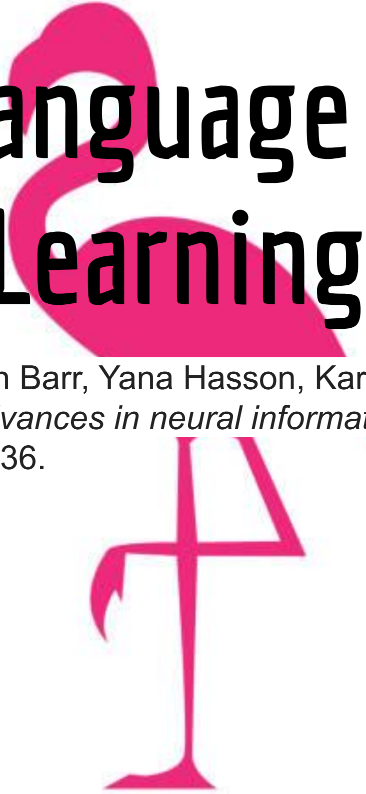
Flamingo: a Visual Language Model for Few-Shot Learning



Alayrac, Jean-Baptiste, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc et al. "Flamingo: a visual language model for few-shot learning." *Advances in neural information processing systems* 35 (2022): 23716-23736.

Presented by:

Qi Zhao, Sumin Lee

Outline

- Research background & Motivation, Challenges
- Key ideas
- Flamingo model & architecture
- Training scheme
- Comparison to state-of-the-art
- Demo
- Discussion (limitation, future, trade-off, benefits)

Research background & Motivation

The few-shot dream

‘One key aspect of intelligence is the ability to quickly learn to perform a new task given a **short instruction**.’

Computer vision:

large-scale pretraining + **task-specific fine-tuning**

But, current fine-tuning requires:

- thousands of training samples
- careful per-task hyperparameter tuning
- resource intensive

Research background & Motivation

Task abilities

Multimodal models (e.g. **Clip**) has shown promising zero-shot performance, but it is inflexible and lacks the ability to generate language.

Flexible models: visually-conditioned language generation (e.g. **VL-T5**) have not demonstrated strong few-shot performance.

Inspired from NLP: large language models (LLM) like **GPT-3** are flexible few shot learners: given a few examples of a task and a new query as input, the LLM generates a continuation to produce a predicted output.

Key factor of their success: **large-scale pretraining**

Can we learn a model capable of open-ended multimodal tasks via pretraining?

Challenges & Approaches

- Training large language models is extremely computationally expensive.
    - To save compute resources, starting from a pre-trained language model.
    - But a text-only model has no build-in ability to take inputs from other modalities

Proposed approach:

Interleave cross-attention layers with frozen pre-trained language self-attention layers

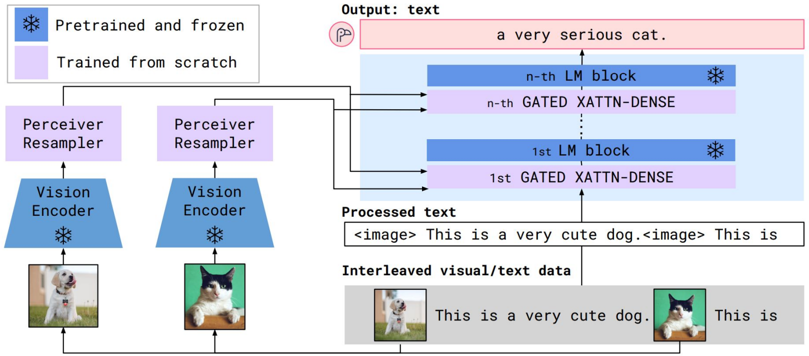
- Images and videos are in high-dimensions, flattening them into 1D sequences is infeasible.
    - Quadratic cost of self-attention makes it worse

Proposed approach:

Perceiver-based architecture with an output of a fixed number of visual tokens

Flamingo: Key Ideas

Flamingo is Visual Language Model (VLM) that accepts interleaved inputs (text + images + videos) and produces free-form text in close/open-ended tasks with few shot prompting.



