Flamingo:
a Visual Language Model for Few-Shot Learning

Andrea Wynn and Xindi Wu

11/21/2022

PRINCETON UNIVERSITY

Overview

Motivation

Flamingo Model Architecture

Training Data & Objective

In-Context Learning & Fine Tuning

Evaluation & Ablation Results

Limitations

Related Work: CM3 & Frozen

Discussion

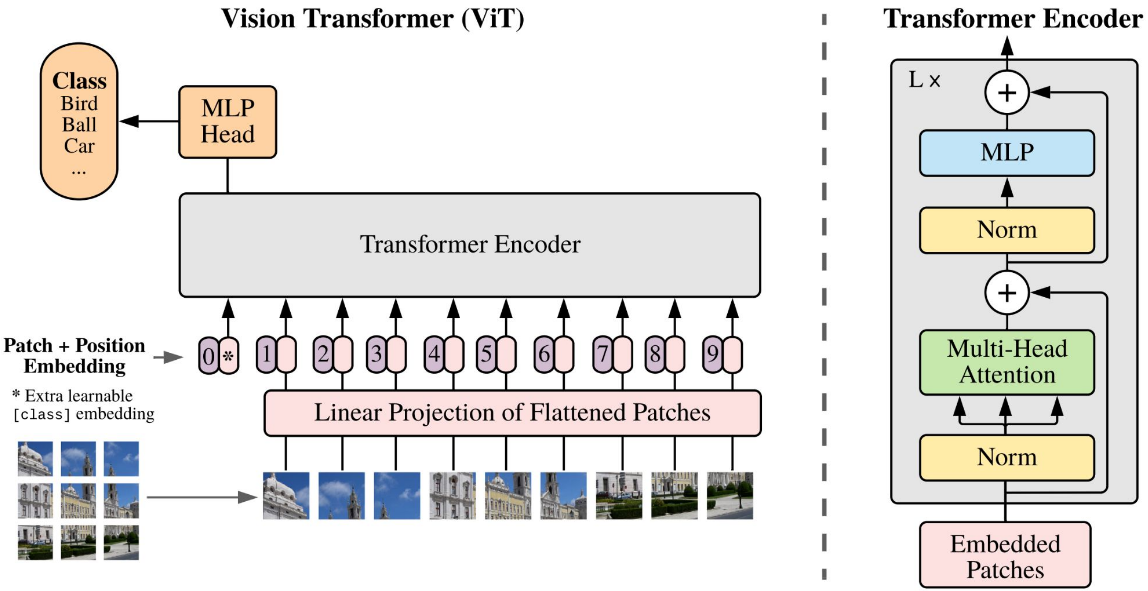
Motivation

GPT-3 VIT VisualBERT CLIP ?

Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

Motivation

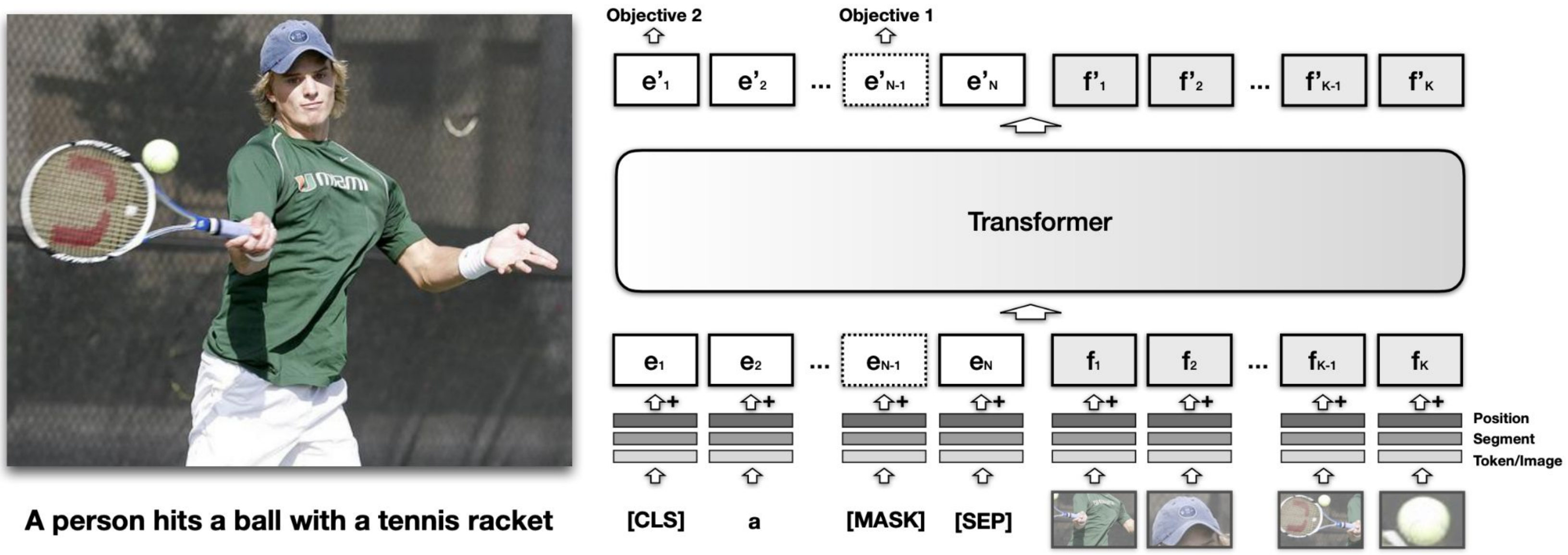
GPT-3 **VIT** VisualBERT CLIP ?



Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).

Motivation

GPT-3 VIT **VisualBERT** CLIP ?

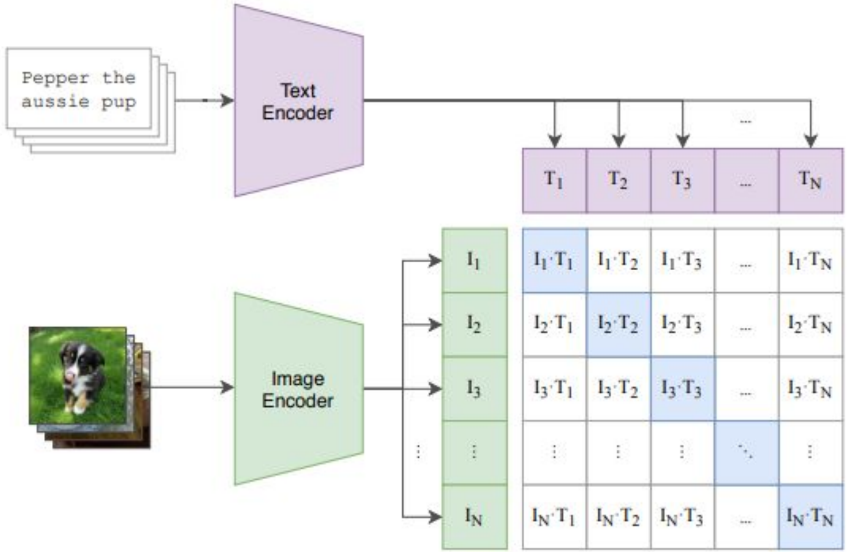


Li, Liunian Harold, et al. "Visualbert: A simple and performant baseline for vision and language." arXiv preprint arXiv:1908.03557 (2019).

Motivation

GPT-3 VIT VisualBERT **CLIP** ?

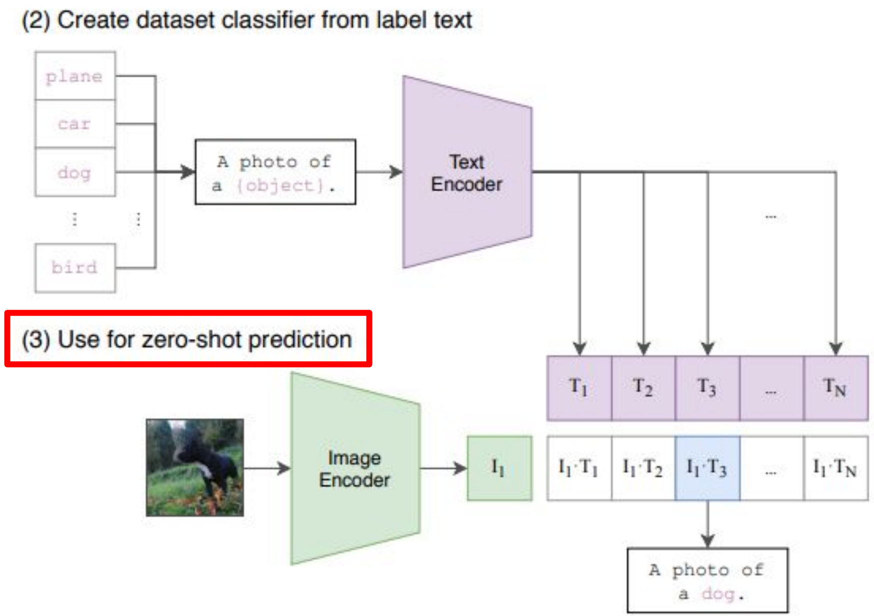
(1) Contrastive pre-training



Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International Conference on Machine Learning*. PMLR, 2021.

