


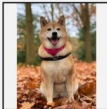







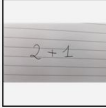
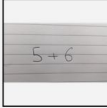
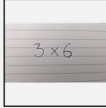
# Flamingo: a Visual Language Model for Few-Shot Learning

## What is Flamingo?

A multimodal model that combines vision and language for few-shot learning, enabling robust performance across diverse tasks without task-specific fine-tuning.

## Motivation:

- Traditional visual models require task specific fine tuning and large annotated datasets.
- Flamingo addresses this by enabling few shot learning across image, video and text tasks

Input Prompt					Completion	
	This is a chinchilla. They are mainly found in Chile.		This is a shiba. They are very popular in Japan.		This is	a flamingo. They are found in the Caribbean and South America.
	What is the title of this painting? Answer: The Hallucinogenic Toreador.		Where is this painting displayed? Answer: Louvres Museum, Paris.		What is the name of the city where this was painted? Answer:	Arles.
	Output: "Underground"		Output: "Congress"		Output:	"Soulomes"
	2+1=3		5+6=11			3x6=18

## References/Image credits

J-B. Alayrac et al., "Flamingo: a Visual Language Model for Few-Shot Learning", arxiv (2022)

# Bridge powerful pretrained vision-only and language-only models

- Training from scratch is extremely **computationally expensive**
- Start from a **pretrained** language model
- Text-only model has no built-in way to incorporate input from other modalities.

**Flamingo:** interleave **cross-attention** layers with language-only self-attention layers (frozen)

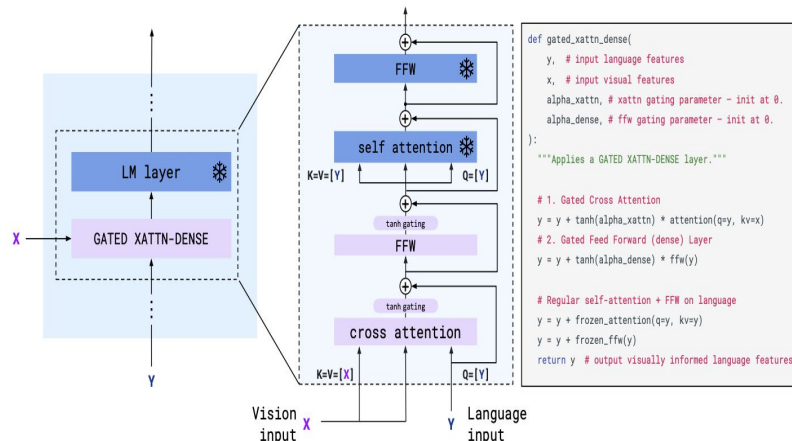


Figure 4: **GATED XATTN-DENSE layers.** To condition the LM on visual inputs, we insert new cross-attention layers between existing pretrained and frozen LM layers. The keys and values in these layers are obtained from the vision features while the queries are derived from the language inputs. They are followed by dense feed-forward layers. These layers are *gated* so that the LM is kept intact at initialization for improved stability and performance.

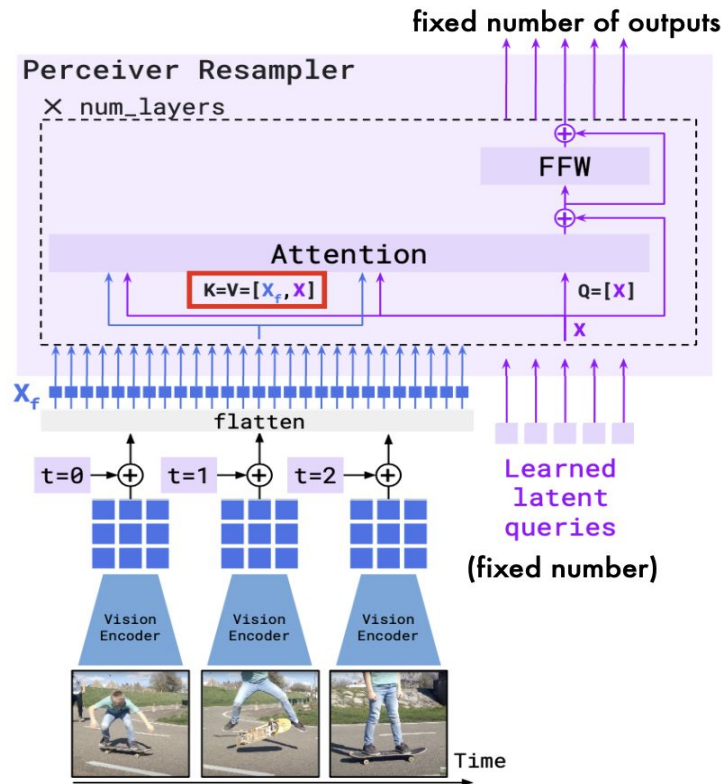
## References/Image credits

J-B. Alayrac et al., "Flamingo: a Visual Language Model for Few-Shot Learning", arxiv (2022)

# Unified handling of images and videos

- Both image and video inputs are high dimensional
- Flattening these inputs into 1-D sequences as is done in text generation, is computationally expensive
- It is exacerbated by the quadratic cost of self-attention layers

**Flamingo:** Uses a **perceiver-resampler** processes inputs into fixed visual tokens, incorporates temporal encodings for videos and combines learned latent queries with visual embeddings through attention.