Key ideas:

- Leverage **pretrained** models to save compute resources

Vision

Language

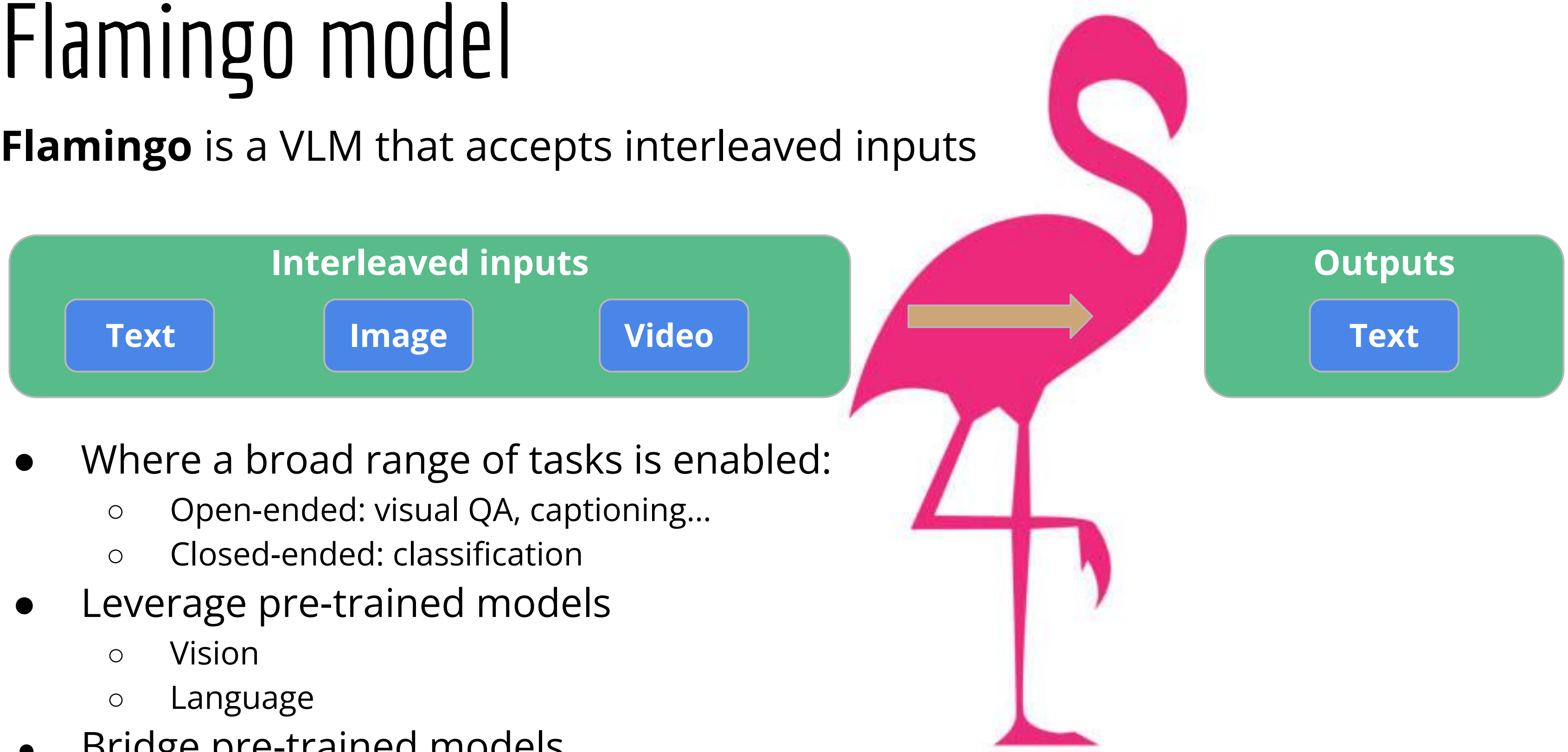
- **Bridge** pretrained models harmoniously