Motivation

GPT-3 VIT VisualBERT **CLIP** ?



Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International Conference on Machine Learning*. PMLR, 2021.

Motivation

GPT-3 VIT VisualBERT CLIP

Flamingo

The first vision-language model that has in-context learning ability

Motivation | Challenges

GPT-3 → VIT → VisualBERT → CLIP → **Flamingo**

Challenges of multimodal generative modelling

- Unifying strong single-modal models
    - Interleave **cross-attention** layers with language only self-attention layers

Motivation | Challenges

GPT-3 → VIT → VisualBERT → CLIP → **Flamingo**

Challenges of multimodal generative modelling

- Unifying strong single-modal models
    - Interleave **cross-attention** layers with language only self-attention layers
- Supporting images and videos
    - **Perceiver-based** architecture with a fixed number of visual tokens

Motivation | Challenges

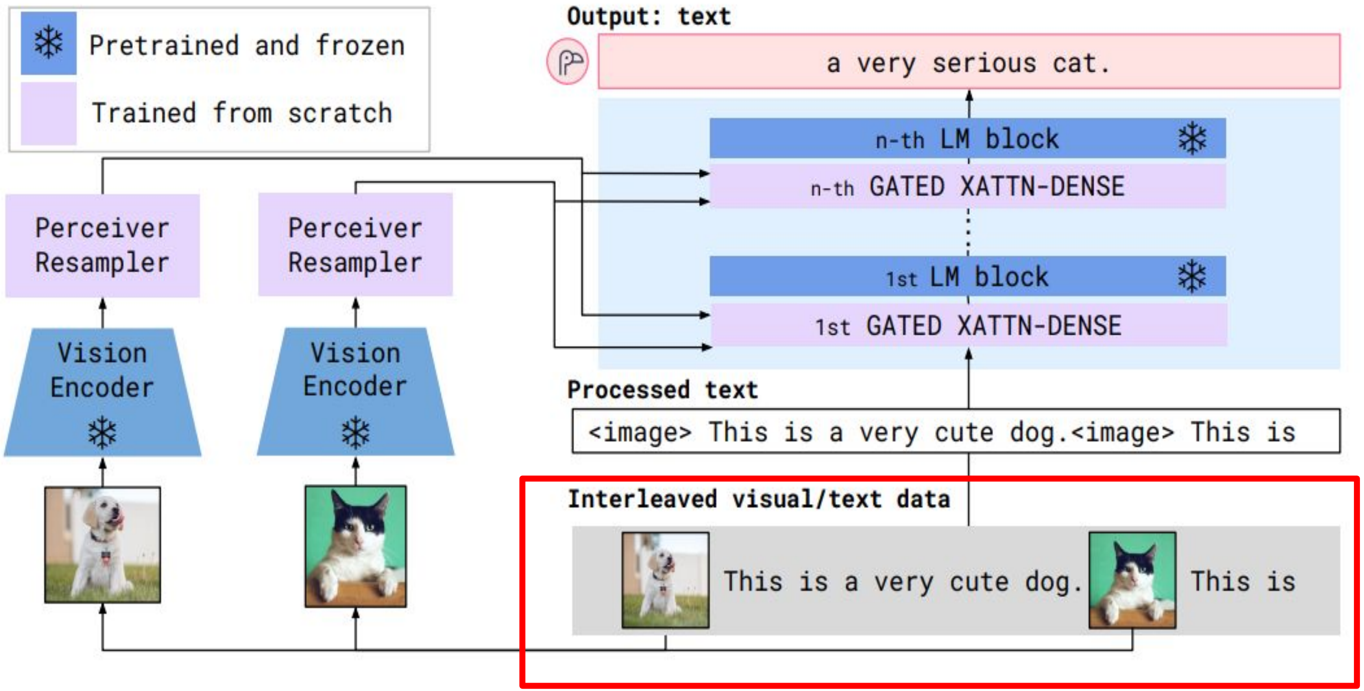
GPT-3 VIT VisualBERT CLIP **Flamingo**

Challenges of multimodal generative modelling

- Unifying strong single-modal models
    - Interleave **cross-attention** layers with language only self-attention layers
- Supporting images and videos
    - **Perceiver-based** architecture with a fixed number of visual tokens
- Heterogeneous training data
    - **Combine** web scraping with existing image-text or video-text datasets.

Flamingo Overview

Separately trained image + language models, with novel layers in between

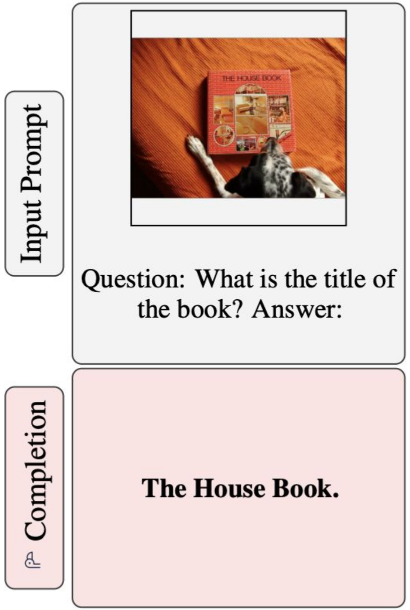


Input/Output

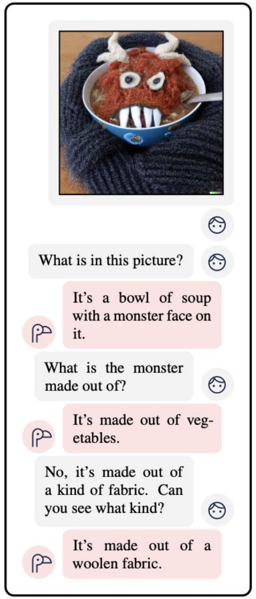
Interleaved inputs: text/images/video

Outputs:free-form text

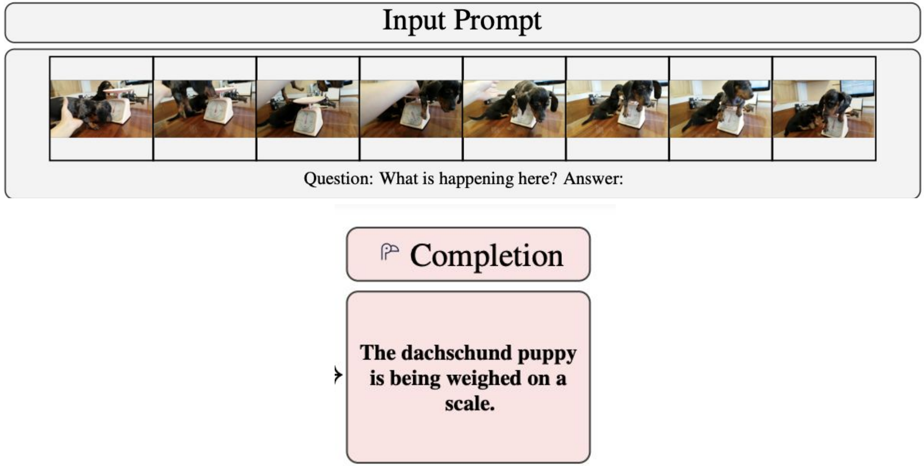
Selected single image samples



Selected dialogue samples

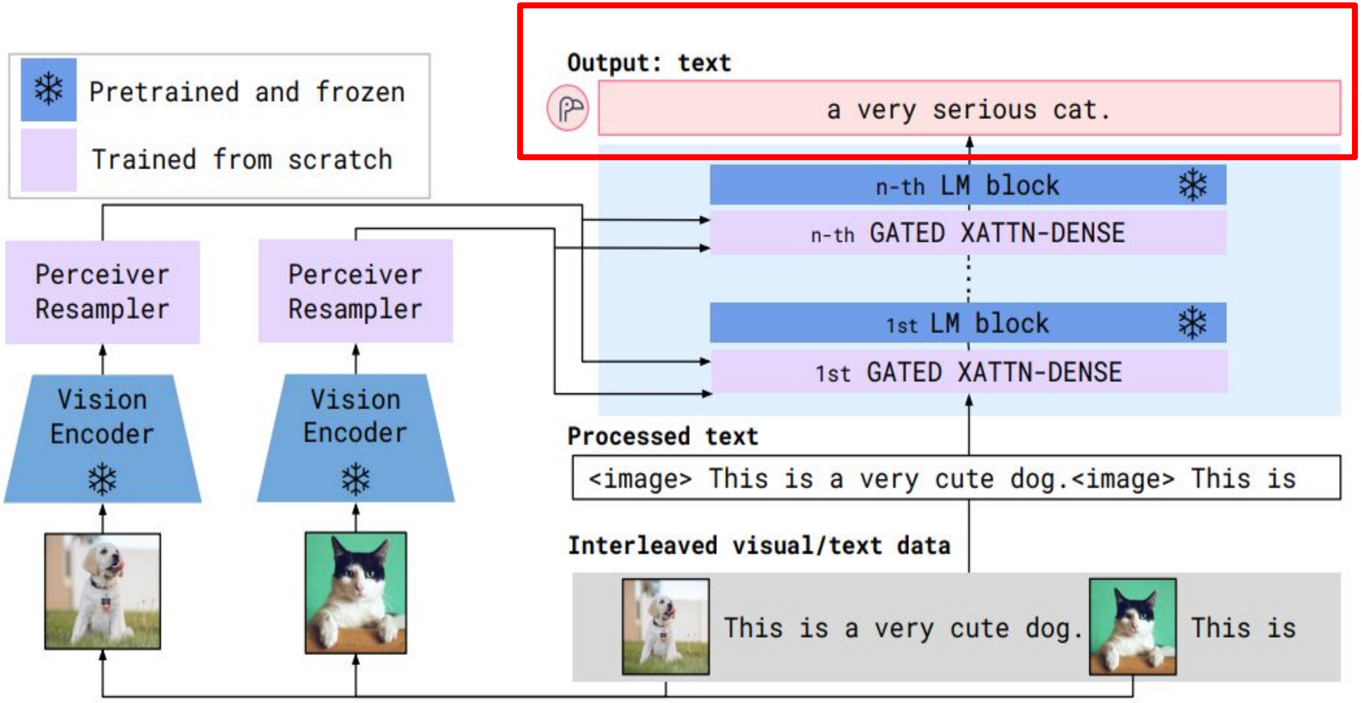


Selected video samples.



