cross



Figure 7 | **Training datasets.** Mixture of training datasets of different nature. N corresponds to the number of visual inputs for a single example. For paired image (or video) and text datasets, $N = 1$. T is the number of video frames with $T = 1$ being the special case of images. H, W, C are height, width and color channels.

attends to. In our case, at training time we use only up to 5 images per sequence when training on our interleaved datasets, but at evaluation our model is able to generalise to sequences of at least 32 “shots” (pairs of images and corresponding text). Moreover, for simplicity we currently perform the full cross-attention before masking the unused outputs, however a careful and more efficient implementation would only compute cross-attentions of the corresponding text tokens against a *single* image representation. Finally, we believe this scheme provides a useful inductive bias for traditional visual understanding tasks involving making predictions about a single image.

3.2. Training on a mixture of vision and language datasets

As illustrated in Figure 7, we train the Flamingo models on a mixture of three kinds of datasets: an interleaved image and text dataset derived from webpages (described in Section 3.2.1), image and text pairs, as well as video and text pairs (both described in Section 3.2.2). It is important to note that we only train on datasets collected from the web that were not annotated for machine learning purposes. In particular, we do not include any downstream task-specific datasets in the training mixture, in order to guarantee the generality of the approach.

3.2.1. Interleaved image and text dataset

Our model has the ability to handle interleaved text and image training data. To feed it such data, crucial for its few-shot capabilities, we collect the *MultiModal MassiveWeb (M3W)* dataset: we extract both text and images from the HTML of approximately 43 million webpages, determining the positions of images relative to the surrounding text based on the relative positions of the text and image elements in the Document Object Model (DOM). In total, *M3W* contains 185 million images (≈ 4.3 per page) and 182 GB of text (≈ 4.4 KB per page). An example processed webpage appears in Figure 7. We detail the collection of *M3W* in Appendix A.1.

We obtain a training example from a given webpage as follows (see Figure 6). We process the text by inserting `<image>` tags at the locations of the images in the webpage. These tags signal to the model the original locations within the text of the images on the page, and are added in plain text so that they do not require any additional special new tokens in the language model. Before any `<image>` tag, and at the end of the document (before `<EOS>`), we also add a special `<EOC>` (*end of chunk*) token. This token is added to the vocabulary of the underlying LM with its embedding randomly initialised and learnt, and is used during sampling and inference to denote the end of the text sequence prediction for a given image. From each document, we sample a random subsequence of $L = 256$ tokens and take up to $N = 5$ images included in the sampled sequence (using only the first N within that sampled subsequence if there are more, or padding to N if fewer). Based on the position of the images with respect to the text, we deduce the function ϕ (a length L sequence), which

indicates for each token the index of the last preceding image. This sequence is used to compute the cross-attention mask described in Section 3.1.3. We provide further details about this processing in Appendix A.1.2. In particular, we explain how we can perform a form of “data augmentation” when constructing ϕ .

In summary, the output of these preprocessing steps is an instance consisting of three parts:

1. `images`: floats, shape $[N = 5, T = 1, H = 320, W = 320, C = 3]$
2. `text`: integers (tokens), shape $[L = 256]$
3. `indices`: $(\phi(i))_{1 \leq i \leq L}$, integers in $[0, N]$, shape $[L = 256]$

3.2.2. Visual data paired with text

Along with our interleaved image and text dataset, we use several paired vision and text web datasets for training. One dataset is ALIGN (Jia et al., 2021), composed of 1.8 billion images paired with alt-text. ALIGN is large, but noisy and limited to images. The images are often poorly described by the corresponding alt-text annotation. For this reason, we augment it with two datasets: LTIP (Long Text & Image Pairs) consists of 312 million images, and VTP (Video & Text Pairs) consists of 27 million short videos (approximately 22 seconds on average). Both datasets are paired with more descriptive captions. For instance, the average number of tokens of an ALIGN text description is 12.4 per image, while it is 20.5 for the LTIP dataset. The LTIP and VTP datasets were collected by crawling fewer than ten websites targeting high-quality and rich image descriptions. These single-image and single-video datasets are preprocessed analogously to the M3W data preprocessing described previously, adding the `<image>` tag at the beginning of the sequence (immediately after `<BOS>`), and the `<EOC>` token after the text (before `<EOS>`). We deduplicated these datasets against all our benchmarks (against both the training and the evaluation sets) using image similarity, as detailed in Appendix A.3. Datasheets for LTIP and VTP are respectively given in Appendix A.2.1 and Appendix A.2.2. A preprocessed instance from one of these datasets again consists of three parts:

1. `images`: floats, shape $[N = 1, T, H = 320, W = 320, C = 3]$ ($T = 1$ for images; $T = 8$ for videos)
2. `text`: integers (tokens), shape $[L]$ (L is dataset-dependent $L = 32$ or $L = 64$)
3. `indices`: integers 1, shape $[L]$ (trivial for these datasets that have a single visual input $N = 1$)

3.2.3. Training objective and optimisation strategy

We train our models by minimizing a weighted sum of dataset specific expected negative log likelihood of text given some visual inputs:

$$\sum_{m=1}^M \lambda_m \cdot \mathbb{E}_{(x,y) \sim \mathcal{D}_m} \left[- \sum_{\ell=1}^L \log p(y_\ell | y_{<\ell}, x_{\leq \ell}) \right], \quad (2)$$

where \mathcal{D}_m and λ_m is the m -th dataset and the positive scalar weighting its influence in the loss, respectively. As the different datasets have different properties (e.g., quality of text, level of correspondence between text and visuals or image versus videos), we found that tuning these weights was very important to overall performance (similar to what was found in the Gopher paper (Rae et al., 2021)). In practice, at each step of optimisation we go over each dataset \mathcal{D}_m in turn, sample a batch of size B_m of visual language sequences from it, compute the gradient of the loss with respect to the minibatch and weight it by λ_m . We then accumulate the gradients over all M datasets before triggering an update step. We found this gradient accumulation strategy to be crucial for high performance compared to a round-robin approach (Cho et al., 2021) as shown in the ablation study (Section 4.4.1).

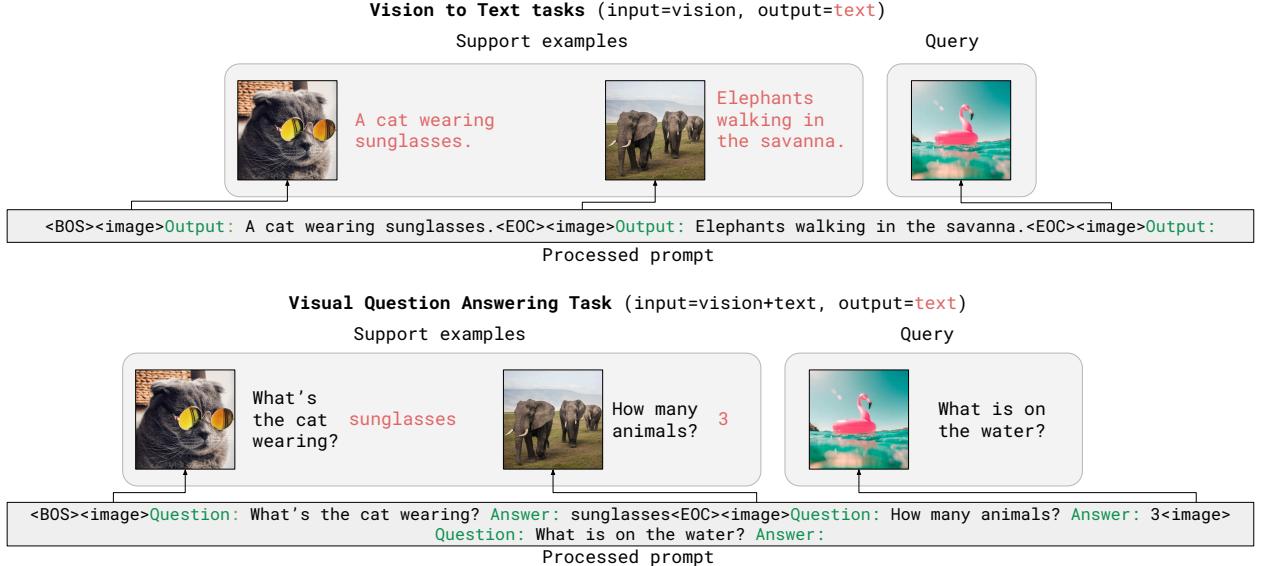


Figure 8 | **Few-shot interleaved prompt generation.** Given some task-specific few-shot examples (a.k.a. support examples) and a query for which Flamingo models have to make a prediction, we build the prompt by interleaving the image before each corresponding text. We introduce some formatting to do this, e.g. we prepend "**Output:**" to the expected response for all vision to text tasks or use a formatting prompt "**Question:** {question} **Answer:** {answer}" for visual question answering tasks.

3.3. Task adaptation with few-shot in-context learning

Once Flamingo is trained, we use it to tackle a visual task by conditioning it on an interleaved prompt, as illustrated in Figure 8.

In-context learning with Flamingo models. We evaluate the ability of our models to rapidly adapt to new tasks using in-context learning, following an analogous approach to the one used in GPT-3 (Brown et al., 2020). In detail, we are given a set of support examples in the form of $(image, text)$ or $(video, text)$ (where the *image* or *video* is the input visual and the *text* is the expected response and/or also contains additional task information e.g. a question) and a single visual query for which we want our model to make a prediction. Given this, we build a multimodal prompt by concatenating the support examples followed by the visual query as illustrated by Figure 8. Unless specified otherwise, we choose the concatenation order at random.

Open-ended and close-ended evaluations. In an open-ended setting, the model’s sampled text following the query image is then taken as its prediction for the image, stopping at the first $\langle EOC \rangle$ (“end of chunk”) token prediction. Unless specified otherwise, we always use beam search with a beam size of 3. In a close-ended setting, all possible outputs are independently appended to the query image, and we score each of the resulting sequences using the log-likelihood estimated by our model. These scores are then used to rank the candidate outputs in decreasing order, from most confident to least confident.

Zero-shot generalization. In the absence of few-shot examples, approaches commonly rely on prompt engineering (Radford et al., 2021) to condition the model at inference using a suitable natural language description of the task. Validation of such prompts can significantly impact performance but

requires access to a number of annotated examples and cannot therefore be considered truly zero-shot. Furthermore, Perez et al. (2021) have shown that such validation procedures are generally not robust with access to only a handful of samples during validation. To report zero-shot performance in our work, we instead build a prompt with two examples from the downstream tasks where we remove their corresponding images or videos. For example, for the task illustrated at the top of Figure 8, the prompt would be "`<BOS>Output: This is a cat wearing sunglasses.<EOC>Output: Three elephants walking in the savanna.<EOC><image>` Output:" and no support images would be fed to the model. We observed that only showing one, instead of two, text examples in the prompt is highly detrimental as the model is biased towards producing text output similar to the single provided text example. Providing more than two text examples helps but only marginally. We hence stick with two text examples for practical reasons. In practice, we believe this is not more cumbersome than finding a good natural text description for a given task. This relates to recent findings on the aspects of demonstrations that are key drivers of performance (Min et al., 2022). For close-ended tasks, where we use the model to score different possible answers, we observe it is not necessary to provide a single text example in the zero-shot prompt.

Retrieval-based In-Context Example Selection (Yang et al., 2021). When the size of the support set exceeds a certain limit, it can become difficult to leverage all the examples with in-context learning: first because it becomes excessively expensive to fit all the examples in the prompt, and second there is a risk of poor generalization when the prompt size exceeds the size of the sequence used during training (Press et al., 2022). In such situations, it is appealing to use a form of prompt selection to both limit the sequence length as well as potentially improve the prompt quality which can in turn lead to better performance (Liu et al., 2021). In particular, we follow the Retrieval-based In-Context Example Selection (RICES in short) approach introduced by Yang et al. (2021). In details, given a query image we retrieve similar images in the support set by comparing the visual features extracted from our frozen pretrained visual encoder. We then build the prompt by concatenating the top- N most similar examples. Since LM are sensitive to the ordering in the prompt due to recency bias (Zhao et al., 2021), we order the examples so that the most similar support example appears right before the query. We notably show the effectiveness of this approach in classification settings with multiple hundreds of classes (see Section 4.2.2) where we are given one or more images/videos per class, yielding a number of examples that would not otherwise all fit in the prompt.

Prompt ensembling. We also explore ensembling the outputs of the model across multiple prompts. This can notably be combined with RICES where ensembling can be done over multiple permutations of the ranked nearest neighbors. Specifically, for a given answer, we ensemble the log likelihoods estimated by the model over 6 random permutations of the selected few-shot prompts.

	Requires model sharding	Frozen		Trainable		Total count
		Language	Vision	GATED XATTN-DENSE	Resampler	
Flamingo-3B	✗	1.4B	435M	1.2B (every)	194M	3.2B
Flamingo-9B	✗	7.1B	435M	1.6B (every 4th)	194M	9.3B
Flamingo	✓	70B	435M	10B (every 7th)	194M	80B

Table 1 | **Parameter counts for Flamingo models.** We focus on increasing the parameter count of the frozen LM and the trainable vision-text GATED XATTN-DENSE modules while maintaining the frozen vision encoder and trainable Resampler to a fixed and small size across the different models. The frequency of the GATED XATTN-DENSE with respect to the original language model blocks is given in parenthesis.

4. Experiments

We first introduce our experiment setting in Section 4.1. We then report the results of the Flamingo models on few-shot learning in Section 4.2. In Section 4.3, we provide the results of the Flamingo models in the fine-tuning regime. Finally we provide an ablation study validating our design choices in Section 4.4.

4.1. Training and evaluation setting

4.1.1. Models

We perform experiments across three model sizes, where we scale the frozen language model from 1.4B to 7B and 70B; and adapt the parameter count of other components accordingly. We keep the pretrained vision encoder frozen across all experiments and use a NFNet-F6 model trained contrastively (see Section 4.1.4), unless explicitly stated otherwise in the ablation study. We use a Perceiver Resampler with approximately 200M parameters across all three model sizes.

The decision on how many GATED XATTN-DENSE layers to interleave is mainly driven by a trade-off between memory constraints and downstream performance. We identified the optimal trade-off at small model scales, before transferring our findings to the large model architecture.

We obtain three models, *Flamingo-3B*, *Flamingo-9B* and *Flamingo-80B*, detailed below:

- The *Flamingo-3B* model builds on top of a **1.4B frozen language model** from Hoffmann et al. (2022). Before each transformer block, we add a GATED XATTN-DENSE layer attending to the visual inputs; this accounts for 1.4B additional learned parameters.
- The *Flamingo-9B* model builds on top of a **7B frozen language model** from Hoffmann et al. (2022). Starting from the very first layer and before every fourth transformer blocks, we add a GATED XATTN-DENSE layer attending to the visual inputs; this accounts for 1.8B additional learned parameters.
- The *Flamingo-80B* model builds on top of the **frozen Chinchilla 70B language model** (Hoffmann et al., 2022). Starting from the very first layer and before every seventh transformer blocks, we add a GATED XATTN-DENSE layer attending to the visual inputs; this accounts for 10B additional learned parameters. For simplicity, we refer to this model as simply *Flamingo* throughout the paper.

We report the parameter count of each component of our models, as well as model sharding requirements, in Table 1 and provide more Transformer architecture details in Appendix C.1. The *Flamingo* model card (Mitchell et al., 2019) is also given in Appendix B.

Dataset	DEV	Gen.	Custom prompt	Task description	Eval set	Metric
ImageNet-1k [103]	✓			Object classification	Val	Top-1 acc.
MS-COCO [18]	✓	✓		Scene description	Test	CIDEr
VQAv2 [3]	✓	✓		Scene understanding QA	Test-dev	VQA acc. [3]
OKVQA [75]	✓	✓		External knowledge QA	Val	VQA acc. [3]
Flickr30k [149]		✓		Scene description	Test (Karpathy)	CIDEr
VizWiz [40]		✓		Scene understanding QA	Test-dev	VQA acc. [3]
TextVQA [108]		✓		Text reading QA	Val	VQA acc. [3]
VisDial [22]				Visual Dialogue	Val	NDCG
HatefulMemes [60]			✓	Meme classification	Seen Test	ROC AUC
Video	Kinetics700 2020 [110]	✓		Action classification	Val	Top-1/5 avg
	VATEX [132]	✓	✓	Event description	Test	CIDEr
	MSVDQA [140]	✓	✓	Event understanding QA	Test	Top-1 acc.
	YouCook2 [161]	✓		Event description	Val	CIDEr
	MSRVTTQA [140]	✓		Event understanding QA	Test	Top-1 acc.
	iVQA [145]		✓	Event understanding QA	Test	iVQA acc. [145]
	RareAct [81]			Composite action retrieval	Test	mWAP
	NextQA [139]		✓	Temporal/Causal QA	Test	WUPS
	STAR [138]			Multiple-choice QA	Test	Top-1 acc.