Perceiver Resampler

cross-attention

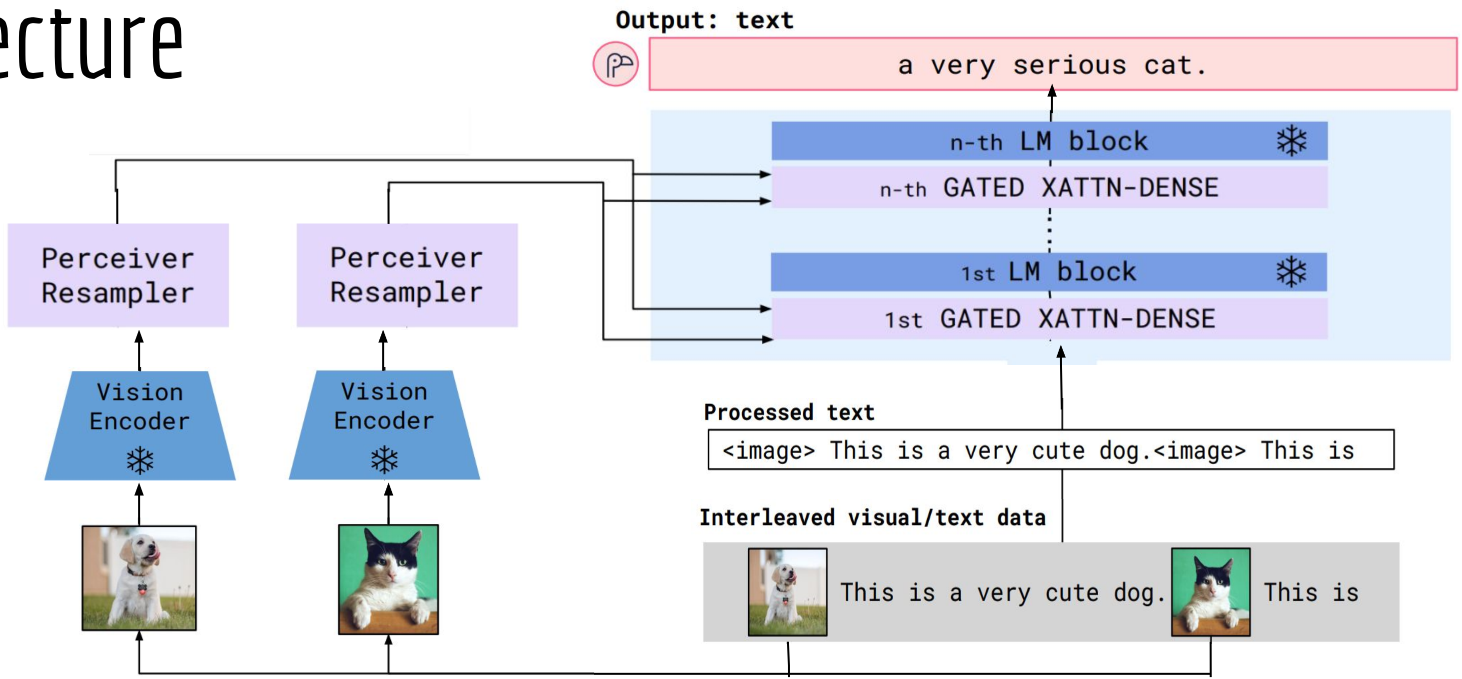
Flamingo model

Flamingo is a VLM that accepts interleaved inputs



- Where a broad range of tasks is enabled:
  - Open-ended: visual QA, captioning...
  - Closed-ended: classification
- Leverage pre-trained models
  - Vision
  - Language
- Bridge pre-trained models
  - Perceiver sampler
  - Cross-attention layers