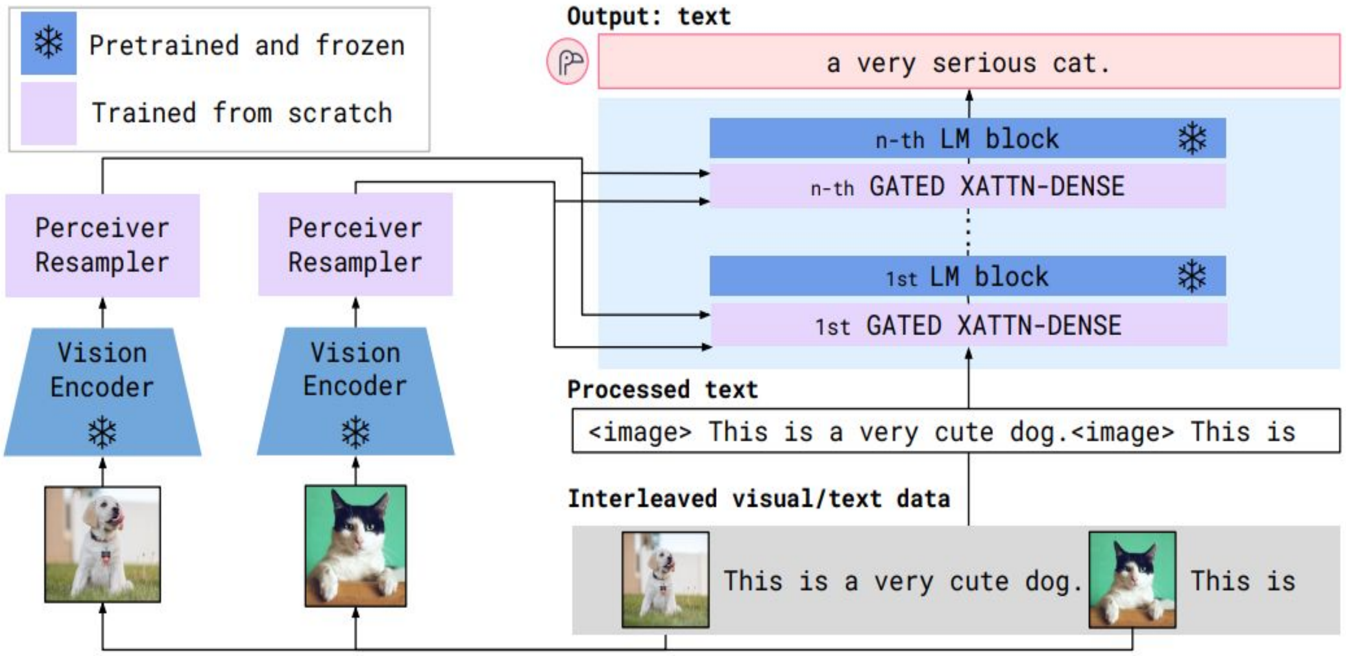
Flamingo Overview

Separately trained image + language models, with novel layers in between



Flamingo Overview

Separately trained image + language models, with novel layers in between

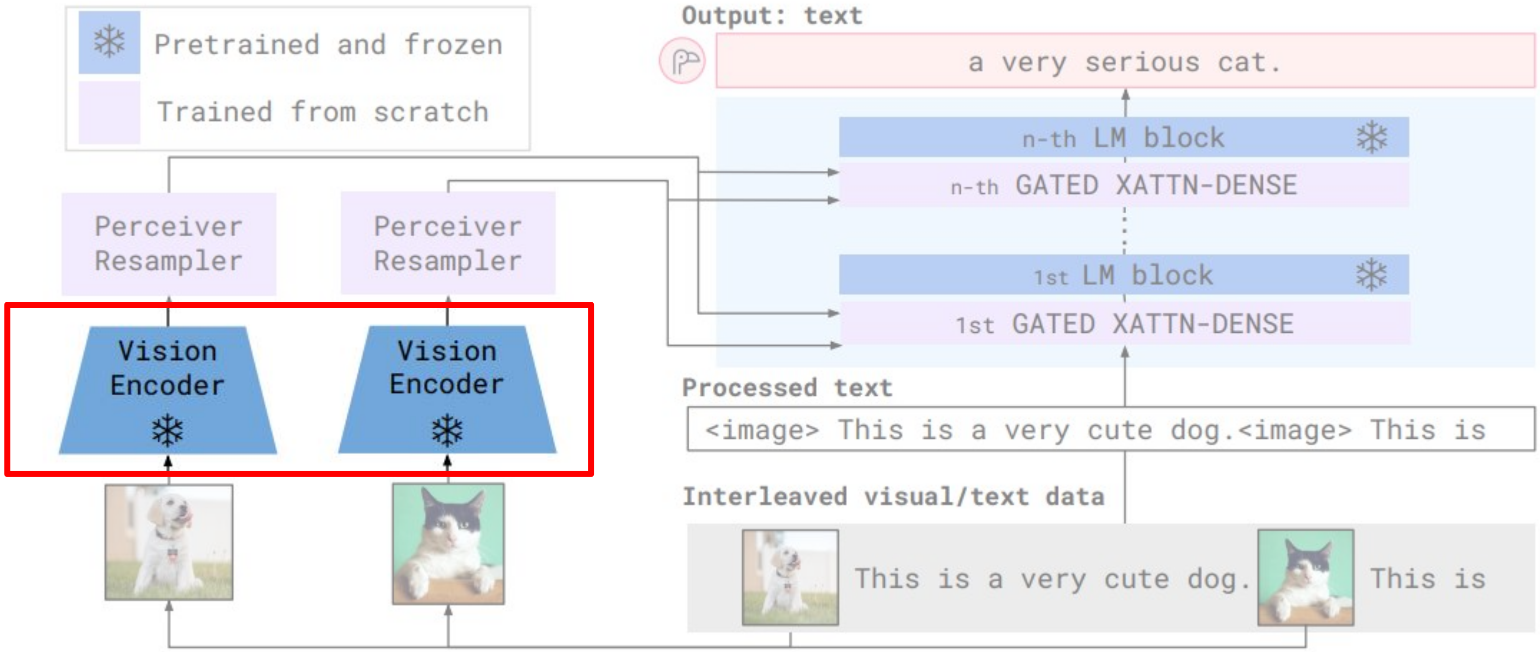


Flamingo Overview

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

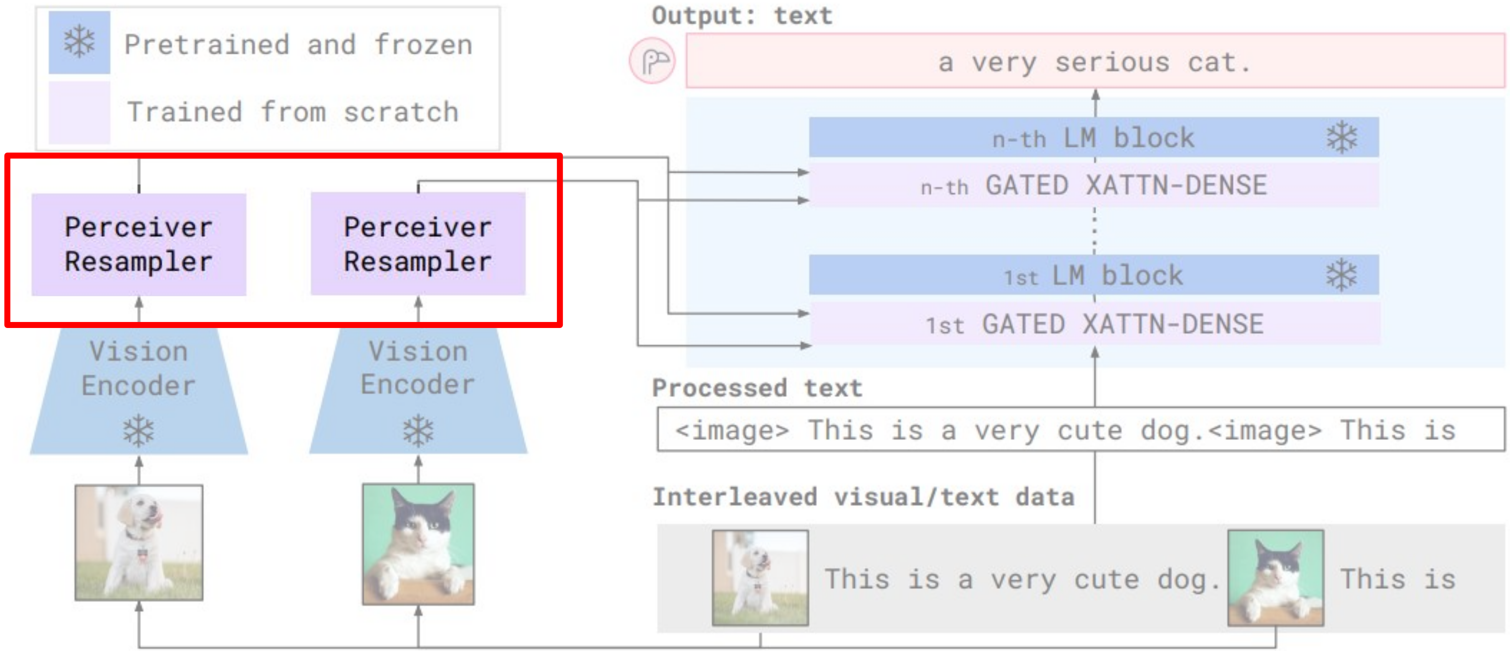
Vision Encoder

Pretrained and frozen Normalizer Free ResNet (NFNet)

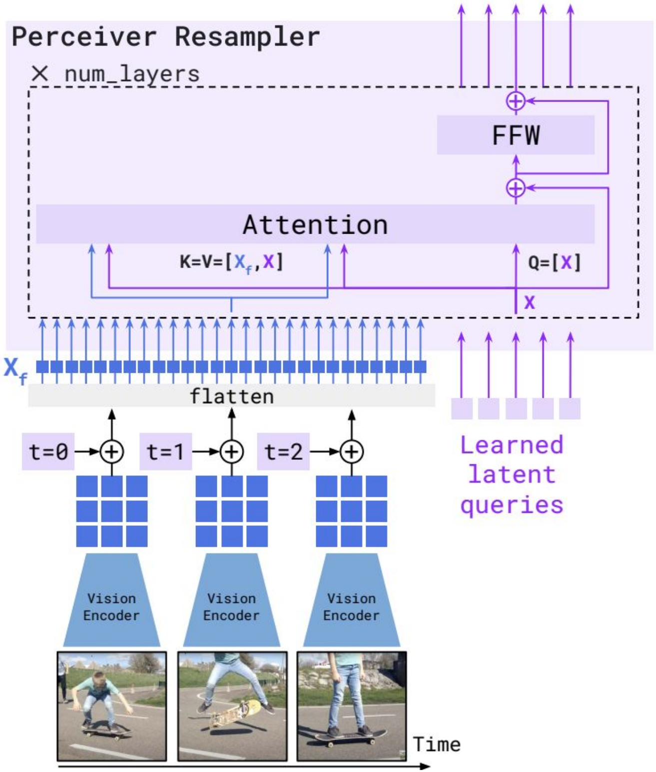


Brock, Andy, et al. "High-performance large-scale image recognition without normalization." *International Conference on Machine Learning*. PMLR, 2021.

Perceiver Resampler