# Training Dataset

- The existing (image, text) datasets used by CLIP or ALIGN are not general enough for few shot learning
- Multimodal datasets are scarce in comparison to abundantly available text-only data
- One approach is to scrape web pages for interleaved images and text however these pairs are often only weakly related
- Flamingo: combines web scraping with existing paired (image, text) and (video, text) datasets



This is an image of a flamingo.

Image-Text Pairs dataset

[ $N=1$ ,  $T=1$ , H, W, C]



A kid doing a kickflip.

Video-Text Pairs dataset

[ $N=1$ ,  $T>1$ , H, W, C]

Welcome to my website!



This is a picture of my dog.



This is a picture of my cat.

Multi-Modal Massive Web (M3W) dataset

[ $N>1$ ,  $T=1$ , H, W, C]

# Flamingo Architecture

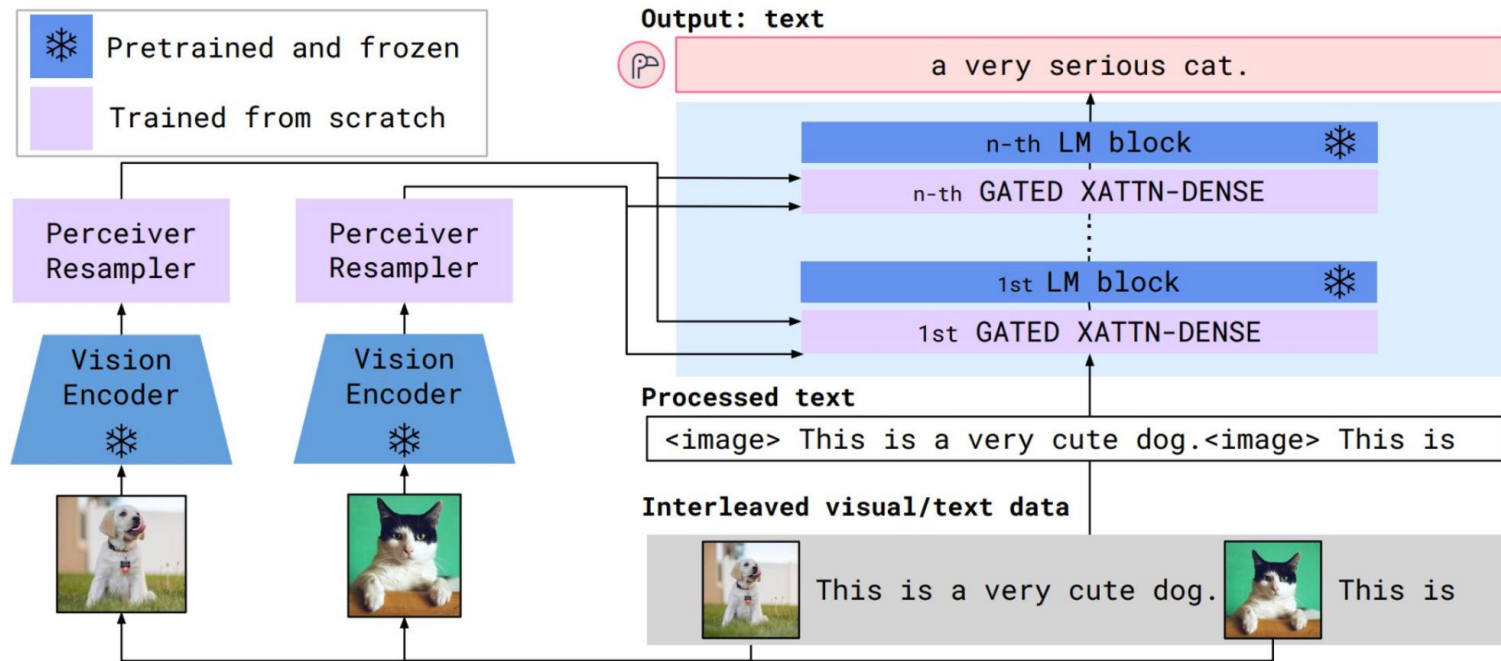


Figure 3: **Flamingo architecture overview.** Flamingo is a family of visual language models (VLMs) that take as input visual data interleaved with text and produce free-form text as output.

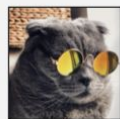
## References/Image credits

J-B. Alayrac et al., "Flamingo: a Visual Language Model for Few-Shot Learning", arxiv (2022)

## Vision to Text tasks (input=vision, output=text)

Support examples

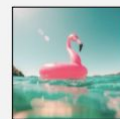
Query



A cat wearing  
sunglasses.



Elephants  
walking in  
the savanna.