Architecture

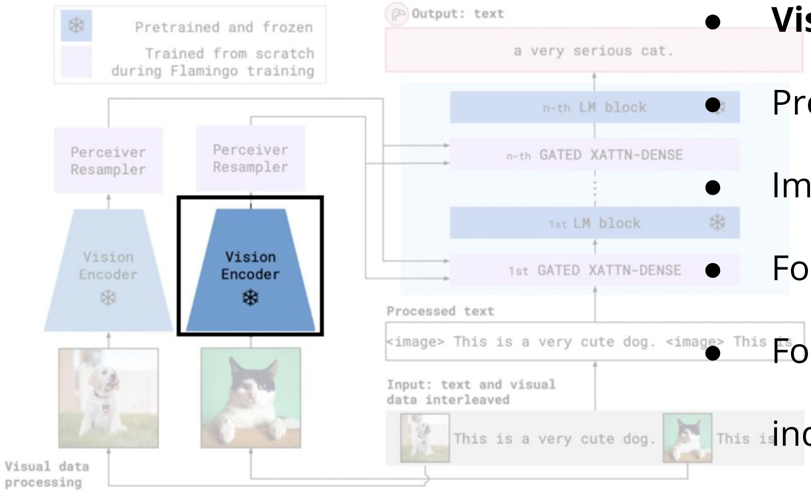


Multimodal likelihood

Flamingo can model the likelihood of text *y* interleaved with a sequence of images/videos *x*:

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

Architecture: vision encoder - pixels to features



- **Vision encoder**: F6 Normalizer-free ResNet (**NFNet**) backbone.
- Pre-trained with **BERT** as dual encoder using **contrastive** loss.
- Image resolution: 288 x 288; Embedding size: 1376
- For images: outputs **2D** spatial grid which is then flatten into **1D**
- For videos: frames are sampled **1FPS** and encoded independently into **3D** grid to which **temporal embedding** are added, then flatten into 1D

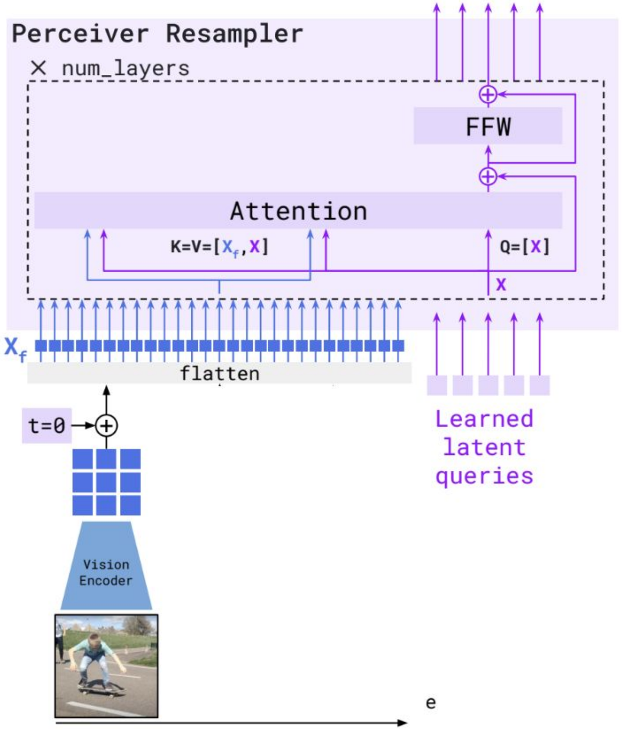
Vision encoder is frozen after pretraining, text encoder is discarded.

Perceiver resampler from varying-size large feature maps to few visual tokens

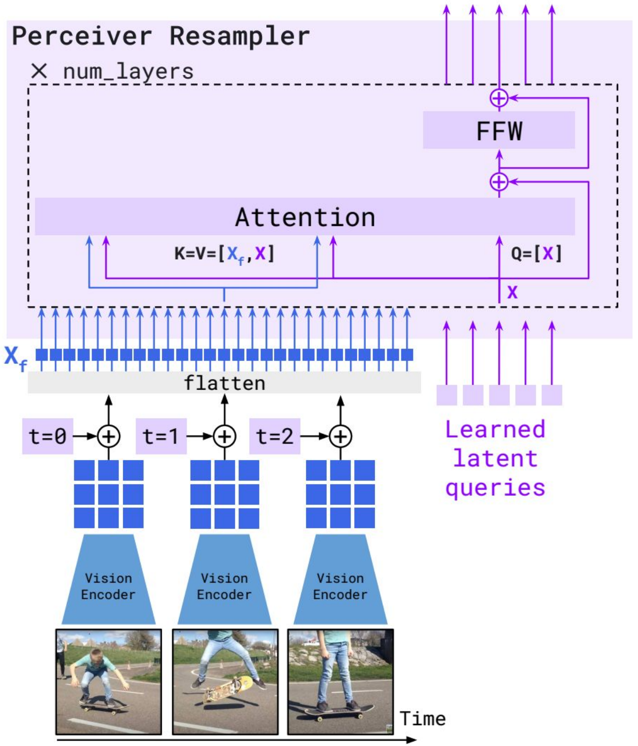
Transform a potentially large and variable size visual features into a smaller fixed number of output tokens