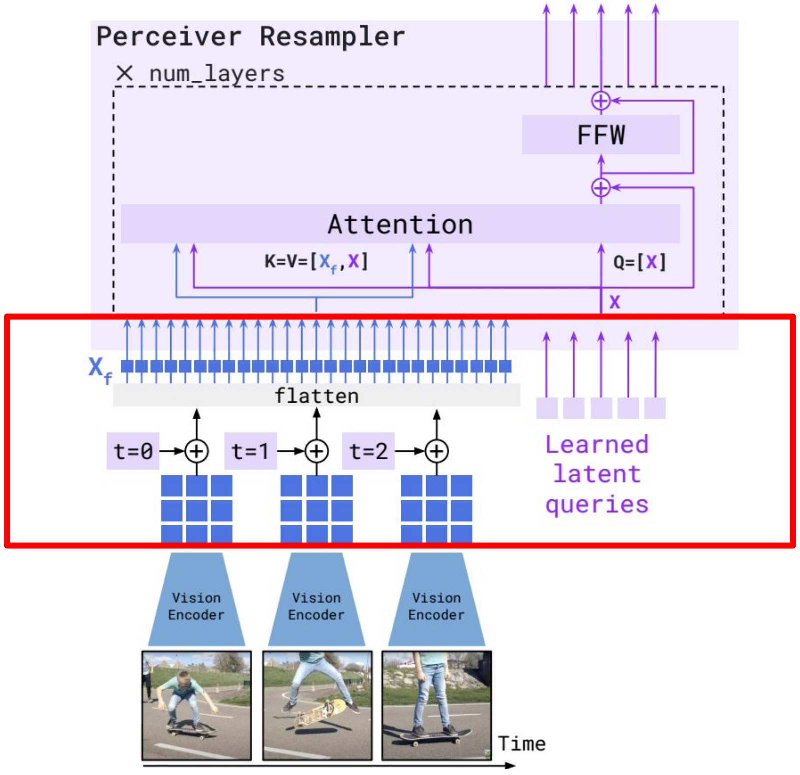


Perceiver Resampler



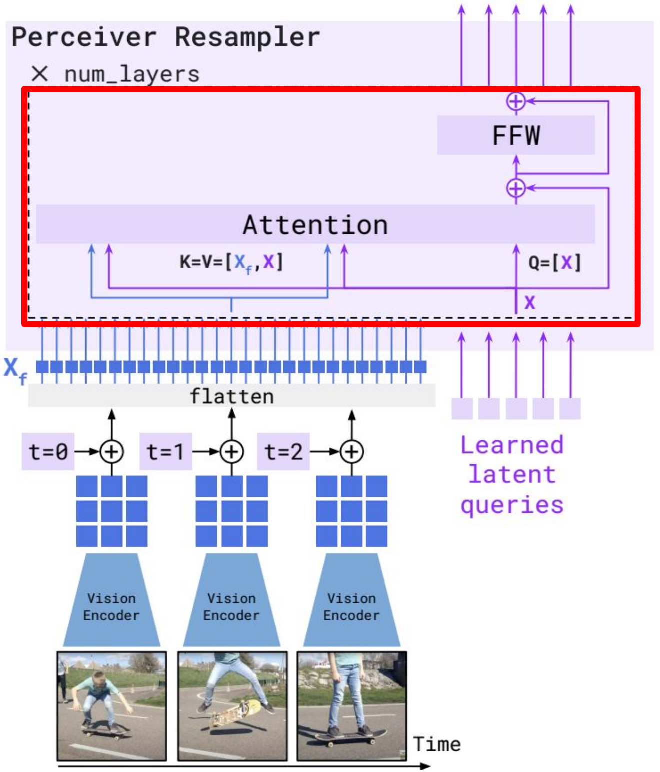

```
def perceiver_resampler(
    x_f,  # The [T, S, d] visual features (T=time, S=space)
    time_embeddings,  # The [T, 1, d] time pos embeddings.
    x,  # R learned latents of shape [R, d]
    num_layers,  # Number of layers
):
  """The Perceiver Resampler model."""

  # Add the time position embeddings and flatten.
  x_f = x_f + time_embeddings
  x_f = flatten(x_f)  # [T, S, d] -> [T * S, d]
  # Apply the Perceiver Resampler layers.
  for i in range(num_layers):
    # Attention.
    x = x + attention_i(q=x, kv=concat([x_f, x]))
    # Feed forward.
    x = x + ffw_i(x)
  return x
```