Table 2 | **Summary of the evaluation benchmarks.** **DEV** benchmarks were used to validate general design decision of the Flamingo models. Gen. stands for generative task where we sample text from the VLM. If a task is non-generative it means that we use VLM to score answers among a given finite set. For most of our tasks we use a common default prompt, hence minimizing task-specific tuning (see Section 4.1.3).

4.1.2. Evaluation benchmarks

Our goal is to develop models that can rapidly adapt to diverse and challenging tasks in the few-shot setting. For this, we consider a wide array of popular image and video benchmarks summarized in Table 2. In total we chose 16 multimodal image/video and language benchmarks spanning tasks that requires some language understanding (visual question answering, captioning, visual dialogue) as well as two standard image and video classification benchmarks (ImageNet and Kinetics). Note that for the video datasets collected from YouTube (i.e. all video datasets except NextQA and STAR), we evaluated our model on all the publicly available video as of April 2022.

DEV benchmarks. In order to validate design decisions of our model during the course of the project (see Section 4.4), we selected seven benchmarks as our development set (referred as **DEV**). To maximise its relevance, we choose the most challenging and widely studied benchmarks for captioning, visual question answering and classification tasks on both images and videos.

Dataset splits for the **DEV benchmarks.** Concretely, estimating few-shot learning performance of a model consists in adapting it on a set of *support* samples and evaluating it on a set of *query* samples. As a result, any evaluation set should be composed of two disjoint subsets containing respectively the support and the query samples. For the **DEV** benchmarks that are used both to validate design decisions and hyperparameters, as well as to report final performance, we therefore use four subsets:

- *validation support*: contains support samples for validation;
- *validation query*: contains query samples for validation;
- *test support*: contains support samples for final performance estimation;
- *test query*: contains query samples for final performance estimation.

In practice, for the *test query* subset, we use the subset that prior works report results on, for apple-to-apple comparison. While the validation set would be a natural choice for the *validation query* subset, we note that this is not possible for all benchmarks, since some benchmarks do not have an official validation set (e.g. OKVQA) and for others, the validation is commonly used to report final performance in place of the test set (e.g. ImageNet or COCO). For simplicity, we use a subset of the original train set as the *validation query* subset. Finally, we also use additional disjoint subsets of the train set as respectively the *validation support* subset and the *test support* subset. We describe in Appendix C.3 how many samples we use for each subset.

Unbiased few-shot performance estimation. Few-shot learning performance estimates on the DEV benchmarks may be biased; in the sense that along the progress of this work, design decision were made based on the performance obtained on these benchmarks. We note that this is the case for prior work which also make use of these benchmarks to validate and ablate their own design decisions. To account for this bias and provide unbiased few-shot learning performance estimates, we report performance on a remaining set of 11 benchmarks. Among those, some spans the same open-ended image and video tasks as our DEV benchmarks (captioning and video question answering). But we also look at more specific benchmarks in order to explore less explored capabilities. This notably includes: TextVQA (Singh et al., 2019) that specifically assesses OCR capability through question answering, VisDial (Das et al., 2017) which is the only visual dialogue benchmark, HatefulMemes (Kiela et al., 2020) which is the only vision and text classification benchmark, NextQA (Xiao et al., 2021) which specially focuses on causality and temporal relation, STAR (Wu et al., 2021) which is the only multiple-choice question answering task and finally RareAct (Miech et al., 2020), the only benchmark measuring compositionality in action recognition. We emphasize that *we do not proceed to any validation of design decisions* on these benchmarks and use them solely to estimate unbiased few-shot learning performance after Flamingo training is done.

4.1.3. Few-shot learning evaluation hyperparameters

In few-shot learning, hyperparameter selection implicitly increases the number of shots as it requires additional validation examples. If those are not taken into account, as is often the case in practice, few-shot performance can be overestimated (Perez et al., 2021). Similarly, cross-validation of benchmark-specific hyperparameters such as the prompt should be considered as a particularly basic few-shot learning method, where one selects the task-specific prompt over the set of shots; but other learning approaches might be more effective in making use of these labelled examples. From this perspective, as pointed out by Perez et al. (2021), the results reported for methods such as CLIP (Radford et al., 2021) or ALIGN (Jia et al., 2021) cannot be considered as true zero-shot. In fact, it is unclear how many “shots” were used or would have been needed to obtain the important gains reported when using prompt engineering. Given the negative results reported by Perez et al. (2021) in terms of the robustness of cross-validation and unless mentioned otherwise, all benchmarks are run using a single set of evaluation hyperparameters, including the prompts. Thus our few-shot learning performance is always “true” in the sense described by Perez et al. (2021). We optimize hyperparameters jointly across the *validation* subsets of the DEV benchmarks and do not proceed to any benchmark-specific cross-validation of hyperparameters. In particular (except for HatefulMemes and RareAct), we always use the prompt "`Output: {output}`" for all non question answering tasks and "`Question: {question} Answer: {answer}`" for all question answering / visual dialogue tasks. In particular for VisDial (Das et al., 2017), we use the previously described prompt to encode each questions/answers of the dialogue and the provided image caption is prepended to the dialogue history without any prompt. For HatefulMemes (Kiela et al., 2020), we use a specific prompt to incorporate the OCR information provided as input which is: "`is an image with written: "{meme_text}"`"

on it. Is it hateful? Answer: {answer}", where the answer is either yes or no. Note that this is the only dataset where we explicitly provide OCR text `meme_text` as input to Flamingo models. In particular, for TextVQA, we do not make use of the provided OCR transcripts and instead directly rely on the off-the-shelf OCR ability of the Flamingo models. For RareAct which is a zero-shot benchmark, we change the verb names to the third person, add an article before each noun and use the prompt "**Caption: a person {verb + object}"**.

4.1.4. Training details for the Flamingo models

Data augmentation and preprocessing. Empirically we find that it is effective to stochastically prepend the paired dataset text samples with a single space character, with probability 0.5. We attribute this to the fact that our subword tokenizer maps the beginning of various words to a different token depending on whether it is preceded with a space or not. This allows us to enforce invariance to this tokenizer artifact, without degrading significantly correctness of the punctuation which is already lacking in many of these samples and leads to substantial improvement across tasks.

The visual inputs are resized to 320×320 while preserving their aspect ratios, padding the image with the mean value if required. Note that this is higher than the 288×288 resolution used for the contrastive pretraining of our Vision Encoder (Appendix C.2.1). The increase in resolution during the final stage training was motivated by [Touvron et al. \(2019\)](#) that show one can obtain improved performance at a higher test-time resolution when using CNNs. This increase in resolution also comes with a moderated computational and memory cost as no backpropagation is performed on the frozen Vision Encoder. We also employ random left/right flip and color augmentation.

For interleaved datasets (Section 3.2.1) we also employ augmentation by lightly randomizing the selected image indices ϕ with a hyperparameter p_{next} when sampling examples from the *M3W* dataset. This augmentation is detailed in Appendix A.1.2 and our choice of $p_{next} = \frac{1}{2}$ is ablated in Section 4.4. For video training, we temporally sample a clip of 8 frames sampled at one frame per second (fps) from each training video. Although our model was trained with a fixed number of 8 frames, at inference time, we input our model with 30 frames at 3 fps. This is achieved by linearly interpolating the learnt time position embedding at test time.

Loss and optimisation. All our models are trained using the AdamW optimizer with global norm clipping of 1, no weight decay to the Perceiver Resampler and a weight decay of 0.1 to the other trainable parameters. The learning rate is increased linearly from 0 to 10^{-4} up over the first 5000 steps then held constant for the duration of training (no improvements were observed from decaying the learning rate). Unless specified otherwise we train our models for 500k steps. Four datasets are used for training: *M3W*, *ALIGN*, *LTIP* and *VTP* with weights λ_m of 1.0, 0.2, 0.2 and 0.03 respectively. Batch sizes depend on the setting and are given in the next sections.

Infrastructure and implementation. Our model and associated infrastructure were implemented using JAX ([Bradbury et al., 2018](#)) and Haiku ([Hennigan et al., 2020](#)). All training and evaluation was performed on TPUs v4 instances. The largest model containing 80 billion parameters is trained on 1536 chips for 15 days and sharded across 16 devices. Megatron type sharding ([Shoeybi et al., 2019](#)) is used to enable 16-way model parallelism for all Embedding / Self-Attention / Cross-Attention / FFW layers, while the NFNet vision layers were unsharded. ZeRO stage 1 ([Rajbhandari et al., 2020](#)) is used to shard the optimizer state. All trained parameters and optimizer accumulators are stored and updated in `float32`; all activations and gradients are computed in `bfloat16` after downcasting of parameters from `float32` to `bfloat16`. Frozen parameters are stored and applied in `bfloat16`.

More because it's QA?
Can few-shot learning?

Method	FT	Shot	OKVQA	VQAv2	COCO	MSVDQA	VATEX	VizWiz	Flick30K	MSRVTTQA	IVQA	YouCook2	STAR	VisDial	TextVQA	NextQA	HatefulMemes	RareAct
Zero/Few shot SOTA	<input checked="" type="checkbox"/>	(X)	[39]	[124]	[134]	[64]	-	-	-	[64]	[145]	[153]	[87]	-	-	[94]	[94]	
		(X)	(16)	(4)	(0)	(0)				(0)	(0)	(0)	(0)			66.1	40.7	
<i>Flamingo</i> -3B	<input checked="" type="checkbox"/>	0	41.2	49.2	73.0	27.5	40.1	28.9	60.6	11.0	32.7	55.8	39.6	46.1	30.1	21.3	53.7	58.4
	<input checked="" type="checkbox"/>	4	43.3	53.2	85.0	33.0	50.0	34.0	72.0	14.9	35.7	64.6	41.3	47.3	32.7	22.4	53.6	-
	<input checked="" type="checkbox"/>	8	44.6	55.4	90.6	37.0	54.5	38.4	71.7	19.6	36.8	68.0	40.6	47.6	32.4	23.9	54.7	-
	<input checked="" type="checkbox"/>	16	45.6	56.7	95.4	40.2	57.1	43.3	73.4	23.4	37.4	73.2	40.1	47.5	31.8	25.2	55.3	-
	<input checked="" type="checkbox"/>	32	45.9	57.1	99.0	42.6	59.2	45.5	71.2	25.6	37.7	76.7	41.6	OOC	30.6	26.1	56.3	-
<i>Flamingo</i> -9B	<input checked="" type="checkbox"/>	0	44.7	51.8	79.4	30.2	39.5	28.8	61.5	13.7	35.2	55.0	41.8	48.0	31.8	23.0	57.0	57.9
	<input checked="" type="checkbox"/>	4	49.3	56.3	93.1	36.2	51.7	34.9	72.6	18.2	37.7	70.8	42.8	50.4	33.6	24.7	62.7	-
	<input checked="" type="checkbox"/>	8	50.0	58.0	99.0	40.8	55.2	39.4	73.4	23.9	40.0	75.0	43.4	51.2	33.6	25.8	63.9	-
	<input checked="" type="checkbox"/>	16	50.8	59.4	102.2	44.5	58.5	43.0	72.7	27.6	41.5	77.2	42.4	51.3	33.5	27.6	64.5	-
	<input checked="" type="checkbox"/>	32	51.0	60.4	106.3	47.2	57.4	44.0	72.8	29.4	40.7	77.3	41.2	OOC	32.6	28.4	63.5	-
<i>Flamingo</i>	<input checked="" type="checkbox"/>	0	50.6	56.3	84.3	35.6	46.7	31.6	67.2	17.4	40.7	60.1	39.7	52.0	35.0	26.7	46.4	60.8
	<input checked="" type="checkbox"/>	4	57.4	63.1	103.2	41.7	56.0	39.6	75.1	23.9	44.1	74.5	42.4	55.6	36.5	30.8	68.6	-
	<input checked="" type="checkbox"/>	8	57.5	65.6	108.8	45.5	60.6	44.8	78.2	27.6	44.8	80.7	42.3	56.4	37.3	32.3	70.0	-
	<input checked="" type="checkbox"/>	16	57.8	66.8	110.5	48.4	62.8	48.4	78.9	30.0	45.2	84.2	41.1	56.8	37.6	32.9	70.0	-
	<input checked="" type="checkbox"/>	32	57.8	67.6	113.8	52.3	65.1	49.8	75.4	31.0	45.3	86.8	42.2	OOC	37.9	33.5	70.0	-
Pretrained FT SOTA	<input checked="" type="checkbox"/>	(X)	54.4 (10K)	80.2 (444K)	143.3 (500K)	47.9 (27K)	76.3 (500K)	57.2 (20K)	67.4 (30K)	46.8 (130K)	35.4 (6K)	138.7 (10K)	36.7 (46K)	75.2 (123K)	54.7 (20K)	25.2 (38K)	75.4 (9K)	-

Table 3 | **Comparison to the state of the art on multimodal benchmarks.** A *single* Flamingo model reaches state-of-the-art on a wide array of image and video tasks with in-context learning from as few as 4 examples per task, beating previous zero-shot or few-shot method by a large margin. More importantly, using only 32 examples and without adapting any model weight, Flamingo *outperforms* the current best methods on 7 tasks, that are fine-tuned on thousands of annotated examples. Best few-shot numbers are in **bold**. Best numbers overall are underlined. See also Figure 2 that illustrate the table. OOC: out-of-context, which happens when the few-shot prompt is longer than the maximum sequence length the model has been trained on.

4.2. Few-shot learning with Flamingo models

This section explores the evaluation of the *Flamingo*-3B, *Flamingo*-9B and *Flamingo* models. In Section 4.2.1 we describe the main results of the paper, i.e. how a *single* Flamingo model can be quickly adapted to a wide array of image / video and language tasks with a handful of annotated examples via in-context learning. Section 4.2.2 explores how the Flamingo models, despite their open-ended nature, can also be made competitive for use in close-ended classification tasks. Finally Section 4.2.3 provides the results of our in-house contrastive pretraining.

4.2.1. State-of-the-art few-shot learning on vision-language tasks

Main results. Results are given in Table 3. Overall, *Flamingo* outperforms by a large margin *all* best previous zero-shot or few-shot methods on the 16 considered benchmarks. This is achieved with as few as four examples per task, hence opening up a practical and high-performing way to adapt vision models to new tasks. More importantly, *Flamingo* is often competitive compared to the state of the art methods which are additionally fine-tuned on up to hundreds of thousands annotated examples. On six tasks, *Flamingo* actually surpasses the best individual fine-tuned SotA despite using a *single* set of weights and only up to 32 task specific examples, orders of magnitude less than the fine-tuned approaches. Note that in Table 3, we reported the SotA methods published in publicly available papers and discarded winning entries of challenges such as TextVQA, VisDial or HatefulMemes, often not published or using ensembles of numerous approaches. Finally, despite the fact we have used the DEV benchmarks to do design decisions, our results generalize well to the other benchmarks, confirming the generality of our approach.

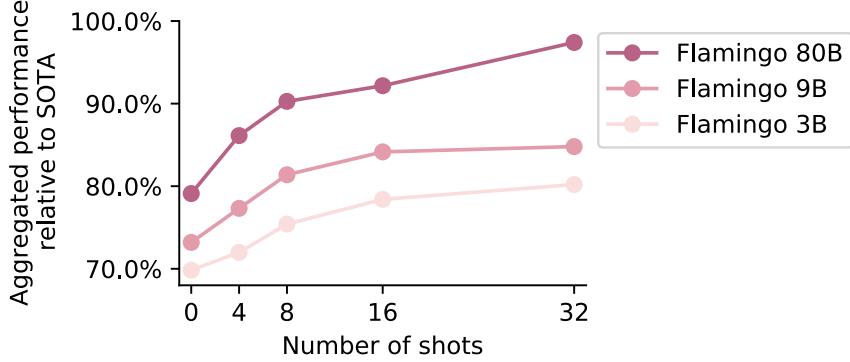


Figure 9 | Overall impact of model scaling and number of shots. The performance of Flamingo models increases with their parametric size and with the number of in-context shot. Performance is reported by averaging the SotA relative score across the 16 benchmarks from Section 4.2.1.