- Encoder provides inputs
- Outputs a fixed number of tokens.
- Temporal encoding
- Learned latent queries
- Attention and FFW layers.

Image

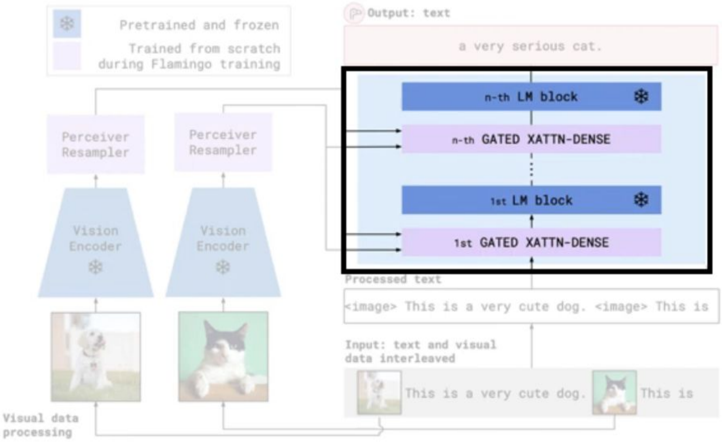


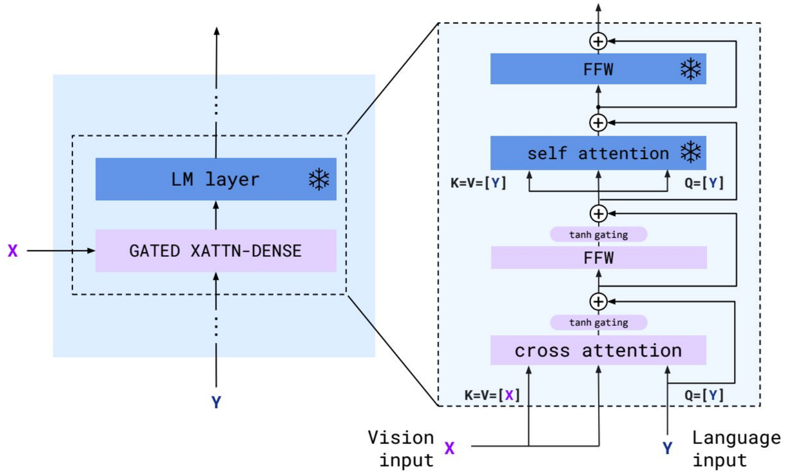
Video



Conditioning the Language model

- Language model: frozen Chinchillas (trained on massive text).
- **Gated xattn dense** blocks (trained from scratch) are inserted between the language model layers
- Each block includes **cross attention, feed-forward**





Training scheme -data

Flamingo is trained on:

- **Image-text** pair
- **Video-text** pair
- **Webpage** data

Image-Text Pairs dataset
[N=1, T=1, H, W, C]

Video-Text Pairs dataset
[N=1, T>1, H, W, C]

Multi-Modal Massive Web (M3W) dataset
[N>1, T=1, H, W, C]

ALIGN dataset & LTIP dataset

- Resolution 320 x 320 pixels
- Text tokens = 32 / 64

VTP dataset

- Resolution 320 x 320 pixels
- Temporal dim = 8
- Text tokens = 32

M3W dataset

- Extract text&images from 43 million webpages
- Resolution 320 x 320 pixels
- Text tokens = 256

Training scheme - Objective

Models are trained with a weighted sum of per-dataset expected negative log-likelihood of text, conditioned on visual inputs.