Perceiver Resampler



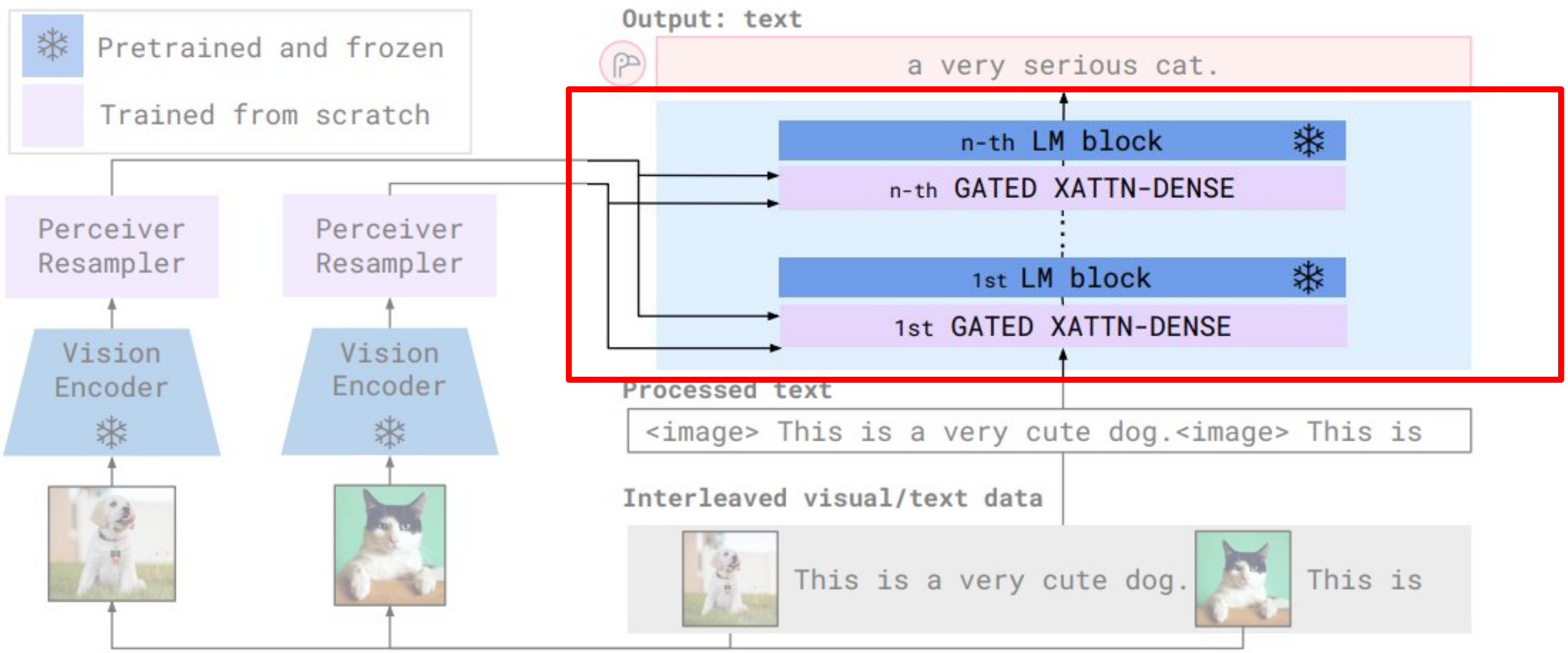

```
def perceiver_resampler(
    x_f,  # The [T, S, d] visual features (T=time, S=space)
    time_embeddings,  # The [T, 1, d] time pos embeddings.
    x,  # R learned latents of shape [R, d]
    num_layers,  # Number of layers
):
  """The Perceiver Resampler model."""

  # Add the time position embeddings and flatten.
  x_f = x_f + time_embeddings
  x_f = flatten(x_f)  # [T, S, d] -> [T * S, d]
  # Apply the Perceiver Resampler layers.
  for i in range(num_layers):
    # Attention.
    x = x + attention_i(q=x, kv=concat([x_f, x]))
    # Feed forward.
    x = x + ffw_i(x)
  return x
```

Perceiver Resampler

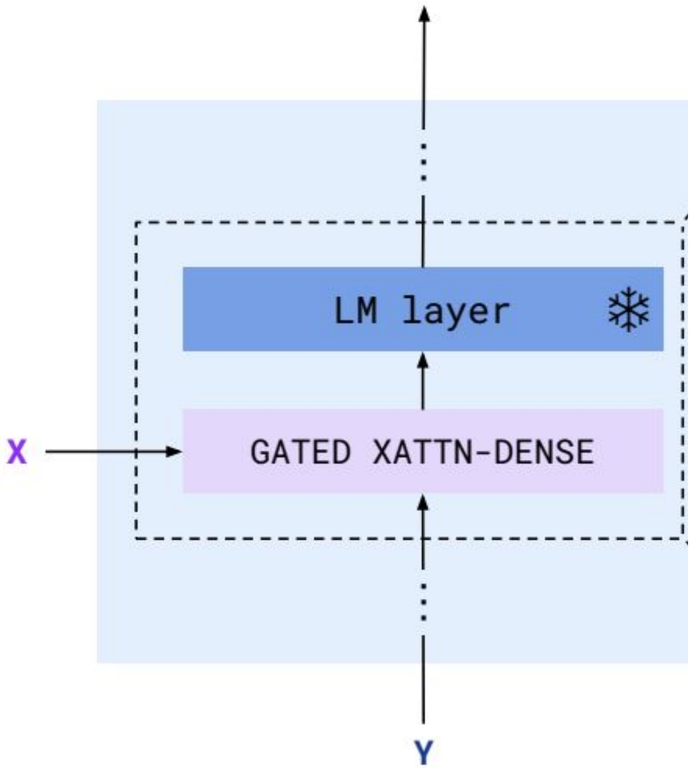



```
def perceiver_resampler(
    x_f,  # The [T, S, d] visual features (T=time, S=space)
    time_embeddings,  # The [T, 1, d] time pos embeddings.
    x,  # R learned latents of shape [R, d]
    num_layers,  # Number of layers
):
  """The Perceiver Resampler model."""

  # Add the time position embeddings and flatten.
  x_f = x_f + time_embeddings
  x_f = flatten(x_f)  # [T, S, d] -> [T * S, d]
  # Apply the Perceiver Resampler layers.
  for i in range(num_layers):
    # Attention.
    x = x + attention_i(q=x, kv=concat([x_f, x]))
    # Feed forward.
    x = x + ffw_i(x)
  return x
```