Scaling with respect to parameters and shots. A general trend we observe is that, similarly to what was observed in GPT-3 (Brown et al., 2020), the larger the model, the better the few-shot performance. We also observe that the overall performance improves with the number of shots. This is illustrated in Figure 9, that plots the average, on the 16 tasks, of the relative performance of each model with respect to the state-of-the-art performance of fine-tuned models. We further find that the largest model better exploits the increase in the number of shots. Interestingly, even though our Flamingo models were trained with sequences limited to only 5 images on *M3W*, they are still able to benefit from 8, 16 and even 32 images or videos at inference. This demonstrates the flexibility of our proposed Flamingo architecture for processing a variable number of videos or images. Finding how to benefit more from more shots with in-context learning is a promising future direction.

4.2.2. Few-shot learning on classification tasks

In this section we consider applying the Flamingo models to well studied classification benchmarks like ImageNet or Kinetics700. Results are given in Table 4. We observe the similar pattern as in the previous section: larger model tend to perform better. Second, given that few-shot classification tasks often come up with more examples to leverage (1000 for ImageNet), using methods to scale to larger support sets is beneficial. RICES (Retrieval In-Context Example Selection (Yang et al., 2021) described in Section 3.3) is much better than simply selecting the examples randomly in the prompt. Indeed, *Flamingo* can get a 9.2% improvement on ImageNet when selecting 16 support examples out of 5000 using RICES compare to choosing the same elements randomly. Ensembling multiple prompts further boost results. However we also note that Flamingo models are still below the current dominant contrastive paradigm for such close-ended tasks. In particular, Flamingo models underperform the same contrastive model they use for the vision encoder (see Limitation Section 6.1 for more details). Finally, state-of-the-art zero-shot models on ImageNet such as BASIC (Pham et al., 2021) and LiT (Zhai et al., 2021) are particularly optimized on classification tasks as trained on JFT-3B (Zhai et al., 2021), which is a dataset with image and labels. A future direction could be to improve the performance of VLMs such as the Flamingo models for such classification tasks.

4.2.3. Zero-shot performance of the pretrained contrastive model

A crucial part of our approach is the Vision Encoder, pretrained separately using contrastive learning and kept frozen when training Flamingo models. We report zero-shot image classification results on ImageNet, Kinetics700 and retrieval results results Flickr30K and COCO. The classification results

Model	Method	Prompt size	shots/class	ImageNet top 1	Kinetics700 avg top1/5
SotA	Fine-tuned	-	full	91.0 [137]	89.0 [144]
SotA	Contrastive	-	0	85.7 [90]	69.6 [94]
NFNetF6	Our contrastive	-	0	77.9	62.9
<i>Flamingo</i> -3B	RICES	8	1	70.9	55.9
		16	1	71.0	56.9
		16	5	72.7	58.3
<i>Flamingo</i> -9B	RICES	8	1	71.2	58.0
		16	1	71.7	59.4
		16	5	75.2	60.9
Random		16	≤ 0.02	66.4	51.2
<i>Flamingo</i> -80B	RICES	8	1	71.9	60.4
		16	1	71.7	62.7
		16	5	76.0	63.5
RICES+ensembling		16	5	77.3	64.2

Table 4 | **Few-shot results on classification tasks.** The Flamingo models can also be used for standard classification tasks. In particular, we explore having access to support sets bigger than what our current prompt can accommodate (using up to 5000 support examples). In that regime, large gains are obtained by using the RICES method (Yang et al., 2021) as well as prompt ensembling. We also observe the same trend as with the vision-language benchmarks: bigger models do better and more shots help.

	Flickr30K						COCO					
	image-to-text			text-to-image			image-to-text			text-to-image		
	R@1	R@5	R@10									
Florence [150]	90.9	99.1	-	76.7	93.6	-	64.7	85.9	-	47.2	71.4	-
ALIGN [56]	88.6	98.7	99.7	75.7	93.8	96.8	58.6	83.0	89.7	45.6	69.8	78.6
CLIP [94]	88.0	98.7	99.4	68.7	90.6	95.2	58.4	81.5	88.1	37.7	62.4	72.2
Ours	89.3	98.8	99.7	79.5	95.3	97.9	65.9	87.3	92.9	48.0	73.3	82.1

Table 5 | **Zero-shot contrastive pretraining evaluation.** Zero-shot image-text retrieval evaluation of our pretrained contrastive model compared to the state-of-the-art dual encoder contrastive models.

are presented in Table 4 while the retrieval results are given in Table 5. For the retrieval tasks, our model outperforms the current state-of-the-art contrastive dual encoder approaches CLIP, ALIGN and Florence. However, we underperform the zero-shot state-of-the-art on Kinetics700 (Radford et al., 2021, CLIP) and the zero-shot state-of-the-art on ImageNet (Pham et al., 2021, BASIC). However, as noted earlier, BASIC is particularly optimized for classification. It is trained on the JFT-3B (Zhai et al., 2021) dataset which has images with labels rather than captions. We have noticed training on image and short text descriptions similar to labels significantly help for ImageNet but is detrimental for retrieval benchmarks which require to capture rich scene text description instead. Since our goal is to use the Vision Encoder as a feature extractor for the Flamingo models in order to provide capture the whole scene and not just the main object, we favor retrieval metrics over classification ones. We provide more details about the contrastive pretraining including model ablations in Appendix C.2.1.

4.3. Fine-tuning *Flamingo* as a pretrained vision-language model

While not the main focus of our work, it is important to verify how the Flamingo models can be adapted to a task when given more data. For this reason we explore fine-tuning our largest model, *Flamingo*, for a given target task given no limit on the annotation budget. To fine-tune *Flamingo* on a downstream task, we train it on data batches from the downstream task of interest in the same format as the single-image/video datasets described in Section 3.2.

*Can we finetune for
our usecases? @Seon*

Freezing and hyperparameters. When fine-tuning *Flamingo*, we keep the underlying LM layers frozen and train the same Flamingo layers as during pretraining. We also increase the resolution of the input images from 320×320 to 480×480 . Unlike in the pretraining phase, we also fine-tune the base visual encoder, finding that this typically improves results, likely due in part to the higher input resolution.

We choose certain hyperparameters on a per-task basis by grid search on a validation subset of the training set (or on the official or standard validation set where available). These hyperparameters include the learning rate (ranging from 3×10^{-8} to 1×10^{-5}) and decay schedule (exponential decay by factors of $10\times$), number of training steps, batch size (either 8 or 16), and whether visual data augmentation (color augmentation, random horizontal flips) is used.

Results. In Table 6, we present our results for per-task *Flamingo* fine-tuning. When provided access to a large-scale task-specific dataset with many thousands of examples, we find that we can indeed improve results over our previously presented in-context few-shot learning results, setting a new state of the art on five tasks: VQAv2, VATEX, VizWiz, MSRVTTQA, and HatefulMemes. For example, on VQAv2, we observe improved results at 82.0%, outperforming our results achieved with 32-shot in-context learning (67.3%) as well as the previous state of the art (81.3%, Yan et al. (2021)).

Although these fine-tuning results come at high computational cost relative to the previously presented in-context few-shot learning results – among other challenges like hyperparameter tuning – they further demonstrate the power of VLM pretraining for visual understanding even in the presence of large amounts of task-specific training data.

In some cases our results likely trail the state of the art due in part to the fact that we simply optimise log-likelihood and do not make use of common task-specific metric optimisation tricks, such as CIDEr optimisation (Liu et al., 2017; Rennie et al., 2017) for COCO captioning, and fine-tuning on dense annotations for VisDial (Murahari et al., 2020). For example, Murahari et al. (2020) report a 10% relative improvement in NDCG on VisDial from such dense annotation fine-tuning.

4.4. Ablation studies

We perform all ablations using the *Flamingo*-3B model on the *validation* subsets of the **DEV** multi-benchmark using 4 shots. We perform the ablation using batch size of 256 for *M3W*, 512 for *ALIGN*, 512 for *LTIP* and 64 for *VTP*. Models are trained for 1 million gradient steps (meaning 250,000 gradient updates, for the base model as we accumulate gradients over four datasets). All results are reported in Table 7. For each run, we provide an **Overall score** to better measure the overall quality of the trained models. This **Overall score** is produced by first normalizing each benchmark score with their respective reported state-of-the-art numbers from Table 3 and Table 4. These SotA normalized scores are then averaged across all 7 benchmarks to produce the **Overall score**. The higher the **Overall score** the better the model.

Method	VQAv2		COCO	VATEX	VizWiz		MSRVTTQA	VisDial		YouCook2	TextVQA		HatefulMemes
	test-dev	test-std	test	test	test-dev	test-std	test	valid	test-std	valid	valid	test-std	test seen
Flamingo - 32 shots	67.6	-	113.8	65.1	49.8	-	31.0	56.8	-	86.8	36.0	-	70.0
SimVLM [134]	80.0	80.3	143.3	-	-	-	-	-	-	-	-	-	-
OFA [129]	79.9	80.0	<u>149.6</u>	-	-	-	-	-	-	-	-	-	-
Florence [150]	80.2	80.4	-	-	-	-	-	-	-	-	-	-	-
Flamingo Fine-tuned	82.0	82.1	138.1	84.2	65.7	65.4	47.4	61.8	59.7	118.6	57.1	54.1	86.6
Restricted SotA [†]	80.2	80.4	143.3	76.3	-	-	46.8	75.2	74.5	138.7	54.7	73.7	75.4
[150]	[150]	[150]	[134]	[165]	-	-	[57]	[87]	[87]	[142]	[147]	[92]	[60]
Unrestricted SotA	81.3	81.3	<u>149.6</u>	81.4	57.2	60.6	-	-	75.4	-	-	-	84.6
[143]	[143]	[129]	[165]	[70]	[70]	-	-	[133]	-	-	-	-	[164]

Table 6 | **Comparison to SotA when fine-tuning Flamingo.** We fine-tune Flamingo on all nine tasks where Flamingo was SotA overall with few-shot learning. Flamingo sets a new SotA on five of these tasks sometimes even beating methods that resorts to known performance optimization tricks such as model ensembling (on VQAv2, VATEX, VizWiz and HatefulMemes). Best numbers among the restricted SotA are in **bold**. Best numbers overall are underlined. Restricted SotA[†]: only includes methods that use a single model (not ensembles) and do not directly optimise the test metric (no CIDEr optimisation).

4.4.1. Importance of the training data mixture

Improvements from additional datasets. Our final model is trained on the combination of four large-scale datasets described in Section 3.2: ALIGN (Jia et al., 2021), Long Text & Image Pairs (LTIP), Video & Text Pairs (VTP), and M3W (M3W). To highlight the contributions of the different types of training data, we investigate how removing datasets from the training mix affects the performance in row (i) of Table 7. Removing the M3W dataset results in important decreases in final scores on all tasks from the validation subsets multi-benchmark, resulting in an important drop in overall score, from 68.4 to 46.9. This demonstrates that interleaved data is crucial to develop the few-shot capability of the model. Removing the two paired image and text datasets (align and Long Text & Image Pairs) from the training set also negatively impacts the performance across tasks, resulting in a 11.9% drop in overall score. In particular, we find that including aligned text and image pairs during training is crucial for captioning and classification tasks. Adding Video & Text Pairs to the combination of image and text training datasets results in further improvement on the overall score. Though it comes with a minor decrease in performance on the some image tasks, including the collected aligned video and text dataset at train time leads to significant improvements on the three video tasks. Namely, removing the VTP dataset from training data decreases the CIDEr score on the VATEX dataset by 7.4%, the MSVD-QA top-1 score by 1.8% and the averaged top1-top5 accuracy on Kinetics by 3.4%. Although our model has never been trained on interleaved video and text data, we achieve high few-shot performance on video tasks. We observed that when trained on some dataset combinations, our model occasionally does not produce the final <EOC> token in the few-shot setting and instead predicts additional prompts for the target task. We therefore trim the prediction to the text preceding the prompt keywords before quantitative evaluation.

Optimisation strategy with multiple datasets. Finally, we ablate the co-training optimisation strategy on this heterogeneous mixture of data in the row (ii) of Table 7. We compare our gradient accumulation approach described in Section 3.2, against a round robin approach that was notably used in Cho et al. (2021) to train a VLM model on a heterogeneous data mixture. In short, the round robin approach sequentially alternates between the different datasets, and a gradient step is taken between each dataset. This leads to significantly worse results as the overall score dropped by almost 8.7%. We view gradient accumulation over different heterogeneous datasets as a mean to stabilize the

Ablated setting	Flamingo 3B value	Changed value	Param. count ↓	Step time ↓	COCO top1↑ CIDEr↑	OKVQA top1↑	VQAv2 top1↑	ImageNet top1↑	MSVDQA top1↑	VATEX CIDEr↑	Kinetics top1-top5↑	Overall score↑
Flamingo 3B model (short training)			3.2B	1.74s	86.5	42.1	55.8	59.9	36.3	53.4	49.4	68.4
(i) Training data	All data	M3W	3.2B	0.68s	58.0	37.2	48.6	35.7	29.5	33.6	34.0	50.7
		w/o VTP	3.2B	1.42s	84.2	43.0	53.9	59.6	34.5	46.0	45.8	65.4
		w/o LTIP/ALIGN	3.2B	0.95s	66.3	39.2	51.6	41.4	32.0	41.6	38.2	56.5
		w/o M3W	3.2B	1.02s	54.1	36.5	52.7	24.9	31.4	23.5	28.3	46.9
(ii) Optimisation	Grad. accumulation	Round Robin	3.2B	1.68s	76.1	39.8	52.1	50.7	33.2	40.8	39.7	59.7
(iii) Tanh gating	✓	✗	3.2B	1.74s	78.4	40.5	52.9	54.0	35.9	47.5	46.4	64.0
(iv) Cross-attention architecture	GATED XATTN-DENSE	VANILLA XATTN GRAFTING	2.4B 3.3B	1.16s 1.74s	80.6 79.2	41.5 36.1	53.4 50.8	59.0 47.5	32.9 32.2	50.7 47.8	46.8 27.9	65.2 57.4
(v) Cross-attention frequency	Every	Single in middle Every 4th Every 2nd	2.0B 2.3B 2.6B	0.87s 1.02s 1.24s	71.5 82.3 83.7	38.1 42.7 41.0	50.2 55.1 55.8	44.0 57.1 59.6	29.1 34.6 34.5	42.3 50.8 49.7	28.3 45.5 47.4	54.6 65.9 66.2
(vi) Resampler	Perceiver	MLP Transformer	3.2B 3.2B	1.85s 1.81s	78.6 83.2	42.2 41.7	54.7 55.6	53.6 59.0	35.2 31.5	44.7 48.3	42.1 47.4	63.3 65.1
(vii) Resampler size	Medium	Small Large	3.1B 3.4B	1.58s 1.87s	81.1 84.4	40.4 42.2	54.1 54.4	60.2 60.4	36.0 35.1	50.2 51.4	48.9 49.4	66.4 67.3
(viii) Multi-Img att.	Only last	All previous	3.2B	1.74s	70.0	40.9	52.0	52.3	32.1	46.8	42.0	60.8
(ix) p_{next}	0.5	0.0 1.0	3.2B 3.2B	1.74s 1.74s	85.0 81.3	41.6 43.3	55.2 55.6	60.3 57.8	36.7 36.8	50.6 52.7	49.9 47.8	67.8 67.6
(x) Vision encoder	NFNet-F6	CLIP ViT-L/14 NFNet-F0	3.1B 2.9B	1.58s 1.45s	76.5 73.8	41.6 40.5	53.4 52.8	49.5 49.8	33.2 31.1	44.5 42.9	42.3 36.6	61.4 58.9
(xi) LM pretraining	MassiveText	C4	3.2B	1.74s	81.3	34.4	47.1	60.6	30.9	53.9	46.9	62.5
(xii) Freezing Vision	✓	✗ (random init) ✗ (pretrained)	3.2B 3.2B	4.70s* 4.70s*	74.5 83.5	41.6 40.6	52.7 55.1	45.2 55.6	31.4 34.6	35.8 50.7	32.6 41.2	56.6 64.5
(xiii) Freezing LM	✓	✗ (random init) ✗ (pretrained)	3.2B 3.2B	2.42s 2.42s	74.8 81.2	31.5 33.7	45.6 47.4	59.5 60.7	26.9 31.0	50.1 53.9	43.4 49.9	58.2 62.9
(xiv) Co-train LM on MassiveText	✗	✓ (random init) ✓ (pretrained)	3.2B 3.2B	5.34s* 5.34s*	69.3 83.0	29.9 42.5	46.1 53.3	59.9 60.9	28.1 35.1	45.5 51.1	46.9 50.1	57.4 67.2

Table 7 | Ablation studies. Each row in this ablation study table should be compared to the baseline Flamingo run reported at the top of the table. The step time measures the time spent to perform gradient updates on all training datasets. (*): Due to higher memory usage, these models were trained using four times more TPU chips. The obtained accumulation step time was therefore multiplied by four.

training as it reduces the gradient variance between each update. Interestingly, gradient accumulation is also the best approach for the contrastive pretraining of the Vision Encoder (see Appendix C.2.2).