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

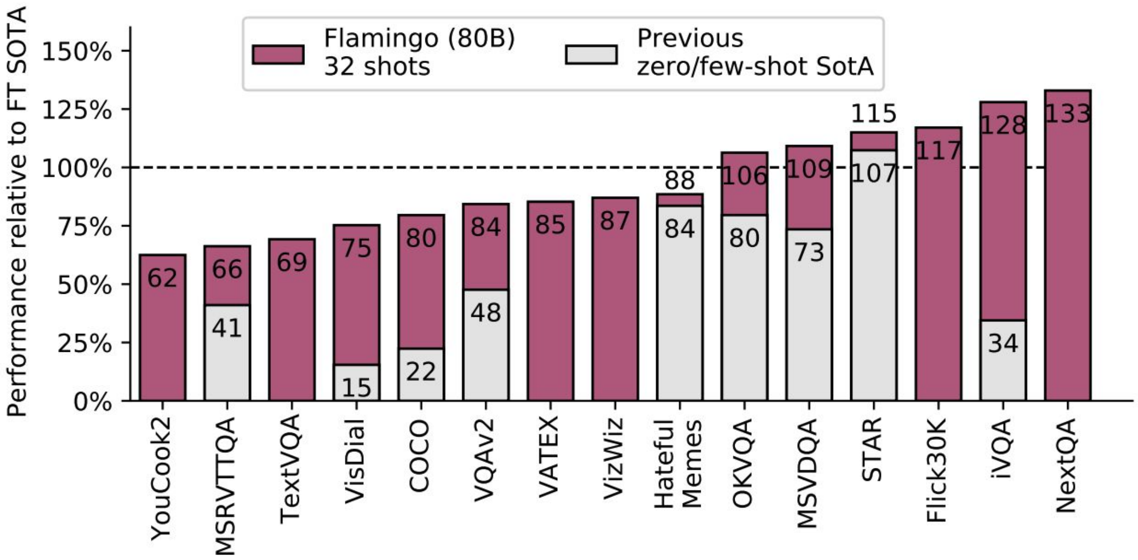
$\mathcal{D}_m$: m-th dataset

λ_m : positive scalar weight for the m-th dataset

Comparison to state-of-the-art

- Outperform SOTA fine-tuned model on 6 out of 16 tasks.