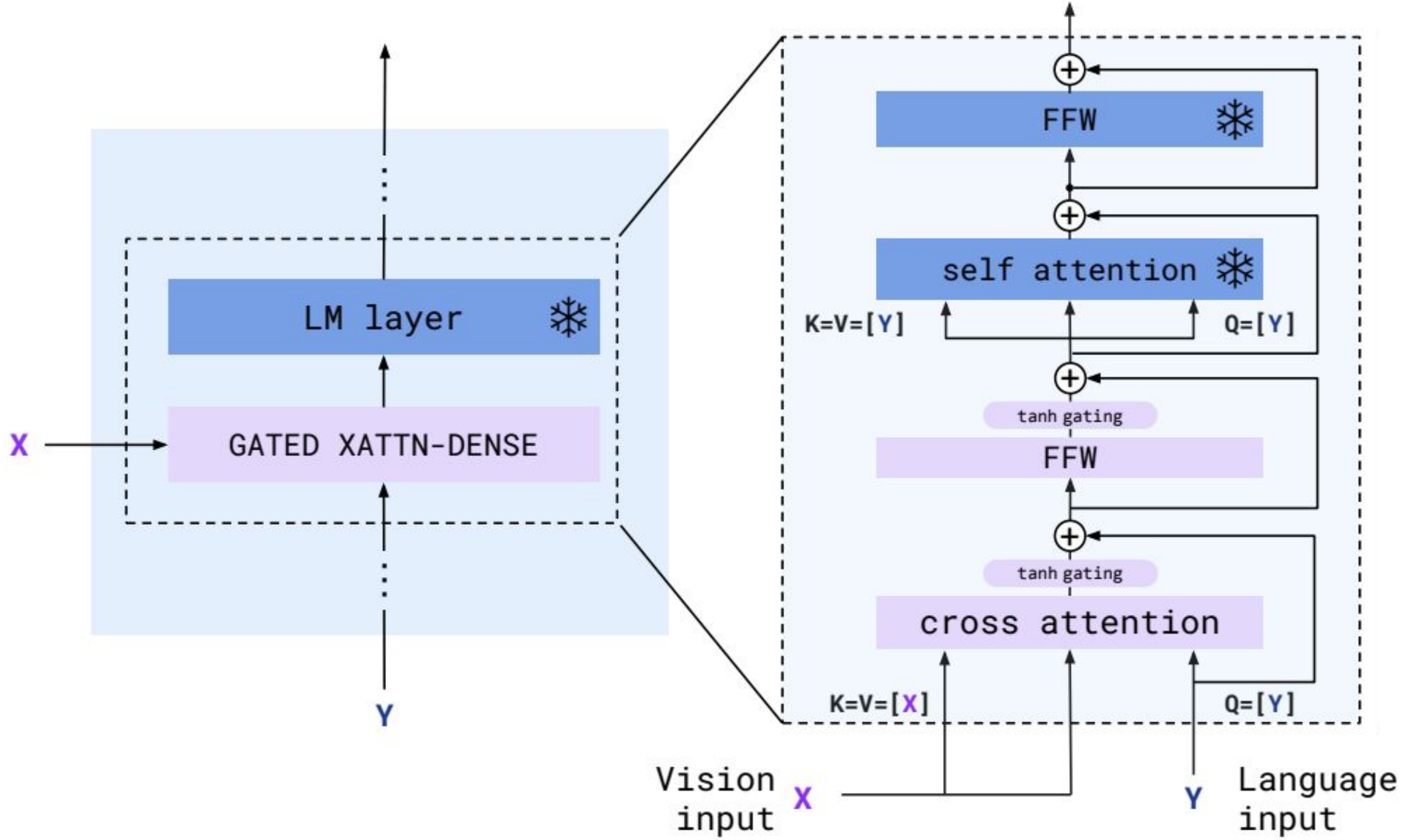
Conditioning the Language Model



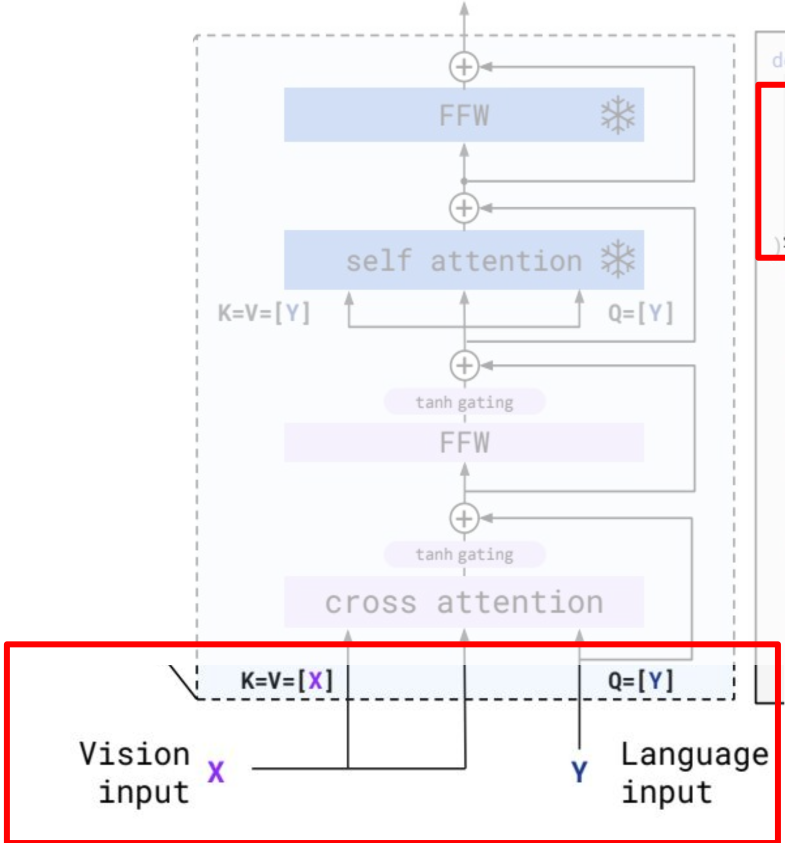
Gated XATTN-Dense layers



Gated XATTN-Dense layers

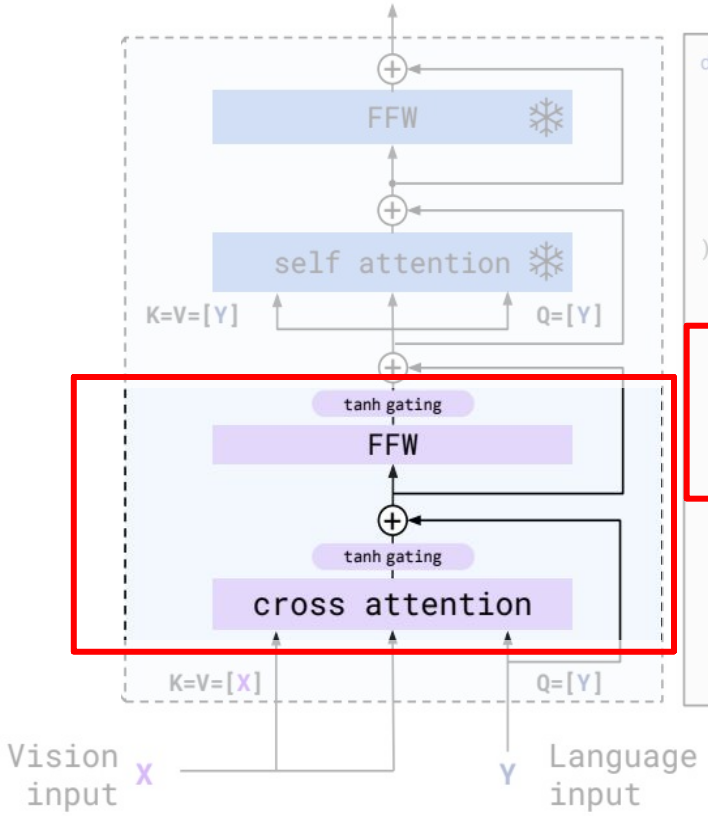


Gated XATTN-Dense layers



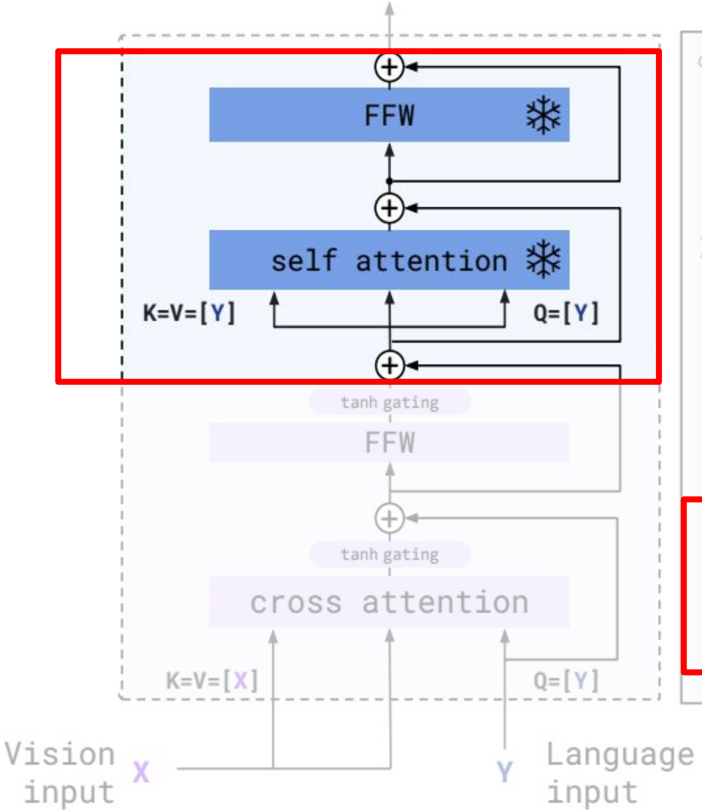

```
def gated_xattn_dense(
    y, # input language features
    x, # input visual features
    alpha_xattn, # xattn gating parameter – init at 0.
    alpha_dense, # ffw gating parameter – init at 0.
):
    """Applies a GATED XATTN-DENSE layer."""

    # 1. Gated Cross Attention
    y = y + tanh(alpha_xattn) * attention(q=y, kv=x)
    # 2. Gated Feed Forward (dense) Layer
    y = y + tanh(alpha_dense) * ffw(y)

    # Regular self-attention + FFW on language
    y = y + frozen_attention(q=y, kv=y)
    y = y + frozen_ffw(y)
    return y  # output visually informed language features
```

Gated XATTN-Dense layers



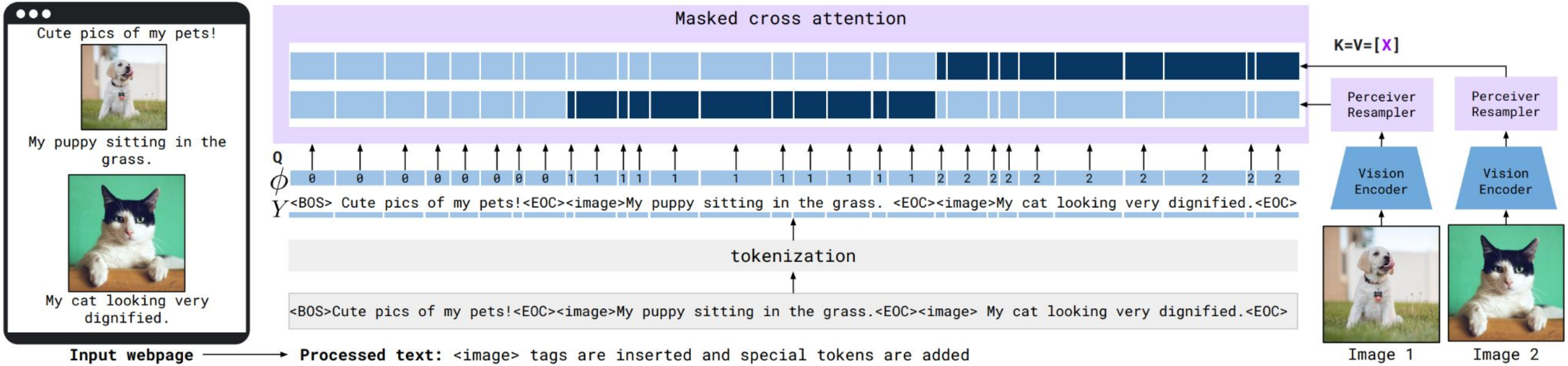

```
def gated_xattn_dense(
    y,  # input language features
    x,  # input visual features
    alpha_xattn, # xattn gating parameter – init at 0.
    alpha_dense, # ffw gating parameter – init at 0.
):
  """Applies a GATED XATTN-DENSE layer."""

  # 1. Gated Cross Attention
  y = y + tanh(alpha_xattn) * attention(q=y, kv=x)
  # 2. Gated Feed Forward (dense) Layer
  y = y + tanh(alpha_dense) * ffw(y)

  # Regular self-attention + FFW on language
  y = y + frozen_attention(q=y, kv=y)
  y = y + frozen_ffw(y)
  return y  # output visually informed language features
```