4.4.2. Key architectural components and training details

Tanh cross-attention gating. We ablate the use of the 0-initialized tanh gating when merging the cross-attention output to the frozen LM output in row (iii) of Table 7. Without it, we see a drop of 4.4% on our overall score. Moreover, we have noticed that disabling the 0-initialized tanh gating lead to training instabilities, with high peaks randomly observed in the losses.

Visual conditioning architectures for the frozen LM. We ablate the architecture used to condition the frozen LM on the vision data. We report the different conditioning architecture results in row (iv) of Table 7. VANILLA XATTN, refers to the original cross-attention architecture from the original Transformer decoder architecture (Vaswani et al., 2017), where a cross-attention layer is interleaved between a self-attention (here from the frozen LM) and an MLP layer (also from the frozen LM). In particular it does not add an extra dense feed forward layer as in our GATED XATTN-DENSE. Although used in recent work (Carion et al., 2020; Desai and Johnson, 2021), the VANILLA XATTN approach under-performs our proposed GATED XATTN-DENSE variant. Note that we added tanh gating to the VANILLA XATTN approach as we found it to be also beneficial for that architecture hence giving a more fair comparison. We also compared to the recent GRAFTING approach from Luo et al. (2022). In this approach, the frozen LM is used as is with no additional layers inserted, and a stack of interleaved self-attention and cross-attention layers that take the frozen LM output are

learnt from scratch.¹ Overall, interleaving the added cross-attention layers within the frozen LM outperforms by 11% the alternative method that grafts the layers on top of the frozen LM.

Compute/capacity vs. performance trade-off of cross-attention. We ablate here the frequency at which we add new GATED XATTN-DENSE blocks to the frozen transformer LM blocks. Results are reported in row (v) of Table 7. Although adding cross-attention at every layer leads to the best overall results, it also significantly increases the number of trainable parameters and time complexity of the model. Notably, halving the number of cross-attention by inserting GATED XATTN-DENSE every second blocks accelerates the training throughput of 40% and reduces the number of parameters by almost 20%. This incurred a moderate decrease of 2.2% in our overall score. This speedup and reduction in the number of trainable parameters are especially important in the regime of larger LMs. In fact, for a 70B frozen LM, halving the number of cross-attention layers reduces the parameter count from roughly 140B to 105B. However, reducing too aggressively the number of added layers is detrimental to performance as demonstrated by the 13.8% decrease in the overall score for the extreme case where a single GATED XATTN-DENSE is added to the middle of the network. In the light of this trade-off, we decide on a frequency for the bigger models by trying to maximize the number of added layers under our hardware constraints (both in terms of memory and time costs). This leads to our choice of adding a GATED XATTN-DENSE every fourth layers for our *Flamingo*-9B model and down to every seventh for our *Flamingo*-80B model. In these cases even though the frequency is decreased the overall added parameter count increases (because the original language model is deeper and wider) as shown in Table 1.

Resampler architecture and size. We ablate the architectural design of the Resampler in row (vi) and (vii) of Table 7. Given a parameter budget, we compare our proposed Perceiver Resampler to the use of a vanilla Transformer and an MLP (row (vi)). We show that both of these approaches lead to a significantly worse overall score, while also decreasing the training throughput. The Perceiver Resampler is more efficient as it compresses the large number of visual features to as few as 64 visual tokens. We also ablate the size of our Resampler with three options: Small, Medium (default value for all Flamingo models), and Large (row (vii)). We see that the best performance is achieved with a medium size Resampler. Moreover, when scaled together with the frozen LM, we observed that increasing the size of the Perceiver Resampler lead to instable trainings. We thus made the conservative choice of keeping the same medium Resampler size for all our Flamingo models.

Effect of how many images are cross-attended to. In the interleaved image-text scenario, we ablate whether or not a text following an image should only be able to attend to the single most recent previous image, or to all the previous images (row (viii) of Table 7). We can see that the single image case leads to significantly better results (8.2% better in the overall score). One potential explanation is that when attending to all previous image, there is no explicit way of disambiguating between different images in the cross-attention inputs. Nonetheless, recent work has shown that such disambiguation is still possible implicitly through the causal attention mechanism (Haviv et al., 2022). We also explored more explicit ways to enable this while attending to all previous images by modifying the image tags to include an index (<image 1>, <image 2>, etc.) and/or learning absolute index embeddings added to the cross-attention features for each image. These strategies were not as robust as our method when the number of images per sequence changes between train

¹Luo et al. (2022) also learn an additional embedding from scratch with a residual connection to add the logits produced by the frozen embedding; in our case, we found this did not work as well as simply using the frozen embedding, as in our main models, so we report the latter.

and test time. Such a property is desirable to reduce the number of images per sequence during training for better efficiency (we use $N = 5$ at train time) while still generalizing to many images for few-shot evaluation (we go up to $N = 32$ at test time). For these reasons, we keep the single image cross-attention strategy for the Flamingo models. Note that while text tokens cannot explicitly attend to all previous images due to this masking strategy, they can still implicitly attend to them from the language-only self-attention that propagates all previous images' features via the previous text tokens.

M3W image placement data augmentation. Given a webpage, we don't know in advance if the text of the page will mention the previous or the next image in the two-dimensional layout of the page DOM. For this reason, we explore a data augmentation on *M3W* controlled by p_{next} which indicates whether a given text token attends to the previous or the next image (see more details in Appendix A.1.2). The default value $p_{next} = \frac{1}{2}$ means that for each webpage sampled, we decide uniformly at random whether the model attends to the previous or next image. $p_{next} = 0$ means the model always attends to the previous image while $p_{next} = 1$ means the model always attends to the following image. The results (row (ix) of Table 7) show that using this randomization is beneficial.

4.4.3. Importance of pretraining and freezing models

Vision encoder pretraining. We evaluate the effect of different pretraining on the vision conditioning and the text generating part of the models. Results are given in row (x) of Table 7. We first compare our NFNet-F6 vision module, trained with a contrastive vision-language objective on our data mixture (details in Appendix C.2.1), to the publicly available CLIP (Radford et al., 2021) model² based on the ViT-L/14 (Dosovitskiy et al., 2020) backbone architecture. Our model has a +7% advantage on the CLIP ViT based model, which highlights the importance of picking a powerful vision backbone. Indeed, we hypothesize that this improvement is likely due to our better contrastive model compared to CLIP (as shown in Section 4.2.3). This trend is also confirmed when we switch to a smaller NFNet-F0 backbone (that is worse than our NFNet-F6 and CLIP when looking at the contrastive evaluation) as it incurs a 9.5% decrease in performance.

Language model pretraining. To measure the importance of text pretraining, we compare the performance of using a frozen decoder-only Transformer either pretrained on MassiveText (our main model) or pretrained on the C4 dataset (Raffel et al., 2019) (row (xi) of Table 7). Using the C4 dataset (which is smaller and less filtered than MassiveText) for training leads to a significant loss in performance (-5.9% overall). We note that the performance notably decreases for tasks that involve more language understanding such as visual question answering tasks (OKVQA, VQAv2 and MSVDQA) while it remains on par for tasks that do not require much language abilities (ImageNet or Kinetics). This highlights the importance of pretraining the LM on a high quality text-only dataset.

Freezing model components prevents catastrophic forgetting. During the Flamingo models training, we freeze the pretrained components (Vision Encoder and LM layers) while training newly added components from scratch. We ablate this choice by unfreezing the Vision Encoder and the LM layers during training starting either from our initialized weights or from scratch. Results are reported in rows (xii) and (xiii) of Table 7 for the Vision Encoder and for the LM layers, respectively. If trained from scratch, the performance decreases by a large margin in both cases (-11.8% for the Vision Encoder and -10.2 for the LM), highlighting again the importance of pretraining. Interestingly, starting from our good initialization while also allowing unfreezing the weights also leads to a drop

²<https://github.com/openai/CLIP>

in performance (-3.9% when unfreezing the Vision Encoder and -5.5% when unfreezing the LM). This is an instance of “catastrophic forgetting” (McCloskey and Cohen, 1989), in which the model progressively forgets its pretraining while training on a new objective. The information acquired during the pretraining on vision and on text turns out to improve downstream performance on multimodal vision and language tasks, and should not be lost. Freezing model parts avoids this issue.

Alternative to freezing the LM by co-training on MassiveText. Another approach for preventing catastrophic forgetting is to co-train on MassiveText (Rae et al., 2021), the dataset that was used to pretrain the language model. Specifically, we add MassiveText to the training mixture, with a weight λ_m of 1.0 (best performing after a small grid search), using a sequence length of 2048 and the exact same setting as the pretraining of Chinchilla (Hoffmann et al., 2022) for computing the text-only training loss. In order to co-train on MassiveText, we need to unfreeze the language model but we keep the vision encoder frozen. We perform two ablations in row (xiv) of Table 7: starting from a pretrained language model (with a learning rate multiplier of 0.1 of the LM weights) versus initializing from scratch (with the same learning rate everywhere). In both cases, the overall scores are worse than our baseline which starts from the language model, pretrained on MassiveText, and is kept frozen throughout training. This indicates that the strategy of freezing the language model to avoid catastrophic forgetting is strong. Even more importantly, freezing the LM is computationally cheaper as no gradient updates of the LM weights are required and we do not need to train on an additional dataset. This computational argument is even more relevant for our largest model, *Flamingo-80B*, where we freeze almost 90% of the overall weights.

5. Qualitative results

We provide some selected samples covering different interaction modalities in Figures 10, 11, and 12. Unlike the quantitative benchmark results (which use beam search with beam width 3 for decoding), all qualitative results presented in this section use greedy decoding for faster sampling.

Figure 10 shows the simplest form of interaction where a single image is provided followed by a text prompt either in the form of a question or the start of a caption. Even though the model is not trained in the question and answer form, the capabilities of the pretrained language model allows this adaptation. In many of these examples, *Flamingo* can do at least one step of implicit inference. Some of the objects are not named in the prompt but their properties are asked directly. In order to answer these questions, the model needs to infer the referred object then recall the relevant knowledge to form the answer. Previously, it has been shown that contrastively trained vision networks can learn some text reading capabilities. We observed that *Flamingo* preserves this capability in the full model, in some cases for text that is rather small with respect to the size of the image.

Since our model can accept inputs in the form of arbitrary sequences of visuals and language, we tested its abilities to hold an extended dialogue with interleaved images and text. Figure 11 shows some samples which are generated by prompting the model with a brief dialogue (Appendix D) followed by user interaction including image insertions. Even after several rounds of interaction *Flamingo* can still successfully attend to the image and reply to questions that can not be guessed by language alone. We observed that multiple images can be separately attended, simple comparisons and inferences are handled properly.

Lastly, we investigated similar capabilities with video inputs as they present some extra challenges compared to images. Figure 12 shows some selected samples. As seen in the figure, in some cases *Flamingo* can successfully integrate information from multiple frames (e.g. videos scanning through a scene or text) and respond to requests involving temporal understanding (e.g. “after”).

Input Prompt	 <p>Question: What do you think the capacities of these are? Answer:</p>	 <p>Question: What is odd about this image? Explain why it is unusual. Answer:</p>	 <p>Question: What country is this? Why do you think so? Answer:</p>
Completion	<p>The floppy disk is 1.44MB and the CD is 700MB.</p>	<p>The image is odd because the elephant is in the back of the truck. It is unusual because elephants are not usually transported in the back of a truck.</p>	<p>It is Canada. I think so because the flag is the Canadian flag.</p>
Input Prompt	 <p>Question: What nutrients is the fruit in the image rich in? Answer:</p>	 <p>Question: What is the purpose of this? Answer:</p>	 <p>Question: What season is this? Explain why. Answer:</p>
Completion	<p>Vitamin C, potassium, and fiber.</p>	<p>To keep the mosquitoes out of the bed.</p>	<p>This is winter season. The Christmas tree and the presents are the main reasons why it is winter.</p>
Input Prompt	 <p>Question: What is the title of the book? Answer:</p>	 <p>The ingredients of this dish are</p>	 <p>The soundtrack includes</p>
Completion	<p>The House Book.</p>	<p>: beef, potatoes, carrots, broccoli, and lemon.</p>	<p>the songs "Let It Go" and "For the First Time in Forever" by Tony Award® winner Idina Menzel, who plays Elsa [...]</p>

Figure 10 | Selected single image samples. Gray boxes are user input and the pink boxes are *Flamingo* output.



This is a dog. It's a white fluffy dog.

What season is this?

It looks like a summer day.

Why do you think so?

The grass is green and the dog is not wearing a sweater.

What breed is the dog?

It's a Samoyed.

Is the dog running?

No, it's sitting.

Can you describe the pose of its back legs?

The back legs are bent and the dog is sitting on its haunches.



Which city is this?

This is a street sign in New York City.

Which street corner is it?

It is the intersection of Madison Avenue and East 42nd Street.

Which subway station is nearby?

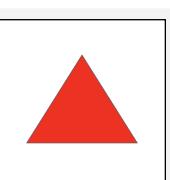
The closest subway station is Grand Central Station.

How about the nearest park?

The nearest park is Bryant Park.

There is another sign. What does it say?

It says "One way".



I would like you to read the color and then say the color it is written in. For example:

PURPLE

Color is "Purple" and it is written in red.

GREEN

Color is "Green" and it is written in blue.

YELLOW

Color is "Yellow" and it is written in green.

BLACK

Color is "Black" and it is written in yellow.

Well done! Do you know the name of the test these images come from?

I think it is called the Stroop test.

Can you explain how humans perform in this test?

Humans are slower when the color of the word and the color of the word are different.

How about you?

I am not affected by this difference.



What is in this picture?

It's a bowl of soup with a monster face on it.

What is the monster made out of?

It's made out of vegetables.

No, it's made out of a kind of fabric. Can you see what kind?

It's made out of a woolen fabric.



This is a picture of a dog. It looks like it is a little bit naughty.

What happened here?

It looks like the dog is destroying the sofa.

How does the person thinking about it?

The person is angry.

What is the person holding?

It looks like a cup of coffee.

How is the weather outside?

It looks like a sunny day.



What is the name of the company on this card?

It's Homebase.

What is its web url?

It's homebase.co.uk.

What are they doing?

They are a home improvement retailer.



What do you think the robot is doing here?

The robot is picking up a ball.

What color is the ball?

It's red.

I want the ball to be hidden. Is there something the robot can use for that?

The robot can use a cloth to hide the ball.

What color is the cloth?

The cloth is blue.

Figure 11 | Selected dialogue samples. Gray boxes are user input and the pink boxes are *Flamingo* output. For dialogue, *Flamingo* is provided with a custom prompt (hidden from the visualization but shown in Appendix D) containing a dialogue with 3 random images, but it is not fine-tuned for dialogue in any other way.

32

Made with Goodnotes

Input Prompt	Completion
	Question: What is happening here? Answer: The dachshund puppy is being weighed on a scale.
	Question: What object is the avatar picking up? Answer: A sword.
	Question: What is the object being lifted? Answer: The object is a small plastic bowl.
	Question: What are the three objects in this video? Answer: A scissors, a pen, and a rubber band.
	Question: What is written here? Answer: Flamingo.
	Question: What happens to the man after hitting the ball? Answer: he falls down.

Figure 12 | Selected video samples. These are all of the frames the model sees. (Best viewed with zoom.)

6. Discussion

6.1. Limitations, failure cases and opportunities

Here, we describe some limitations and failure cases of our models; as well as opportunities for further improving our models and extending their abilities.