- In all tasks that has published few-shot result, Flamingo sets the new SOTA



Demo

1. Interleaved embedding visualization (link to Colab)

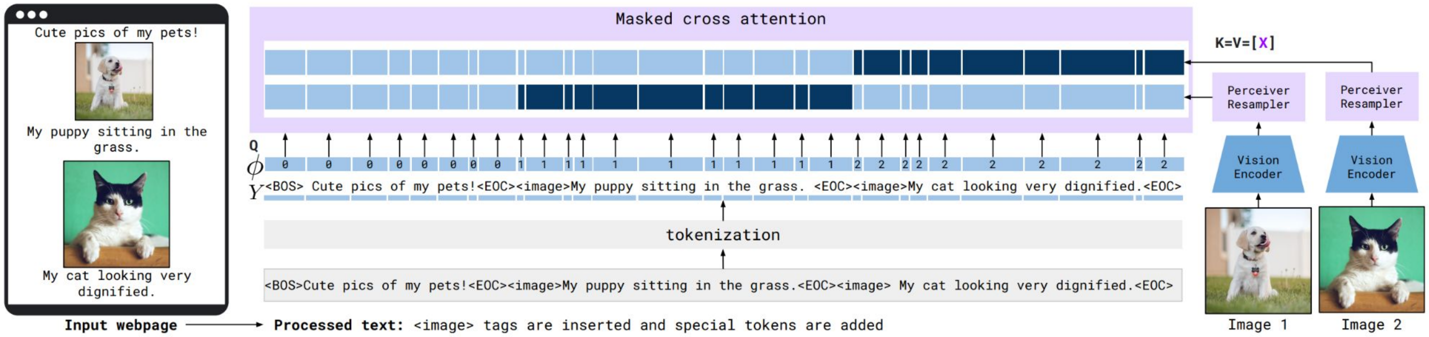


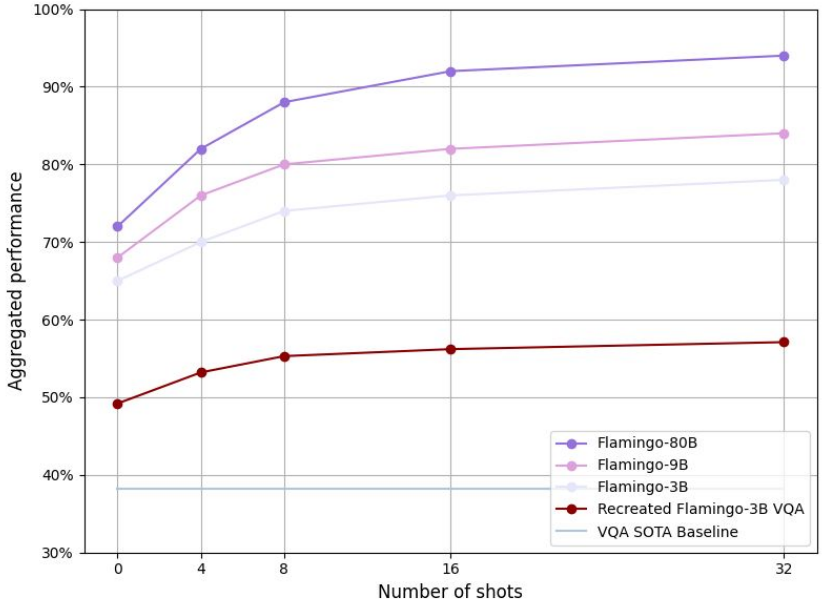
Figure 7: **Interleaved visual data and text support.** Given text interleaved with images/videos,

2. Image captioning demo with OpenFlamingo* (link to Colab)

*https://huggingface.co/openflamingo/OpenFlamingo-3B-vitl-mpt1b-langinstruct

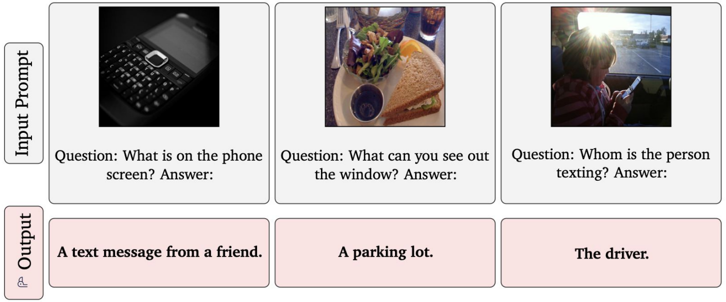
Demo

3. Recreating n-shot performance on VQA task with Flamingo-3B* (link to Colab)



*https://huggingface.co/openflamingo/OpenFlamingo-3B-vitl-mpt1b-langinstruct

Discussion: Trade-offs

	Benefits	Limitations
Uses pretrained LLM	Computationally efficient - Less trainable parameters - Requires less data Competitive performance - Full capacity of trained LLM	Hallucination & ungrounded guesses 
Easily adapted to various VL tasks	Can handle open-ended tasks	Limited performance on classification tasks
In-context learning	Simple deployment - only requires inference - no hyperparameter tuning	Sensitive to demonstration Poor scaling of inference compute cost and absolute performance with the number of shots

Discussion: Impacts on Future Work

1. Frozen LLM backbone + Trainable Adapter + Cross attention
    a. **Gemini (Google DeepMind, 2023)** → Expands on Flamingo's architecture by incorporating *video and audio* understanding alongside images and text.
    b. **BLIP-2 (Salesforce, 2023)** → Uses a similar frozen LLM + vision encoder strategy but introduces Q-Former, a *cross-modal query transformer*.
    c. **LLaVA (2023)** → Uses a frozen LLaMA model with trainable image-text projection layers.

2. Scaling Multimodal Models with Web-Scale Data
    a. **DeepSeek-VL (DeepSeek AI, 2024)** → Uses *web-scale multimodal data for training*, inspired by Flamingo's large-scale pre-training approach.
    b. **OpenFlamingo (OpenAI, 2023)** → An open-source re-implementation of Flamingo, trained on massive datasets for broad adaptability.