Gated XATTN-Dense layers

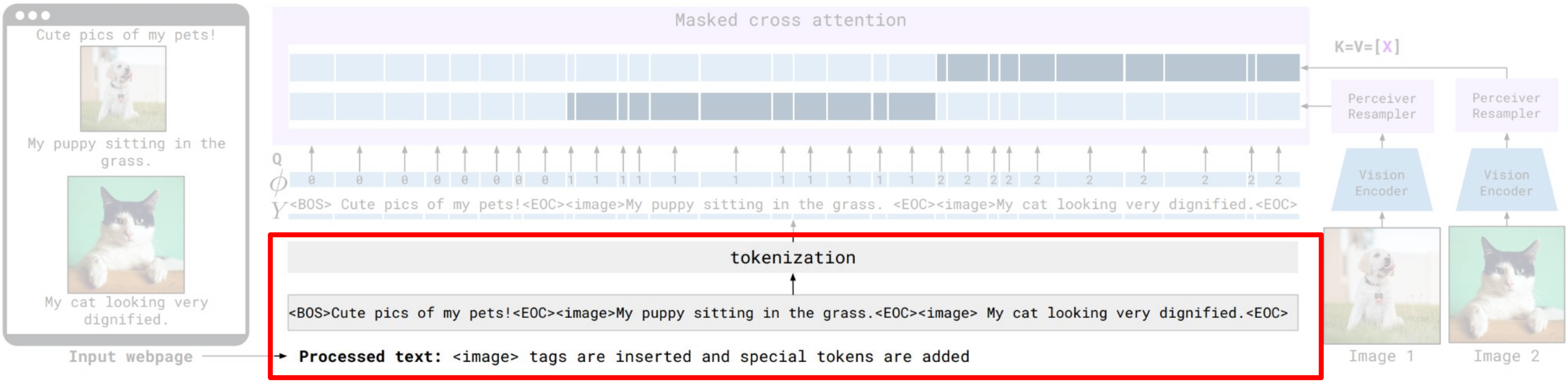



```
def gated_xattn_dense(
    y,  # input language features
    x,  # input visual features
    alpha_xattn, # xattn gating parameter – init at 0.
    alpha_dense, # ffw gating parameter – init at 0.
):
  """Applies a GATED XATTN-DENSE layer."""

  # 1. Gated Cross Attention
  y = y + tanh(alpha_xattn) * attention(q=y, kv=x)
  # 2. Gated Feed Forward (dense) Layer
  y = y + tanh(alpha_dense) * ffw(y)

  # Regular self-attention + FFW on language
  y = y + frozen_attention(q=y, kv=y)
  y = y + frozen_ffw(y)
  return y  # output visually informed language features
```

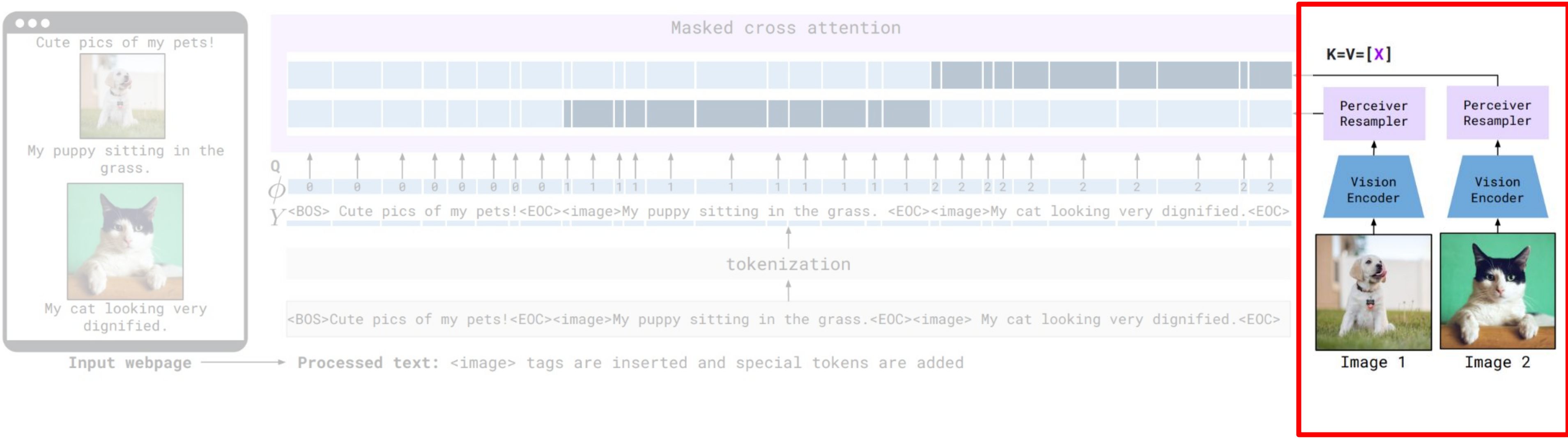
Multi-Visual Input Support



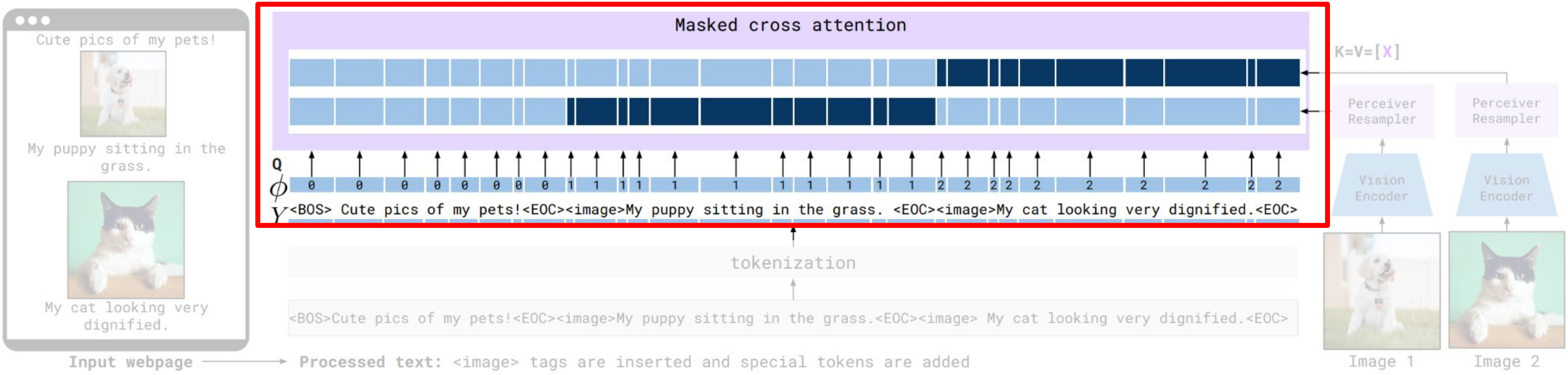
Multi-Visual Input Support



Multi-Visual Input Support



Multi-Visual Input Support



Pre-Lecture Question

Describe how Flamingo handles input sequences of arbitrarily interleaved textual and visual data, and combines pre-trained text-only and vision-only models.

Answer:

For example, the input contains an image of a dog together with a text description and an image of a cat with an incomplete text description. The text is parsed from the input with images replaced with placeholders the images are also extracted from the input passed through a frozen vision encoder and then mapped through the perceiver resampler to produce a fixed number of visual tokens per input.

Training Data

Mixture of Datasets

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]

- N: Number of visual inputs for a single example
- T: Number of video frames
- H, W, C: height, width, color channels

Interleaved Image/Text: MultiModal MassiveWeb (M3W)

- Interleaved text and image training data
- Compiled from webpage HTML
- Randomly sample 256 token subsequence and extract first 5 images

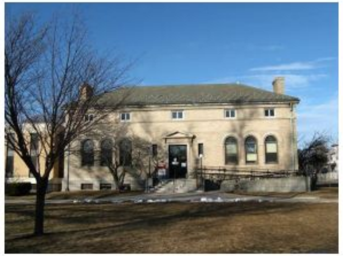
Example:

Multi-Modal Massive Web (M3W) dataset

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

Image-Text Pairs: ALIGN

"motorcycle front wheel"



"*thumbnail for version as of 21 57 29 june 2010*"

"file frankfurt airport skyline 2017 05 jpg"

"file london barge race 2 jpg"