Classification performance. Although our visual language models have important advantages over contrastive models (e.g., few-shot learning and open-ended generation capabilities), their performance lags behind that of contrastive models on classification tasks. We believe this is because the contrastive training objective directly optimizes for text-image retrieval, and in practice, the evaluation procedure for classification can be thought of as a special case of image-to-text retrieval (Radford et al., 2021). This is not the case for the language modeling objective we use to train our visual language models and this may contribute to the observed performance gap on classification tasks. In particular, Zhao et al. (2021) have shown that language models suffer from various biases arising from the training data distribution, the set of samples used in the prompt, and their order. They also show that such issues can be mitigated with calibration techniques, provided one can assume a certain prior distribution (e.g., uniform) over the label space. This assumption doesn't hold in general, and further research is needed to develop techniques to address these issues in the few-shot setting. More generally, seeking objectives, architectures, or evaluation procedures that could bridge the gap between these two classes of models is a promising research direction.

Legacies of language models. Our models build on powerful pretrained causal language models, and as a side effect, directly inherit their weaknesses. For instance, causal modeling of the conditioning inputs is strictly less expressive than bidirectional modeling. In this direction, recent work has shown that non-causal masked language modeling adaptation (Wang et al., 2022) followed by multitask

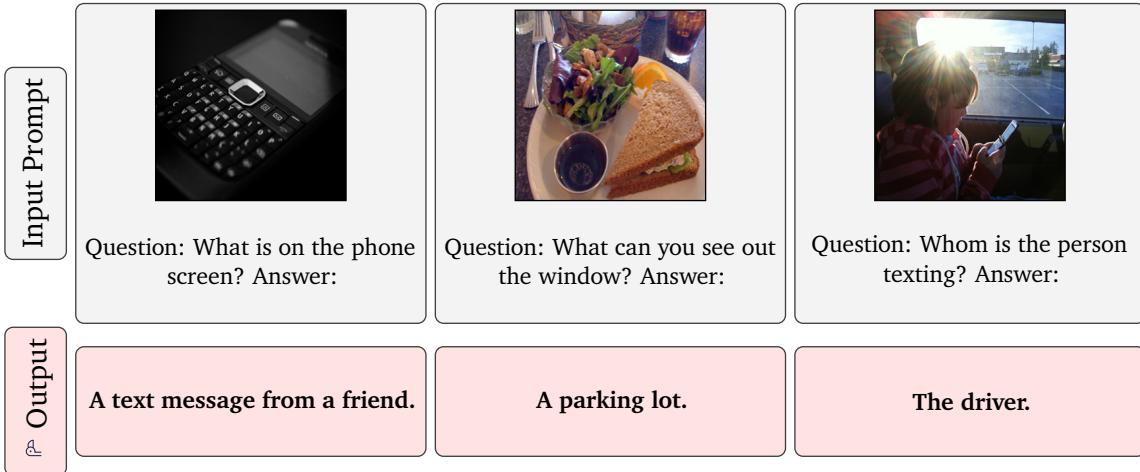


Figure 13 | **Hallucinations and ungrounded guesses in open-ended visual question answering.** *Left:* The model occasionally hallucinates by producing answers that seem likely given the text only, but are wrong given the image as additional input. *Middle:* Similar hallucinations can be provoked by adversarially prompting the model with an irrelevant question. *Right:* A more common pitfall arises when the model makes ungrounded guesses when the answer cannot be determined based on the inputs. Few-shot examples and more sophisticated prompt design may be used to mitigate these issues. More broadly, addressing these issues is an important research direction towards improving our models’ applications in open-ended visual dialogue settings.

fine-tuning ([Sanh et al., 2022](#); [Wei et al., 2021](#); [Xu et al., 2022](#)) can efficiently improve the zero-shot performance of causal decoder-only language models. Furthermore, transformer-based language models tend to generalize poorly to test sequences significantly longer than the training ones ([Press et al., 2022](#)). In settings where the expected text output is too long, the ability of the models to leverage enough shots for few-shot learning can be affected. For instance, for the VisDial dataset ([Das et al., 2017](#)), a single shot consists of an image followed by a long dialogue composed of 21 different sentences. A sequence of 32 VisDial shots is thus composed of at least $32 \times 21 = 672$ sentences, which in practice means that the prompt length ranges from 4096 to 8192 tokens. This is significantly longer than the maximum sequence length (2048) our LMs have been trained on ([Hoffmann et al., 2022](#)). Empirically, we observe a large drop of roughly 30% in relative performance on VisDial when going from 16-shots to 32-shot, likely due to this limitation. On another note, while our ablations demonstrate the importance of the language model priors inherited from frozen language models, we suspect that they may play a role in occasional hallucinations and ungrounded guesses observed in open-ended dialogue settings. We provide and analyze examples of such behaviours in Figure 13. Finally, language modeling suffers from poor sample efficiency during pretraining ([Brown et al., 2020](#)). Mitigating this issue has the potential to greatly accelerate progress in the field, by improving turnaround of large-scale training runs and in turn increasing feasibility of more systematic exploration of design decisions at larger scales. Further discussion on typical weaknesses observed for large LMs can be found in ([Brown et al., 2020](#); [Rae et al., 2021](#)).

Trade-offs of few-shot learning methods. In the paper, we use in-context learning as our “go-to” few-shot learning method (see Section 4.2). This method has notable advantages over gradient-based approaches such as fine-tuning (see Section 4.3). Indeed, in-context learning requires almost no hyperparameter tuning, works reasonably well in the very low data regime (dozens of examples), and only requires inference, simplifying deployment. In contrast, gradient-based approaches require carefully tuned design choices to avoid overfitting (either by proper learning rate schedule or architecture

design (Houlsby et al., 2019)) and often need more data (thousands) to work well. This motivated our focus on in-context learning; however, this approach also has drawbacks we discuss next.

Inference compute cost. The compute cost of in-context learning with transformer models scales linearly with the number of shots if one can reuse the few-shot prompt for multiple query samples (by caching the keys and values) and quadratically otherwise. In contrast, gradient-based few-shot learning approaches (Houlsby et al., 2019) have constant complexity with respect to the number of shots during inference.

Prompt sensitivity. In-context learning has also been shown to be disconcertingly sensitive to various aspects of the demonstrations, such as the order of the samples (Zhao et al., 2021) or their format.

Leveraging more shots. When using in-context learning, performance plateaus rapidly as the number of few-shot samples increases beyond 32. This proves a striking contrast with typical gradient-based methods, for which the amount of correctly paired training data is a critical factor of performance. We note that RICES (Retrieval In-Context Example Selection (Yang et al., 2021) described in Section 3.3) effectively mitigates this issue for classification tasks (Section 4.2.2), but still faces similar issues beyond a small number of example per class.

Task location. Recent work on understanding what makes in-context learning effective sheds some light on a possible explanation for why more shots do not always help (Min et al., 2022; Reynolds and McDonell, 2021). In more detail, Brown et al. (2020) raise the question of whether in-context learning actually “learns” new tasks at inference time based on the provided input-output mappings, or simply recognizes and identifies tasks learned during training. On this question, the findings of Reynolds and McDonell (2021) suggest that the latter is the key driver of performance across diverse settings, and refer it as *task location*. Similarly, Min et al. (2022) show that the mapping from input to output generally has limited impact on few-shot performance, as opposed to specifying the overall format of the examples. In line with these findings, we also observe non-trivial zero-shot performance using prompt without any images, hence also highlighting that the format of the task matters a lot. Intuitively, a handful of samples may often be enough to perform task location well, but the model may generally not be able to leverage further samples at inference time to refine its behaviour.

In summary, there is no “golden” few-shot method that would work well in all scenarios. In particular, the best choice of few-shot learning approach highly depends on characteristics of the application, an important one being the number of annotated samples. On this point, in our work, we demonstrate high effectiveness of in-context learning for the data-starved regime (i.e. below 32 samples). There may be opportunities to combine different methods to leverage their complementary benefits; in particular when targeting less data-constrained data regimes (e.g. hundreds of samples).

Extending the visual and text interface. Natural language is a powerful and versatile input/output interface to provide descriptions of visual tasks to the model and generate outputs or estimate conditional likelihoods over possible outputs. However, it may be a cumbersome interface for tasks that involve conditioning on or predicting more structured outputs such as bounding boxes (or their temporal and spatio-temporal counterparts); as well as making spatially (or temporally and spatio-temporally) dense predictions. Furthermore, some vision tasks, such as predicting optical flow, involve predicting in continuous space, which is not something our model is designed to handle out of the box. Finally, one may consider additional modalities besides vision, that may be complementary, such as audio. All of these directions have the potential to extend the range of tasks that our models can handle; and even improve performance on the ones we focus on, thanks to synergies between the corresponding abilities.

Scaling laws for vision-language models. In this work, we scale Flamingo models up to 80B parameters and provide some initial insights on their scaling behaviour across evaluation benchmarks, summarized in Figure 9. In the language space, an important line of work has focused on establishing scaling laws for language models (Hoffmann et al., 2022; Kaplan et al., 2020). In the vision domain, Zhai et al. (2021) take a step in this direction. Similar efforts have yet to be made for vision-language models, including contrastive models, as well as visual language models such as the ones we propose. While language modeling scaling law research has focused on perplexity as the golden metric, we speculate that it may be more directly useful for our purposes to establish such trends in terms of aggregate downstream evaluation task performance.

6.2. Benefits, risks and mitigation strategies

6.2.1. Benefits

Accessibility. A system like Flamingo offers a number of potential societal benefits, some of which we will discuss in this section. Broadly, the fact that Flamingo is capable of task generalisation makes it suitable for use cases that have not been the focus of vision research historically. Typical vision systems are trained to solve a particular problem by training on large databases of manually annotated task-specific examples, making them poorly suited for applications outside of the narrow use cases for which they were deliberately trained. On the other hand, Flamingo is trained in a minimally constrained setting, endowing it with strong few-shot task induction capabilities. As we've shown in our qualitative examples (Section 5), Flamingo can also be used through a "chat"-like interface for open-ended dialogue. Such capabilities could enable non-expert end users to apply models like Flamingo even to low-resource problems for which little to no task-specific training data has been collected, and where queries might be posed in a variety of formats and writing styles. In this direction, we have shown that *Flamingo* achieves strong performance on the VizWiz challenge³, which promotes visual recognition technologies to assist visually impaired people. A dialogue interface could also promote better understanding and interpretability of visual language models. It could help highlight issues with bias, fairness, and toxicity the model may pick up on from the training data. Overall, we believe that Flamingo represents an important step towards making state-of-the-art visual recognition technology more broadly accessible and useful across a myriad of diverse applications.

Model recycling. From a modeling perspective, although Flamingo is computationally expensive to train, it importantly leverages pretrained frozen language models and visual encoders. We demonstrated that new modalities can be introduced into frozen models, thereby avoiding expensive retraining. As such models continue to grow in size and computational demands, "recycling" them will become increasingly important from an environmental perspective (as well as a practical one), as explored in Strubell et al. (2019) for language models. We hope such results may inspire further research into how existing models can be repurposed efficiently rather than trained from scratch.

6.2.2. Risks and mitigation strategies

This section provides some early investigations of the potential risks of models like Flamingo. This study is preliminary and we foresee that further research efforts should be undertaken to better assess those risks. We also discuss potential mitigation strategies towards safely deploying these models. Note that as explained in our Model Card (Mitchell et al., 2019) in Appendix B, this model was developed for research purposes only and should not be used in specific applications before proper risk analyses are conducted and mitigation strategies are explored.

³<https://vizwiz.org/>

		CIDEr difference female - male = Δ	CIDEr difference darker - lighter = Δ	CIDEr overall
AoANet [52]		-	+0.0019	1.198
Oscar [67]		-	+0.0030	1.278
Flamingo, 0 shot	0.899 – 0.870 = +0.029 ($p = 0.52$)	0.955 – 0.864 = +0.091 ($p = 0.25$)		0.843
Flamingo, 32 shots	1.172 – 1.142 = +0.030 ($p = 0.54$)	1.128 – 1.152 = -0.025 ($p = 0.76$)		1.138

Table 8 | **Bias evaluation of Flamingo for COCO captioning.** We report results on the COCO dataset splits over gender and skin tone provided by Zhao et al. (2021).

By construction, Flamingo inherits the risks of Large LMs. Recall that a large part of our model is obtained by freezing the weights of an existing language model (Hoffmann et al., 2022). In particular, if provided with no images Flamingo falls back to language model behavior. As such Flamingo is exposed to the same risks of large language models: it can output potentially offensive language, propagate social biases and stereotypes, as well as leaking private information (Weidinger et al., 2021). In particular, we refer to the analysis presented in the Chinchilla paper (Hoffmann et al. (2022), Section 4.2.7) in terms of gender bias on the Winogender dataset (Rudinger et al., 2018) which demonstrate that even though this model is less biased towards gender than previous models (Rae et al., 2021), gender biases are still present. In terms of unprompted toxicity, we also refer to the analysis from Chinchilla (Hoffmann et al., 2022) which highlights that overall the propensity of the model to produce toxic outputs when not prompted to do so is rather low, as measured by computing the PerspectiveAPI toxicity score on 25,000 samples. Weidinger et al. (2021) detail possible long-term mitigation strategies for these risks. They include social or public policy interventions, such as the creation of regulatory frameworks and guidelines; careful product design, for instance relating to user interface decisions; and research at the intersection between AI Ethics and NLP, such as building better benchmarks and improving mitigation strategies. In the short term, effective approaches include relying on prompting to mitigate any biases and harmful outputs (Rae et al., 2021). Next, we explore the additional risks incurred by Flamingo’s additional visual input capabilities.

Gender and racial biases when prompted with images. Previous work has studied biases that exist in captioning systems (Hendricks et al., 2018; Zhao et al., 2021). Such modeling biases can result in real-world harms if deployed without care. For AI systems to be useful to society as a whole, their performance should not depend on the perceived skin tone or gender of the subjects – they should work equally well for all populations. However, current automatic vision system performance has been reported to vary with race, gender or when applied across different demographics and geographic regions (Buolamwini and Gebru, 2018; De Vries et al., 2019; Schwemmer et al., 2020). As a preliminary study assessing how Flamingo’s performance varies between populations, we follow the study proposed in Zhao et al. (2021) and report how the captioning performance of our model varies on COCO as a function of gender and race. Note that we use a different evaluation protocol from the one proposed by Zhao et al. (2021); in that work, they measure results across 5 pretrained models and compute confidence intervals across aggregated per-model scores. Here, we have just one copy of our model (due to its high training cost), and we instead perform statistical tests on the per-sample CIDEr scores across the splits from Zhao et al. (2021). We report the results in Table 8.

Overall, when comparing the CIDEr scores aggregated among images labeled as *female* versus *male*, as well as when comparing *darker skin* versus *lighter skin*, we find there are no statistically significant differences in the per-sample CIDEr scores. To compare the two sets of samples, we use a two-tailed *t*-test with unequal variance, and among the four comparisons considered, the lowest *p*-value we find is $p = 0.25$, well above typical statistical significance thresholds (e.g. a common rejection threshold

might be $p < \alpha = 0.05$). This implies that the differences in scores are indistinguishable from random variation under the null hypothesis that the mean scores are equal. We note that a failure to reject the null hypothesis and demonstrate a significant difference does not imply that there are no significant differences; it is possible that a difference exists that could be demonstrated with larger sample sizes, for example. However, these preliminary results are nonetheless encouraging.

Toxicity when prompted with images. We also evaluate the toxicity of *Flamingo* using the *Perspective API*⁴ to evaluate the toxicity of the model’s generated captions when prompted with images from the COCO test set. We observe that some captions are labelled as toxic by the classifier; however, when examining them manually, we do not observe any clear toxicity – output captions are appropriate for the images provided. Overall, based on our own experiences interacting with the system throughout the course of the project, we have not observed toxic outputs when given “safe-for-work” imagery. However this does not mean the model is not capable of toxic outputs, especially if probed with “not-safe-for-work” images and/or toxic text. A more thorough exploration and study would be needed if such a model were put in production.