<BOS><image>Output: A cat wearing sunglasses.<EOC><image>Output: Elephants walking in the savanna.<EOC><image>Output:

Processed prompt

## Visual Question Answering Task (input=vision+text, output=text)

Support examples

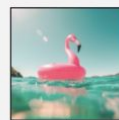
Query



What's  
the cat  
wearing? sunglasses



How many  
animals? 3

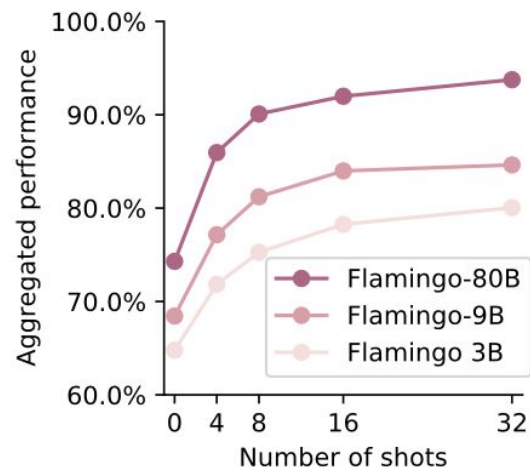
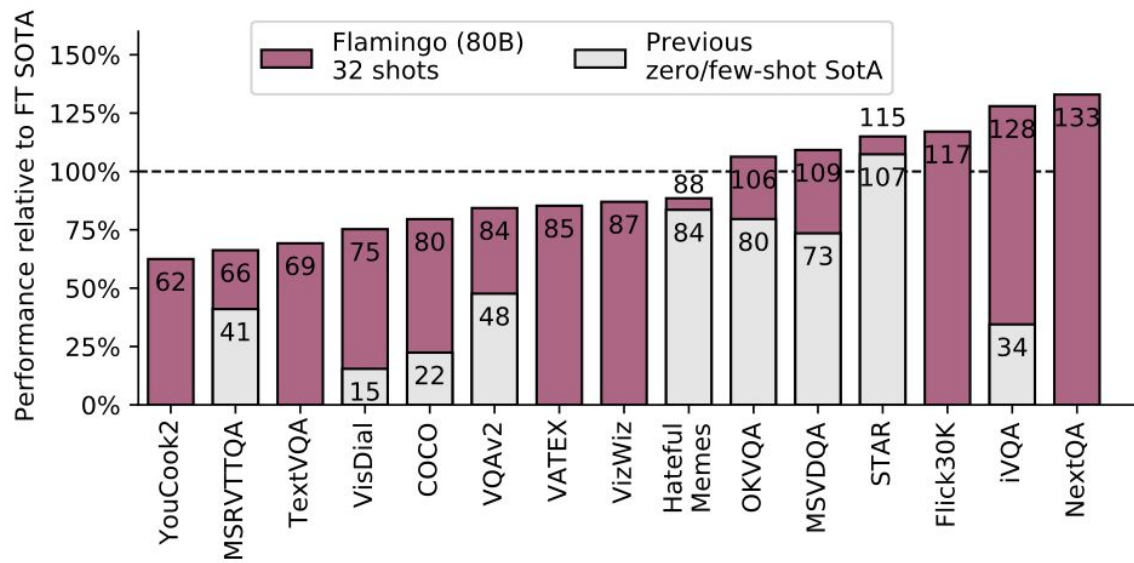


What is on  
the water?

<BOS><image>Question: What's the cat wearing? Answer: sunglasses<EOC><image>Question: How many animals? Answer: 3<EOC><image>  
Question: What is on the water? Answer:

Processed prompt

# Results





# With Finetuning

- During fine-tuning, Flamingo keeps the language model layers frozen
- The base vision encoder is also fine-tuned (unlike Flamingo pretraining)
- Hyperparameters are set by grid search on validation subsets of the training sets
- Search over: learning rate, decay schedule, training steps, batch size, augmentation

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
🔗 <i>Flamingo</i> - 32 shots	67.6	-	113.8	65.1	49.8	-	31.0	56.8	-	86.8	36.0	-	70.0
SimVLM [124]	80.0	80.3	<b>143.3</b>	-	-	-	-	-	-	-	-	-	-
OFA [119]	79.9	80.0	<u>149.6</u>	-	-	-	-	-	-	-	-	-	-
Florence [140]	80.2	80.4	-	-	-	-	-	-	-	-	-	-	-
🔗 <i>Flamingo</i> Fine-tuned	<b>82.0</b>	<b>82.1</b>	138.1	<b>84.2</b>	<u>65.7</u>	<b>65.4</b>	<b>47.4</b>	61.8	59.7	118.6	<b>57.1</b>	54.1	<b>86.6</b>
Restricted SotA <sup>†</sup>	80.2	80.4	<b>143.3</b>	76.3	-	-	46.8	<b>75.2</b>	<b>74.5</b>	<b>138.7</b>	54.7	<b>73.7</b>	79.1
	[140]	[140]	[124]	[153]	-	-	[51]	[79]	[79]	[132]	[137]	[84]	[62]
Unrestricted SotA	81.3	81.3	<u>149.6</u>	81.4	57.2	60.6	-	-	75.4	-	-	-	84.6
	[133]	[133]	[119]	[153]	[65]	[65]	-	-	[123]	-	-	-	[152]