Applying Flamingo for mitigation strategies. Thanks to its ability to rapidly adapt in low-resource settings, *Flamingo* could itself be applied in addressing some of the issues described above. For instance, following [Thoppilan et al. \(2022\)](#), adequately conditioned or fine-tuned *Flamingo* models could be used for filtering purposes of toxic or harmful samples in the training data. In their work, they observe significant improvements relating to safety and quality when fine-tuning on the resulting data. Furthermore, during evaluation, such adapted models could be used to down-rank or exclude outputs that might be classified as offensive, promoting social biases and stereotypes or leaking private information, thus accelerating progress in this direction even for low-resource tasks. Our results on the HatefulMemes benchmark represent a promising step in this direction. Recent work in the language modeling space has also shown success of training an LM to play the role of a red team, and generate test cases, so as to automatically find cases where another target LM behaves in a harmful way ([Perez et al., 2022](#)). A similar approach could be derived for our setting. Enabling the model to support outputs with reference to particular locations within the visual inputs, or to external verified quotes is also interesting direction ([Menick et al., 2022](#); [Thoppilan et al., 2022](#)). Finally, in Figure 11, we provide qualitative examples demonstrating that *Flamingo* can explain its own outputs, suggesting avenues to explainability and interpretability using the model’s text interface.

7. Conclusion

We propose a general-purpose family of *Flamingo* models that can be applied to image and video understanding tasks with minimal task-specific training data. With just a few examples presented to the model, a single *Flamingo* model can achieve state-of-the-art results on a wide array of tasks, performing competitively with approaches requiring task-specific fine-tuning on orders of magnitude more examples, and often requiring hand-engineered “tricks”. We’ve further presented qualitative examples showing interesting interactive abilities, allowing users to “chat” with the model, querying it for arbitrary information about input images and videos, demonstrating our models’ flexibility beyond traditional vision and language benchmarks. Our results suggest that the *Flamingo* model, a visual language model that bridges the gap between large language models and powerful visual representations, represents an important step towards general-purpose visual understanding.

⁴<https://perspectiveapi.com/>

References

- [1] Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, and Luke Zettlemoyer. CM3: A causal masked multimodal model of the internet. *arXiv:2201.07520*, 2022.
- [2] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *Conference on Neural Information Processing Systems*, 2020.
- [3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *International Conference on Computer Vision*, 2015.
- [4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv:1607.06450*, 2016.
- [5] Thomas Bachlechner, Bodhisattwa Prasad Majumder, Henry Mao, Gary Cottrell, and Julian McAuley. ReZero is all you need: Fast convergence at large depth. In *Uncertainty in Artificial Intelligence*, 2021.
- [6] Max Bain, Arsha Nagrani, Gülcin Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *International Conference on Computer Vision*, 2021.
- [7] Luca Bertinetto, João F. Henriques, Jack Valmadre, Philip Torr, and Andrea Vedaldi. Learning feed-forward one-shot learners. *Conference on Neural Information Processing Systems*, 2016.
- [8] Luca Bertinetto, Joao F. Henriques, Philip H. S. Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. *arXiv:1805.08136*, 2018.
- [9] 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. JAX: composable transformations of Python+NumPy programs, 2018. URL <http://github.com/google/jax>.
- [10] John S. Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neurocomputing*, 1990.
- [11] Andrew Brock, Soham De, Samuel L. Smith, and Karen Simonyan. High-performance large-scale image recognition without normalization. *arXiv:2102.06171*, 2021.
- [12] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Conference on Neural Information Processing Systems*, 2020.
- [13] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *ACM Conference on Fairness, Accountability, and Transparency*, 2018.

- [14] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, 2020.
- [15] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *IEEE Computer Vision and Pattern Recognition*, 2021.
- [16] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. *arXiv:2102.10407*, 2021.
- [17] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. *arXiv:2109.10852*, 2021.
- [18] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv:1504.00325*, 2015.
- [19] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: Universal image-text representation learning. In *European Conference on Computer Vision*, 2020.
- [20] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, 2021.
- [21] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. PaLM: Scaling language modeling with pathways. *arXiv:2204.02311*, 2022.
- [22] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *IEEE Computer Vision and Pattern Recognition*, 2017.
- [23] Terrance De Vries, Ishan Misra, Changhan Wang, and Laurens Van der Maaten. Does object recognition work for everyone? In *IEEE Computer Vision and Pattern Recognition*, 2019.
- [24] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *IEEE Computer Vision and Pattern Recognition*, 2021.
- [25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2018.
- [26] Carl Doersch, Ankush Gupta, and Andrew Zisserman. Crosstransformers: spatially-aware few-shot transfer. *Conference on Neural Information Processing Systems*, 2020.

- [27] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *IEEE Computer Vision and Pattern Recognition*, 2015.
- [28] 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. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*, 2020.
- [29] Constantin Eichenberg, Sidney Black, Samuel Weinbach, Letitia Parcalabescu, and Anette Frank. MAGMA—multimodal augmentation of generative models through adapter-based finetuning. *arXiv:2112.05253*, 2021.
- [30] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006.
- [31] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, 2017.
- [32] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. VIOLET: End-to-end video-language transformers with masked visual-token modeling. *arXiv:2111.12681*, 2021.
- [33] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In *Conference on Neural Information Processing Systems*, 2020.
- [34] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv:2101.00027*, 2020.
- [35] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 2021.
- [36] Jonathan Gordon, John Bronskill, Matthias Bauer, Sebastian Nowozin, and Richard E. Turner. Meta-learning probabilistic inference for prediction. *arXiv:1805.09921*, 2018.
- [37] Alex Graves. Generating sequences with recurrent neural networks. *arXiv:1308.0850*, 2013.
- [38] Thomas L. Griffiths, Frederick Callaway, Michael B. Chang, Erin Grant, Paul M. Krueger, and Falk Lieder. Doing more with less: meta-reasoning and meta-learning in humans and machines. *Current Opinion in Behavioral Sciences*, 2019.
- [39] Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. KAT: A knowledge augmented transformer for vision-and-language. *arXiv:2112.08614*, 2021.
- [40] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. VizWiz grand challenge: Answering visual questions from blind people. In *IEEE Computer Vision and Pattern Recognition*, 2018.
- [41] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *International Conference on Computer Vision*, 2017.

- [42] Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, and Omer Levy. Transformer language models without positional encodings still learn positional information. *arXiv:2203.16634*, 2022.
- [43] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision*, 2018.
- [44] Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. Decoupling the role of data, attention, and losses in multimodal transformers. *Annual Meeting of the Association for Computational Linguistics*, 2021.
- [45] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv:1606.08415*, 2016.
- [46] Tom Hennigan, Trevor Cai, Tamara Norman, and Igor Babuschkin. Haiku: Sonnet for JAX, 2020. URL <http://github.com/deepmind/dm-haiku>.
- [47] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 1997.
- [48] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Eric Noland, Tom Hennigan, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *arXiv:2203.15556*, 2022.
- [49] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, 2019.
- [50] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv:1801.06146*, 2018.
- [51] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. *arXiv:2111.12233*, 2021.
- [52] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *International Conference on Computer Vision*, 2019.
- [53] Md Amirul Islam, Matthew Kowal, Sen Jia, Konstantinos G. Derpanis, and Neil D. B. Bruce. Global pooling, more than meets the eye: Position information is encoded channel-wise in CNNs. In *International Conference on Computer Vision*, 2021.
- [54] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning*, 2021.
- [55] Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. MURAL: multimodal, multitask retrieval across languages. *arXiv:2109.05125*, 2021.
- [56] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv:2102.05918*, 2021.

- [57] Alex Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. All in one: Exploring unified video-language pre-training. *arXiv:2203.07303*, 2022.
- [58] Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling. *arXiv:1602.02410*, 2016.
- [59] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv:2001.08361*, 2020.
- [60] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The Hateful Memes Challenge: Detecting hate speech in multimodal memes. *Conference on Neural Information Processing Systems*, 2020.
- [61] Brenden Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua Tenenbaum. One shot learning of simple visual concepts. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2011.
- [62] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv:2104.08691*, 2021.
- [63] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Conference on Neural Information Processing Systems*, 2021.
- [64] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv:2201.12086*, 2022.
- [65] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. HERO: Hierarchical encoder for video+language omni-representation pre-training. *arXiv:2005.00200*, 2020.
- [66] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv:2101.00190*, 2021.
- [67] Xiuju Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, 2020.
- [68] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? *arXiv:2101.06804*, 2021.
- [69] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Optimization of image description metrics using policy gradient methods. In *International Conference on Computer Vision*, 2017.
- [70] Yu Liu, Lianghua Huang, Liuyihang Song, Bin Wang, Yingya Zhang, and Pan Pan. Enhancing textual cues in multi-modal transformers for vqa. *VizWiz Challenge 2021*, 2021.
- [71] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *International Conference on Computer Vision*, 2021.

- [72] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Conference on Neural Information Processing Systems*, 2019.
- [73] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. UniVL: A unified video and language pre-training model for multimodal understanding and generation. *arXiv:2002.06353*, 2020.
- [74] Ziyang Luo, Yadong Xi, Rongsheng Zhang, and Jing Ma. VC-GPT: Visual conditioned GPT for end-to-end generative vision-and-language pre-training. *arXiv:2201.12723*, 2022.
- [75] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Computer Vision and Pattern Recognition*, 2019.
- [76] Ellen M. Markman. *Categorization and naming in children: Problems of induction*. MIT Press, 1989.
- [77] Daniela Massiceti, Luisa Zintgraf, John Bronskill, Lida Theodorou, Matthew Tobias Harris, Edward Cutrell, Cecily Morrison, Katja Hofmann, and Simone Stumpf. ORBIT: A real-world few-shot dataset for teachable object recognition. In *International Conference on Computer Vision*, 2021.
- [78] Michael McCloskey and Neil J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *The Psychology of Learning and Motivation*, 1989.
- [79] Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. Teaching language models to support answers with verified quotes. *arXiv:2203.11147*, 2022.
- [80] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In *International Conference on Computer Vision*, 2019.
- [81] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. RareAct: A video dataset of unusual interactions. *arxiv:2008.01018*, 2020.
- [82] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *IEEE Computer Vision and Pattern Recognition*, 2020.
- [83] Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. Recurrent neural network based language model. *Interspeech*, 2010.
- [84] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv:2202.12837*, 2022.
- [85] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *ACM Conference on Fairness, Accountability, and Transparency*, 2019.
- [86] Ron Mokady, Amir Hertz, and Amit H. Bermano. ClipCap: CLIP prefix for image captioning. *arXiv:2111.09734*, 2021.

- [87] Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In *European Conference on Computer Vision*, 2020.
- [88] Ethan Perez, Douwe Kiela, and Kyunghyun Cho. True few-shot learning with language models. *Conference on Neural Information Processing Systems*, 2021.
- [89] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. *arXiv:2202.03286*, 2022.
- [90] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Hanxiao Liu, Adams Wei Yu, Minh-Thang Luong, Mingxing Tan, and Quoc V. Le. Combined scaling for zero-shot transfer learning. *arXiv:2111.10050*, 2021.
- [91] Ofir Press, Noah Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2022.
- [92] Yixuan Qiao, Hao Chen, Jun Wang, Yihao Chen, Xianbin Ye, Ziliang Li, Xianbiao Qi, Peng Gao, and Guotong Xie. Winner team Mia at TextVQA Challenge 2021: Vision-and-language representation learning with pre-trained sequence-to-sequence model. *arXiv:2106.15332*, 2021.
- [93] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *Preprint*, 2019.
- [94] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv:2103.00020*, 2021.
- [95] Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling language models: Methods, analysis & insights from training Gopher. *arXiv:2112.11446*, 2021.
- [96] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv:1910.10683*, 2019.

- [97] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: Memory optimizations toward training trillion parameter models. In *International Conference for High Performance Computing, Networking, Storage and Analysis*, 2020.
- [98] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv:2204.06125*, 2022.
- [99] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *IEEE Computer Vision and Pattern Recognition*, 2017.
- [100] James Requeima, Jonathan Gordon, John Bronskill, Sebastian Nowozin, and Richard E. Turner. Fast and flexible multi-task classification using conditional neural adaptive processes. *Conference on Neural Information Processing Systems*, 2019.
- [101] Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.
- [102] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. *arXiv:1804.09301*, 2018.
- [103] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015.
- [104] Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M. Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *International Conference on Learning Representations*, 2022.
- [105] Carsten Schwemmer, Carly Knight, Emily D. Bello-Pardo, Stan Oklobdzija, Martijn Schoonvelde, and Jeffrey W. Lockhart. Diagnosing gender bias in image recognition systems. *Socius*, 2020.
- [106] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics*, 2018.
- [107] Mohammad Shoeybi, Mostafa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-LM: Training multi-billion parameter language models using model parallelism. *arXiv:2104.08691*, 2019.
- [108] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *IEEE Computer Vision and Pattern Recognition*, 2019.
- [109] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A foundational language and vision alignment model. *arXiv:2112.04482*, 2021.

- [110] Lucas Smaira, João Carreira, Eric Noland, Ellen Clancy, Amy Wu, and Andrew Zisserman. A short note on the Kinetics-700-2020 human action dataset. *arXiv:2010.10864*, 2020.
- [111] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Conference on Neural Information Processing Systems*, 2017.
- [112] David R So, Wojciech Mańke, Hanxiao Liu, Zihang Dai, Noam Shazeer, and Quoc V. Le. Primer: Searching for efficient transformers for language modeling. *arXiv:2109.08668*, 2021.
- [113] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. *arXiv:1906.02243*, 2019.
- [114] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: Pre-training of generic visual-linguistic representations. *arXiv:1908.08530*, 2019.
- [115] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. VideoBERT: A joint model for video and language representation learning. In *International Conference on Computer Vision*, 2019.
- [116] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. VL-Adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *IEEE Computer Vision and Pattern Recognition*, 2021.
- [117] Ilya Sutskever, James Martens, and Geoffrey E. Hinton. Generating text with recurrent neural networks. In *International Conference on Machine Learning*, 2011.
- [118] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformer. In *Conference on Empirical Methods in Natural Language Processing*, 2019.
- [119] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 2016.
- [120] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehai Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. LaMDA: Language models for dialog applications. *arXiv:2201.08239*, 2022.
- [121] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B. Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *European Conference on Computer Vision*, 2020.
- [122] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou. Fixing the train-test resolution discrepancy. *Conference on Neural Information Processing Systems*, 2019.

- [123] Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, and Hugo Larochelle. Meta-dataset: A dataset of datasets for learning to learn from few examples. *arXiv:1903.03096*, 2019.
- [124] Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Conference on Neural Information Processing Systems*, 2021.
- [125] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Conference on Neural Information Processing Systems*, 2017.
- [126] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *International Conference on Computer Vision*, 2015.
- [127] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. *Conference on Neural Information Processing Systems*, 2016.
- [128] Jianfeng Wang, Xiaowei Hu, Zhe Gan, Zhengyuan Yang, Xiyang Dai, Zicheng Liu, Yumao Lu, and Lijuan Wang. UFO: A unified transformer for vision-language representation learning. *arXiv:2111.10023*, 2021.
- [129] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv:2202.03052*, 2022.
- [130] Thomas Wang, Adam Roberts, Daniel Hesslow, Teven Le Scao, Hyung Won Chung, Iz Beltagy, Julien Launay, and Colin Raffel. What language model architecture and pretraining objective work best for zero-shot generalization? *arXiv:2204.05832*, 2022.
- [131] Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. VLMO: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv:2111.02358*, 2021.
- [132] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. VATEX: A large-scale, high-quality multilingual dataset for video-and-language research. In *International Conference on Computer Vision*, 2019.
- [133] Yue Wang, Shafiq Joty, Michael Lyu, Irwin King, Caiming Xiong, and Steven Hoi. Vd-bert: A unified vision and dialog transformer with bert. In *Conference on Empirical Methods in Natural Language Processing*, 2020.
- [134] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. SimVLM: Simple visual language model pretraining with weak supervision. *arXiv:2108.10904*, 2021.
- [135] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. *arXiv:2109.01652*, 2021.
- [136] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Ethical and social risks of harm from language models. *arXiv:2112.04359*, 2021.

- [137] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. *arXiv:2203.05482*, 2022.
- [138] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B. Tenenbaum, and Chuang Gan. STAR: A Benchmark for Situated Reasoning in Real-World Videos. In *Conference on Neural Information Processing Systems*, 2021.
- [139] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-QA: Next phase of question-answering to explaining temporal actions. In *IEEE Computer Vision and Pattern Recognition*, 2021.
- [140] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueling Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*, 2017.
- [141] Hanwei Xu, Yujun Chen, Yulun Du, Nan Shao, Yanggang Wang, Haiyu Li, and Zhilin Yang. Ze-roprompt: Scaling prompt-based pretraining to 1,000 tasks improves zero-shot generalization. *arXiv:2201.06910*, 2022.
- [142] Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. VLM: Task-agnostic video-language model pre-training for video understanding. *arXiv:2105.09996*, 2021.
- [143] Ming Yan, Haiyang Xu, Chenliang Li, Junfeng Tian, Bin Bi, Wei Wang, Weihua Chen, Xianzhe Xu, Fan Wang, Zheng Cao, Zhicheng Zhang, Qiyu Zhang, Ji Zhang, Songfang Huang, Fei Huang, Luo Si, and Rong Jin. Achieving human parity on visual question answering. *arXiv:2111.08896*, 2021.
- [144] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. *arXiv:2201.04288*, 2022.
- [145] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *International Conference on Computer Vision*, 2021.
- [146] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of GPT-3 for few-shot knowledge-based VQA. In *National Conference on Artificial Intelligence (AAAI)*, 2021.
- [147] Zhengyuan Yang, Yijuan Lu, Jianfeng Wang, Xi Yin, Dinei Florencio, Lijuan Wang, Cha Zhang, Lei Zhang, and Jiebo Luo. TAP: Text-aware pre-training for text-VQA and text-caption. In *IEEE Computer Vision and Pattern Recognition*, 2021.
- [148] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: Fine-grained interactive language-image pre-training. *arXiv:2111.07783*, 2021.
- [149] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Annual Meeting of the Association for Computational Linguistics*, 2014.

- [150] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision. *arXiv:2111.11432*, 2021.
- [151] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv:2106.10199*, 2021.
- [152] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. MERLOT: Multimodal neural script knowledge models. *Conference on Neural Information Processing Systems*, 2021.
- [153] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. MERLOT reserve: Neural script knowledge through vision and language and sound. In *IEEE Computer Vision and Pattern Recognition*, 2022.
- [154] Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv:2204.00598*, 2022.
- [155] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. *arXiv:2106.04560*, 2021.
- [156] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. LiT: Zero-shot transfer with locked-image text tuning. *arXiv:2111.07991*, 2021.
- [157] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. VinVL: Revisiting visual representations in vision-language models. In *IEEE Computer Vision and Pattern Recognition*, 2021.
- [158] Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. In *IEEE Computer Vision and Pattern Recognition*, 2021.
- [159] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, 2021.
- [160] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *arXiv:2109.01134*, 2021.
- [161] Luwei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In *National Conference on Artificial Intelligence (AAAI)*, 2018.
- [162] Luwei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *National Conference on Artificial Intelligence (AAAI)*, 2020.
- [163] Linchao Zhu and Yi Yang. ActBERT: Learning global-local video-text representations. In *IEEE Computer Vision and Pattern Recognition*, 2020.

- [164] Ron Zhu. Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution. *arXiv:2012.08290*, 2020.
- [165] Xinxin Zhu, Longteng Guo, Peng Yao, Shichen Lu, Wei Liu, and Jing Liu. Vatex video captioning challenge 2020: Multi-view features and hybrid reward strategies for video captioning. *arXiv:1910.11102*, 2019.
- [166] Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Xiaogang Wang, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. Uni-Perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. *arXiv:2112.01522*, 2021.
- [167] Luisa Zintgraf, Kyriacos Shiarli, Vitaly Kurin, Katja Hofmann, and Shimon Whiteson. Fast context adaptation via meta-learning. In *International Conference on Machine Learning*, 2019.

Acknowledgements

We would like to thank many of our colleagues for useful discussions, suggestions, feedback, and advice, including: Relja Arandjelović, Kareem Ayoub, Lorrayne Bennett, Adria Recasens Continente, Tom Eccles, Nando de Freitas, Sander Dieleman, Conor Durkan, Aleksa Gordić, Raia Hadsell, Will Hawkins, Lisa Anne Hendricks, Felix Hill, Jordan Hoffman, Geoffrey Irving, Drew Jaegle, Koray Kavukcuoglu, Agustin Dal Lago, Mateusz Malinowski, Soňa Mokrá, Gaby Pearl, Toby Pohlen, Jack Rae, Laurent Sifre, Francis Song, Maria Tsimpoukelli, Gregory Wayne, and Boxi Wu.

Credit for visual content

- Figure 1:
 - Row 1: All images are provided under license by Unsplash.
 - Row 2: All images are under the public domain.
 - Row 3: First two images are provided under license by Unsplash; third one is courtesy of Brigitte Alayrac.
 - Row 5: Available from DALL·E 2 [98].
 - Row 6: First two are provided under license by Unsplash, the third one is provided by Wikimedia Commons, licensed under CC BY-ND 2.0.
 - Row 7: The images are provided by Wikimedia Commons, licensed under CC BY-ND 2.0.
 - Row 8: The images are provided by Wikimedia Commons, licensed under CC BY-ND 2.0.
 - Row 9: This video is from YFCC100M, licensed under CC BY-ND 2.0.
 - Dialogue 1: Available from DALL·E 2 [98].
 - Dialogue 2: The first icon is provided under license by Flaticon, the second image is provided under license by Unsplash, the third one is provided under license by Sketchfab.
 - Third dialogue: Available from CLIP [94].
- Model Figures 3, 6, 7 and 8: All images are provided under license by Unsplash.
- Qualitative Figures 10, 11, 12, and 13: All visuals are sourced from various sources including the COCO dataset, Wikimedia Commons, licensed under CC BY-ND 2.0 or available from DALL·E 2 [98].

Funding

This research was funded by DeepMind.

8. Appendix

Datasets. We provide more details about our datasets in Appendix A. *M3W* is described in Appendix A.1, including details about its collection in Appendix A.1.1, the data augmentation we employ controlled by the p_{next} parameter in Appendix A.1.2, and its datasheet in Appendix A.1.3. Datasheets for LTIP and VTP are respectively given in Appendix A.2.1 and Appendix A.2.2. The deduplication process for our training datasets against the evaluation datasets is provided in Appendix A.3.

Model card. The *Flamingo* model card is provided in Appendix B.

Experiment details. Additional experiment details are provided in Appendix C. This includes details about our transformer architecture (Appendix C.1), our contrastive pretraining (Appendix C.2), the subsets creation for the `DEV` benchmarks (Appendix C.3) as well as how the gating values evolve during training (Appendix C.4).

Dialogue prompt. Finally we provide the dialogue prompt that is used to generate our dialogue qualitative figures in Appendix D.

A. Datasets

A.1. *M3W*

A.1.1. Collection

The selection and scraping of web pages for *M3W* follows a similar process to that used for collecting the *MassiveWeb* dataset [95]. We start by filtering out non-English documents. We also remove those that do not pass Google’s SafeSearch filter⁵, which identifies explicit content across images, videos, and text. We use a custom scraper to extract salient content from the remaining documents, in the form of plain text interleaved with images, as described in section 3.2.1. This is implemented as an extension to the *MassiveWeb* scraper: the text is collected in a similar fashion, but we also collect any images present at the same level in the HTML tree (if available). We discard documents for which the scraping process does not yield any images.

We then apply similar text filtering heuristics, to remove low quality documents and reduce repetition, as well as some image filters to remove images that are too small (either width or height less than 64 pixels), too wide or narrow (aspect ratio greater than 3 in either direction), or unambiguously low quality (e.g. single-colour images). We discard documents that no longer contain any images following this filtering step.

A.1.2. Image placement data augmentation

During evaluation of Flamingo models, we prompt the model with an image and ask it to generate text for that image. This lends itself to a natural sequencing at inference time in which the image comes before the corresponding text output. When training on single-image (LTIP) or single-video (VTP) datasets (Section 3.2.2) we use this same (*image, text*) sequencing to maximize the model’s effectiveness during inference.

However, the correspondence between images and text in our interleaved M3W dataset (Section 3.2.1) is in general unknown (and potentially not well-defined in certain cases). As a motivating example, a simple webpage might be structured in either of the following ways:

⁵<https://support.google.com/websearch/answer/510>

- | | |
|----------------------------------|------------------------------|
| (a) This is my dog! <dog image> | This is my cat! <cat image> |
| (b) <dog image> That was my dog! | <cat image> That was my cat! |

The text-aligned image indices (`indices`) might “ideally” be chosen such that at each point in the text, the index points to the most semantically relevant image for that text – i.e., the next image in example (a), and the previous image in example (b). In the absence of a general way of determining semantic correspondence between text and images on webpages “in the wild”, we make a simplifying assumption that the most relevant image at any given point in the text is either the last image appearing before the text token, or the image immediately following it (as in the simple examples above), and choose `indices` accordingly.

During training, for each webpage sampled, we sample with probability $p_{next} = \frac{1}{2}$ whether `indices` are chosen to map text to the previous or next image. This inevitably means we make the semantically “unnatural” choice – e.g., associating the text “This is my cat!” with the dog image in (a) above – around half of the time. We ablate this choice in Section 4.4, finding a small advantage to setting $p_{next} = \frac{1}{2}$ over either 0 (always the previous image index) or 1 (always the next image index). This suggests that there may be a beneficial “data augmentation” effect to this randomisation.

A.1.3. Datasheet

We follow the framework defined by Gebru et al. [35] and provide the datasheet for *M3W* in Table 9.

Motivation	
For what purpose was the dataset created? Who created the dataset? Who funded the creation of the dataset?	The dataset was created for pre-training multimodal models of language and images by researchers at DeepMind.
Any other comments?	None.
Composition	
What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?	All instances of the dataset are documents from the web containing interleaved text and images.
How many instances are there in total (of each type, if appropriate)?	There are 43.3M instances (documents) in total, with a total of 185M images and 182 GB of text.
Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?	The dataset is a sample from a larger set.
What data does each instance consist of?	Each instance is made up of a sequence of UTF-8 bytes encoding the document’s text, as well as a sequence of integers indicating the positions of images in the text, and the images themselves in compressed format (see Section 3.2.1).
Is there a label or target associated with each instance?	No, there are no labels associated with each instance.

Is any information missing from individual instances?	No.
Are relationships between individual instances made explicit?	There are no relationships between the different instances in the dataset.
Are there recommended data splits?	We use random splits for the training and development sets.
Are there any errors, sources of noise, or redundancies in the dataset?	There is significant redundancy at the sub-document level.
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.
Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?	The dataset likely contains some data that might be considered offensive, insulting or threatening, as such data is prevalent on the web. We do not try to filter out such content, with the exception of explicit content, which we identify using Google's SafeSearch filter.

Collection Process

How was the data associated with each instance acquired?	The data is available publicly on the web.
What mechanisms or procedures were used to collect the data?	The data was collected using a variety of software programs to extract and clean the raw text and images.
If the dataset is a sample from a larger set, what was the sampling strategy?	We randomly subsample documents.
Who was involved in the data collection process?	Researchers at DeepMind.
Over what timeframe was the data collected?	The dataset was collected over a period of several months in 2021. We do not filter the sources based on creation date.
Were any ethical review processes conducted?	No.

Preprocessing/cleaning/labeling

Was any preprocessing/Cleaning/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)?	Yes — the pre-processing details are discussed in (A.1.1).
--	--

Is the software used to preprocess/-clean/label the instances available?	No.
--	-----

Uses

Has the dataset been used for any tasks already?	Yes, we use the dataset for pre-training multimodal language and vision models.
Is there a repository that links to any or all papers or systems that use the dataset?	No, the dataset has only been used to train the models in this paper.
What (other) tasks could the dataset be used for?	We do not foresee other usages of the dataset at this stage.
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 static and thus will become progressively more “stale”. For example, it will not reflect new language and norms that evolve over time. However, due to the nature of the dataset it is relatively cheap to collect an up-to-date version.
Are there tasks for which the dataset should not be used?	The dataset described in this paper contains English language text almost exclusively and therefore should not be used for training models intended to have multilingual capabilities.

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 9 | **M3W Datasheet**. We follow the framework as presented in Gebru et al. [35].**A.2. Image and video text pair datasets****A.2.1. Datasheet for LTIP****Motivation**

For what purpose was the dataset created? Who created the dataset? Who funded the creation of the dataset?	The dataset was created for pre-training multimodal models of language and images by researchers at DeepMind.
Any other comments?	None.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?	All instances of the dataset are image-text pairs.
How many instances are there in total (of each type, if appropriate)?	The dataset contains 312M image-text pairs.
Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?	The dataset is a sample from a larger set.
What data does each instance consist of?	Each instance is made up of a sequence of UTF-8 bytes encoding the document's text, and an image in compressed format (see Section 3.2.2).
Is there a label or target associated with each instance?	No, there are no labels associated with each instance.
Is any information missing from individual instances?	No.
Are relationships between individual instances made explicit?	There are no relationships between the different instances in the dataset.
Are there recommended data splits?	We use random splits for the training and development sets.
Are there any errors, sources of noise, or redundancies in the dataset?	The data is relatively high quality but there is a chance that some instances are repeated multiple times.
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.
Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?	The websites that were used for this dataset were carefully selected to avoid such content. However given the scale of the data it is possible that some data could be considered offensive or insulting.

Collection Process

How was the data associated with each instance acquired?	The data is available publicly on the web.
What mechanisms or procedures were used to collect the data?	The data was collected using a variety of software programs to extract and clean the raw text and images.
If the dataset is a sample from a larger set, what was the sampling strategy?	N.A.
Who was involved in the data collection process?	Researchers at DeepMind.

Over what timeframe was the data collected?	The dataset was collected over a period of several months in 2021. We do not filter the sources based on creation date.
Were any ethical review processes conducted?	No.

Preprocessing/cleaning/labeling

Was any preprocessing/Cleaning/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)?	Some automatic text formatting was applied to remove from the captions dates and locations that were not relevant to the training objective.
Is the software used to preprocess/-clean/label the instances available?	No.

Uses

Has the dataset been used for any tasks already?	Yes, we use the dataset for pre-training multimodal language and vision models.
Is there a repository that links to any or all papers or systems that use the dataset?	No, the dataset has only been used to train the models in this paper.
What (other) tasks could the dataset be used for?	We do not foresee other usages of the dataset at this stage.
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 static and thus will become progressively more “stale”. For example, it will not reflect new language and norms that evolve over time. However, due to the nature of the dataset it is relatively cheap to collect an up-to-date version.
Are there tasks for which the dataset should not be used?	The dataset described in this paper contains English language text almost exclusively and therefore should not be used for training models intended to have multilingual capabilities.

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 10 | **LTIP Datasheet**. We follow the framework as presented in Gebru et al. [35].

A.2.2. Datasheet for VTP

Motivation

For what purpose was the dataset created? Who created the dataset? Who funded the creation of the dataset?	The dataset was created for pre-training multimodal models of language and images by researchers at DeepMind.
Any other comments?	None.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?	All instances of the dataset are video-text pairs.
How many instances are there in total (of each type, if appropriate)?	The dataset contains 27M video-text pairs.
Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?	The dataset is a sample from a larger set.
What data does each instance consist of?	Each instance is made up of a sequence of UTF-8 bytes encoding the document's text, and a video in compressed format (see Section 3.2.2).
Is there a label or target associated with each instance?	No, there are no labels associated with each instance.
Is any information missing from individual instances?	No.
Are relationships between individual instances made explicit?	There are no relationships between the different instances in the dataset.
Are there recommended data splits?	We use random splits for the training and development sets.
Are there any errors, sources of noise, or redundancies in the dataset?	The data is relatively high quality but there is a chance that some instances are repeated multiple times.
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.
Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?	The websites that were used for this dataset were carefully selected to avoid such content. However given the scale of the data it is possible that some data could be considered offensive or insulting.

Collection Process

How was the data associated with each instance acquired?	The data is available publicly on the web.
What mechanisms or procedures were used to collect the data?	The data was collected using a variety of software programs to extract and clean the raw text and videos.
If the dataset is a sample from a larger set, what was the sampling strategy?	N.A.
Who was involved in the data collection process?	Researchers at DeepMind.
Over what timeframe was the data collected?	The dataset was collected over a period of several months in 2021. We do not filter the sources based on creation date.
Were any ethical review processes conducted?	No.

Preprocessing/cleaning/labeling

Was any preprocessing/Cleaning/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)?	Some automatic text formatting was applied to remove from the captions dates and locations that were not relevant to the training objective.
Is the software used to preprocess/-clean/label the instances available?	No.

Uses

Has the dataset been used for any tasks already?	Yes, we use the dataset for pre-training multimodal language and vision models.
Is there a repository that links to any or all papers or systems that use the dataset?	No, the dataset has only been used to train the models in this paper.
What (other) tasks could the dataset be used for?	We do not foresee other usages of the dataset at this stage.
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 static and thus will become progressively more “stale”. For example, it will not reflect new language and norms that evolve over time. However, due to the nature of the dataset it is relatively cheap to collect an up-to-date version.
Are there tasks for which the dataset should not be used?	The dataset described in this paper contains English language text almost exclusively and therefore should not be used for training models intended to have multilingual capabilities.

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 11 | **VTP Datasheet**. We follow the framework as presented in Gebru et al. [35].

A.3. Dataset deduplication against evaluation tasks

We used a Google internal deduplication tool to deduplicate our training datasets. This deduplication pipeline relies on a trained visual encoder which maps embedding closer together when they are potential duplicates. Once the image embeddings have been computed, a fast approximate nearest neighbor search⁶ is performed on the training images to retrieve duplicates candidates from the validation datasets. For the paired image-text dataset, we have deduplicated our LTIP and ALIGN training images against: ImageNet (train, val), COCO (train, valid, test), OK-VQA (train, valid, test), VQAv2 (train, valid, test), Flickr30k (train, valid, test), VisDial (train, valid, test).

We did not deduplicate our image datasets against VizWiz, HatefulMemes and TextVQA as we performed these evaluation only after having trained our Flamingo models. However, we believe this had no impact on our results as the images from these datasets are unlikely to be scraped from the web; VizWiz images were obtained using a specific mobile app and only available for download, HatefulMemes memes were created by researchers instead of being scraped on the web and finally TextVQA images are from OpenImages.

Note that we did not run the deduplication on the *M3W* dataset as one training example is a full webpage of interleaved paragraph with several images, unlikely to contain images from our benchmark suite. To verify this hypothesis, we have obtained near-duplicates statistics on the 185M individual images from *M3W* and the results are the following: 1314 potential duplicates were found from the validation and test splits of ImageNet, COCO, OK-VQA, VQAv2, Flickr30k and VisDial. Out of the 1314 candidates, only 125 are exact duplicates.

For the video datasets, we did not perform any deduplication of VTP (27M videos) as none of the collected VTP videos were obtained from YouTube or Flickr, which are the source of all of our video evaluation datasets collected on the Internet.

B. Flamingo Model Card

We present a model card for Flamingo in Table 12, following the framework presented by Mitchell et al. [85].

Model Details

Organization Developing the Model	DeepMind
Model Date	March 2022

⁶<https://github.com/google-research/google-research/tree/master/scann>

Model Type	Transformer-based autoregressive language model, conditioned on visual features from a convnet-based encoder. Additional transformer-based cross-attention layers incorporate vision features into the language model’s text predictions. (See Section 3 for details.)
Feedback on the Model	jalayrac@google.com

Intended Uses

Primary Intended Uses	The primary use is research on visual language models (VLM), including: research on VLM applications like classification, captioning or visual question answering, understanding how strong VLMs can contribute to AGI, advancing fairness and safety research in the area of multimodal research, and understanding limitations of current large VLMs.
Primary Intended Users	DeepMind researchers. We will not make this model available publicly.
Out-of-Scope Uses	Uses of the model for visually conditioned language generation in harmful or deceitful settings. Broadly speaking, the model should not be used for downstream applications without further safety and fairness mitigations specific to each application.

Factors

Card Prompts – Relevant Factor	Relevant factors include which language is used. Our model is trained on English data. Our model is designed for research. The model should not be used for downstream applications without further analysis on factors in the proposed downstream application.
Card Prompts – Evaluation Factors	<i>Flamingo</i> is based on Chinchilla (a large proportion of the weights of Chinchilla are used as this) and we refer to the analysis provided in [48, 95] for the language only component of this work. We refer to our study presented in Section 6.2.2 for a toxicity analysis when the model is conditioned on an image.

Metrics

Model Performance Measures	We principally focus on the model’s ability to predict relevant language when given an image. For that we used a total of 18 different benchmarks described in Section 4.1.2 spanning various vision and language tasks such as classification (ImageNet, Kinetics700, HatefulMemes), image and video captioning (COCO, VATEX, Flickr30K, YouCook2, RareAct), visual question answering (OKVQA, VizWiz, TextVQA, VQAv2, MSRVTTQA, MSVDQA, iVQA, STAR, NextQA) and visual dialog (VisDiag). This was tested either in an open ended setting where <i>Flamingo</i> generate language and we compare the outputs with the ground truth or in a close ended setting where we directly score various outcomes using the likelihood of the model.
Decision thresholds	N/A
Approaches to Uncertainty and Variability	Due to the costs of training <i>Flamingo</i> , we cannot train it multiple times. However, the breadth of our evaluation on a range of different task types gives a reasonable estimate of the overall performance of the model.

Evaluation Data

Datasets	See Table 2 for a detailed list.
Motivation	We chose our evaluation datasets to span an important range of vision and language tasks to correctly assess the ability of <i>Flamingo</i> to produce relevant text given an image.
Preprocessing	Input text is tokenized using a SentencePiece tokenizer with a vocabulary size of 32,000. Images are processed so that their mean and variance are 0 and 1 respectively.

Training Data

See [56], the Datasheet in Appendix A.1.3, Appendix A.2.1, Appendix A.2.2

Quantitative Analyses

Unitary Results	<i>Flamingo</i> sets a new state of the art in few-shot learning on a wide range of open-ended vision and language tasks. On the 16 tasks we consider, Flamingo also surpasses the fine-tuned state-of-art in 6 of the cases despite using orders of magnitude less task-specific training data. We refer to Section 4 for the full details of our quantitative study.
Intersectional Results	We did not investigate intersectional biases.

Ethical Considerations

Data	The data is sourced from a variety of sources, some of it from web content. Sexually explicit content is filtered out, but the dataset does include racist, sexist or otherwise harmful content.
Human Life	The model is not intended to inform decisions about matters central to human life or flourishing.
Mitigations	Apart from removing sexual explicit content we did not filter out toxic content, following the rationale of Rae et al. [95]. More work is needed on mitigation approaches to toxic content and other types of risks associated with language models, such as those discussed in Weidinger et al. [136].
Risks and Harms	The data is collected from the internet, and thus undoubtedly toxic and biased content is included in our training dataset. Furthermore, it is likely that personal information is also in the dataset that has been used to train our models. We defer to the more detailed discussion in Weidinger et al. [136].
Use Cases	Especially fraught use cases include the generation of factually incorrect information with the intent of distributing it or using the model to generate racist, sexist or otherwise toxic text with harmful intent. Many more use cases that could cause harm exist. Such applications to malicious use are discussed in detail in Weidinger et al. [136].

Table 12 | **Flamingo Model Card.** We follow the framework presented in Mitchell et al. [85].

	Resampler				xattn dense				Frozen LM			
	L	D	H	Act.	L	D	H	Act.	L	D	H	Act.
Flamingo-3B	6	1536	16	Sq. ReLU	24	2048	16	Sq. ReLU	24	2048	16	GeLU
Flamingo-9B	6	1536	16	Sq. ReLU	10	4096	32	Sq. ReLU	40	4096	32	GeLU
Flamingo	6	1536	16	Sq. ReLU	12	8192	64	Sq. ReLU	80	8192	64	GeLU

Table 13 | Flamingo models transformer hyper-parameters. All transformer blocks have a same constant key and value size of 128 and the hidden size of each MLP is $4 \times D$. L: number of layers, D: transformer hidden size, H: number of heads, Act.: FFW activation, Sq. ReLU: Squared ReLU [112].

C. Additional experiment details

C.1. Transformer architecture details.

We list in Table 13, the number of layers (L), the hidden dimension (D), the number of head (H) as well as the FFW activation (Act.) used for each transformer part of our Flamingo models. All transformer blocks have a same constant key and value size of 128 and the hidden size of each MLP is $4 \times D$. Note that the frozen LM was trained with the GeLU [45] activation while we have noticed that

training the remaining trainable transformers with the Squared ReLU [112] activation outperforms GeLU.

Dataset	Combination strategy	ImageNet accuracy top-1	COCO					
			image-to-text			text-to-image		
			R@1	R@5	R@10	R@1	R@5	R@10
LTIP	None	40.8	38.6	66.4	76.4	31.1	57.4	68.4
ALIGN	None	35.2	32.2	58.9	70.6	23.7	47.7	59.4
LTIP + ALIGN	Accumulation	45.6	42.3	68.3	78.4	31.5	58.3	69.0
LTIP + ALIGN	Data merged	38.6	36.9	65.8	76.5	15.2	40.8	55.7
LTIP + ALIGN	Round-robin	41.2	40.1	66.7	77.6	29.2	55.1	66.6

Table 14 | **Effect of contrastive pretraining datasets and combination strategies.** The first two rows show the effect of training a small model on LTIP and ALIGN only, the final three show the results of a small model trained on combinations of these datasets, comparing different combination strategies.

C.2. Contrastive model details

C.2.1. Contrastive model training details.

To pretrain the vision encoder we use a contrastive loss on the ALIGN and LTIP datasets, similar to CLIP [94] and ALIGN [56]. The vision encoder is trained from scratch, together with a language encoder. Using these encoders, images and text pairs are separately encoded and projected to a shared embedding space and L2 normalized. From these embeddings, we maximise the similarity of paired embeddings and minimize the similarity of unpaired embeddings, using a multi-class cross-entropy loss, where the paired image-texts are treated as positive examples and the rest of the batch as negative examples. We use the same loss as in CLIP [94], which consists of two contrastive losses, one from text to image and another one from image to text. We use a learnable temperature parameter in the final log-softmax [10] layer. The result is two losses, one from texts-to-images:

$$L_{\text{contrastive:im2txt}} = -\frac{1}{N} \sum_i^N \log \left(\frac{\exp(L_i^\top V_i \beta)}{\sum_j^N \exp(L_i^\top V_j \beta)} \right) \quad (3)$$

and a similar one for images-to-text:

$$L_{\text{contrastive:txt2im}} = -\frac{1}{N} \sum_i^N \log \left(\frac{\exp(V_i^\top L_i \beta)}{\sum_j^N \exp(V_i^\top L_j \beta)} \right) \quad (4)$$

the sum of which is optimized. Here, V_i and L_i are, respectively, the normalized embedding of the vision and language component of the i -th element of a batch. β is a trainable inverse temperature parameter and N is the number of elements in the batch. We use the BERT [25] architecture for the language encoder. The outputs of the language and vision encoders are mean-pooled before being projected to the shared embedding space. We only use the weights from the contrastive vision encoder in the main Flamingo model. Our pretrained experiments are discussed in more detail in Section 4.1.4.

Experiment details. The training image resolution is 288×288 , the joint embedding space is size 1376 and the batch size is 16,384. It is trained for 1.2 million parameter update steps, each of which consist of two gradient calculation steps (more details below) on 512 TPUv4 chips. The learning rate is decayed linearly from 10^{-3} to zero over the course of training. Images have random color

augmentation and horizontal flips applied during training. We use the same tokenizer as [56]. The Adam optimizer is used to optimize the network, and we apply label smoothing of 0.1. We apply 10^{-2} adaptive gradient clipping (AGC) [11] to the NFNet encoder and global norm gradient clipping of 10 for the BERT encoder.

To evaluate the pretrained model, we track zero shot image classification and retrieval. For zero shot image classification, we use image-text retrieval between the images and the class names. Following [94] we use “prompt-emsembling” in which we embed multiple texts using templates such as “A photo of a {class_name}” and average the resulting embedding.

C.2.2. Ablation study on different dataset mixing strategies for the contrastive pretraining

One key to achieving strong results was the inclusion of our new dataset LTIP alongside ALIGN for training. Despite being a smaller dataset ALIGN by a factor of 6, a contrastive model trained on only LTIP outperforms one trained only on ALIGN on our evaluation metrics, suggesting that dataset quality may be more important than scale in the regimes in which we operate. We also find that a model trained on both ALIGN and LTIP outperforms those trained on the two datasets individually and that how the datasets are combined is important.

To demonstrate this, we train a small model with an NFNet-F0 vision encoder, BERT-mini language encoder and batch size 2048 for 1 million gradient-calculation steps on ALIGN, LTIP and a mixture of the two. The results are presented in Table 14. It shows the results of training models on the combined datasets using three different merging regimes:

- Data merged: Batches are constructed by merging examples from each dataset into one batch.
- Round-robin: We alternate batches of each dataset, updating the parameters for each batch.
- Accumulation: We compute a gradient on a batch from each dataset. These gradients are then weighted and summed and use to update the parameters.

Across all evaluation metrics, we find that the Accumulation method outperforms other methods of combining the datasets. Although the LTIP dataset is $5 \times$ smaller than the ALIGN dataset, this ablation study suggests that the quality of the training data can be more important than its abundance.

C.3. Details of subsets used for the `DEV` benchmarks

Here, we describe in more detail how we create the *validation* subsets for each benchmark. For captioning tasks, open-ended evaluation is efficient so we evaluate on a large number of samples. Specifically, for COCO, we use the same number of samples as used in the splits of Karpathy for evaluation sets (5000). For VATEX, because the training set is of limited size, we only evaluate over 1024 samples, reserving the rest for support sets. For question answering tasks, we evaluate over 1024 samples; chosen to make both open- and close-ended evaluation reasonably fast. For image classification tasks, we evaluate over 10 images per class, that is 10000 samples for ImageNet, and 7000 samples for Kinetics700. As for the support sets, for both validation and final performance estimation, we use 2048 samples across all tasks, except for classification tasks where we scale this to 32 samples per class, to better estimate expected performance for each class.

C.4. Evolution of the tanh gating values throughout training.

We plot in Figure 14, the evolution of the absolute value of the tanh gating values as a function of the training progress (from 0% to 100%) at different layers of the LM stack for the *Flamingo-3B* model

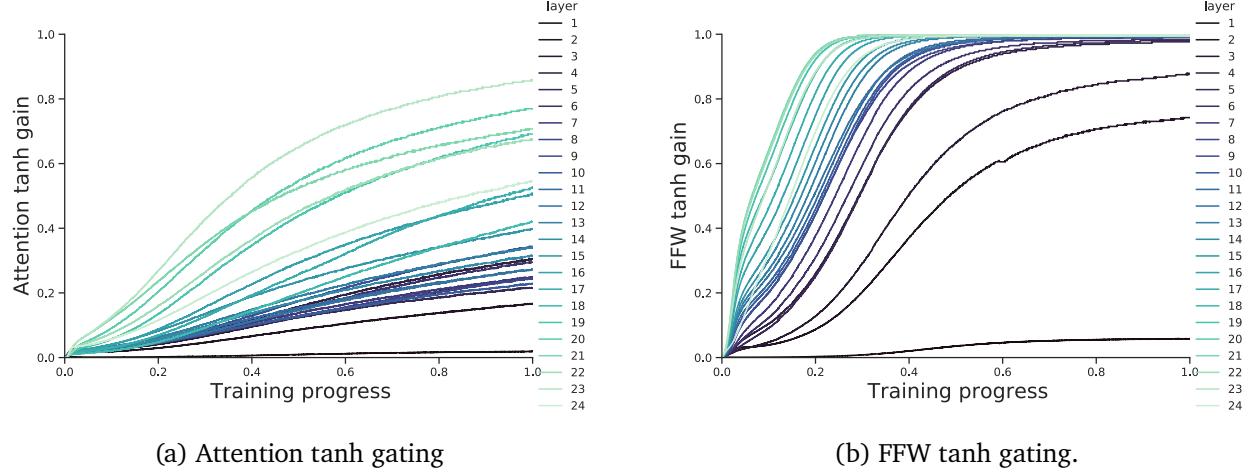


Figure 14 | Evolution of the absolute value of the tanh gating at different layers of *Flamingo-3B*.

composed of 24 LM layers. All layers of the frozen LM stack seem to utilize the visual information as the tanh gating absolute values quickly go away from the 0 initializations. We also note that the absolute values seem to grow with the depth. It is however difficult to draw strong conclusions from this observation as the scale of activations might also vary with the depth even though layer normalization is present. Future work is required to better understand the effect of these added layers on the optimization dynamic and on the language model itself.

D. Dialogue prompt

For the dialogue examples in Figure 11 the following prompt is used:

This is a conversation between a human, User, and an intelligent visual AI, Flamingo. User sends images, and Flamingo describes them.

User: <a cat image>

Flamingo: That is a cat. It's a tiny kitten with really cute big ears.

User: <a dinner image>

Flamingo: This is a picture of a group of people having dinner. They are having a great time!

User: Can you guess what are they celebrating?

Flamingo: They might be celebrating the end of a successful project or maybe a birthday?

User: <a graph image>

Flamingo: This is a graph, it looks like a cumulative density function graph.