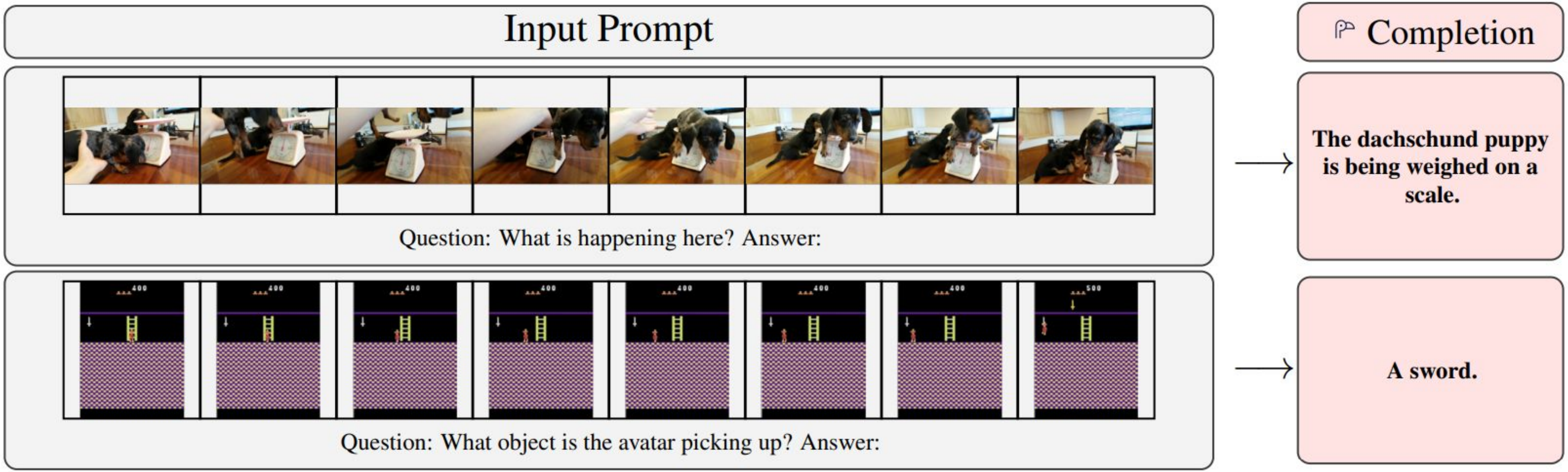
"moustache seamless wallpaper design"

"st oswalds way and shops"

Source: https://arxiv.org/pdf/2102.05918v2.pdf

Image-Text Pairs: Long Text & Image Pairs (LTIP)

Video & Text Pairs (VTP)



Data Augmentation & Preprocessing

- Visual inputs resized to 320x320
- M3W Data Augmentation: Randomizing image placement

(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!

Training Objective

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

- Weighted sum of dataset specific expected negative log likelihood of text, given some visual inputs
- AdamW optimizer
- No weight decay for Perceiver Resampler
- Weight decay of 0.1 for other parameters

Pre-Lecture Question

Describe what datasets are used for mixed training. How important is each type of dataset empirically?

Answer:

Datasets - M3W (interleaved images and text), ALIGN (large, lower quality image + text pairs), LTIP (image + text pairs), VTP (video + text pairs)

Importance (lambda weights) - 1.0 (M3W), 0.2 (ALIGN), 0.2 (LTIP), 0.03 (VTP)

Number of datasets (M) - 4

Flamingo Evaluation

Benchmark Tasks

	Dataset	DEV	Gen.	Custom prompt	Task description
Image	ImageNet-1k [94]	✓			Object classification
	MS-COCO [15]	✓	✓		Scene description
	VQAv2 [3]	✓	✓		Scene understanding QA
	OKVQA [69]	✓	✓		External knowledge QA
	Flickr30k [139]		✓		Scene description
	VizWiz [35]		✓		Scene understanding QA
	TextVQA [100]		✓		Text reading QA
	VisDial [20]				Visual Dialogue
	HatefulMemes [54]			✓	Meme classification
Video	Kinetics700 2020 [102]	✓			Action classification
	VATEX [122]	✓	✓		Event description
	MSVDQA [130]	✓	✓		Event understanding QA
	YouCook2 [149]		✓		Event description
	MSRVTTQA [130]		✓		Event understanding QA
	iVQA [135]		✓		Event understanding QA
	RareAct [73]			✓	Composite action retrieval
	NextQA [129]		✓		Temporal/Causal QA
	STAR [128]				Multiple-choice QA

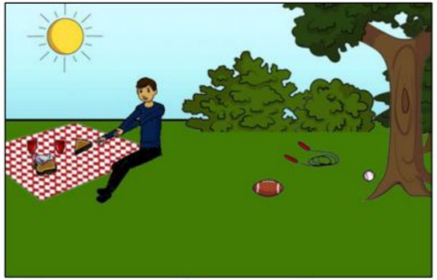
Benchmark Tasks: ImageNet-1k

Source: https://link.springer.com/content/pdf/10.1007/s11263-015-0816-y.pdf

Benchmark Tasks: Visual Question Answering (VQA)

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

Source: https://link.springer.com/content/pdf/10.1007/s11263-016-0966-6.pdf

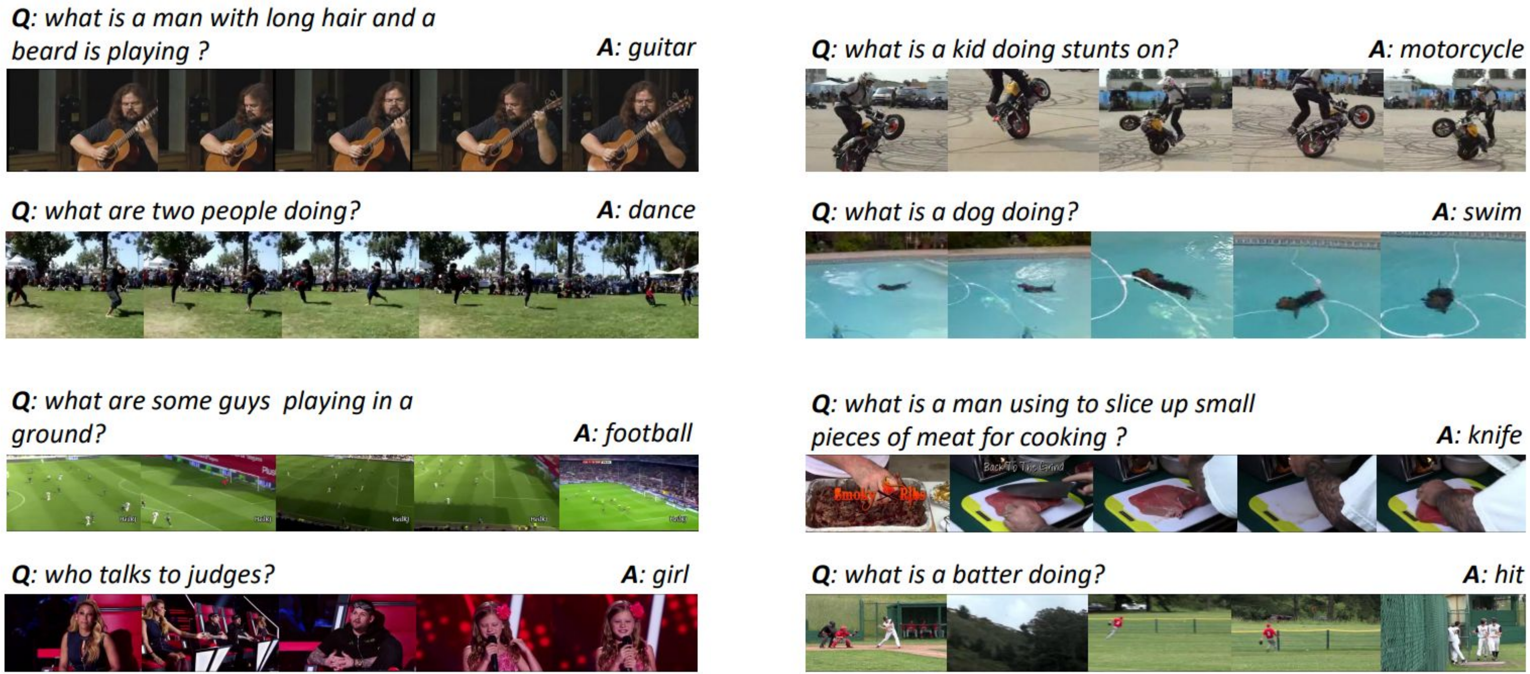
Benchmark Tasks: Kinetics700 2020

- Taken from YouTube videos
- Format: label, youtube_id, start time, end time

label	youtube_id	time_start	time_end
clay pottery making	---0dWlqevI	19	29
javelin throw	--07WQ2iBlw	1	11
climbing a rope	--0NTAs-fA0	29	39
sipping cup	--0I35AkU34	68	78
flipping pancake	--33Lscn6sk	4	14
tickling	--3OAstUWtU	45	55

Source: https://arxiv.org/pdf/2210.10864.pdf

Benchmark Tasks: MSVDQA



Source: https://dl.acm.org/doi/pdf/10.1145/3123266.3123427

Classification Task Results

Model	Method	Prompt size	shots/class	ImageNet top 1	Kinetics700 avg top1/5
SotA	Fine-tuned	-	full	90.9 [127]	89.0 [134]
SotA	Contrastive	-	0	**85.7** [82]	**69.6** [85]
NFNetF6	Our contrastive	-	0	77.9	62.9
Flamingo-3B	RICES	8	1	70.9	55.9
		16	1	71.0	56.9
		16	5	72.7	58.3
Flamingo-9B	RICES	8	1	71.2	58.0
		16	1	71.7	59.4
		16	5	75.2	60.9
Flamingo-80B	Random	16	≤ 0.02	66.4	51.2
	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

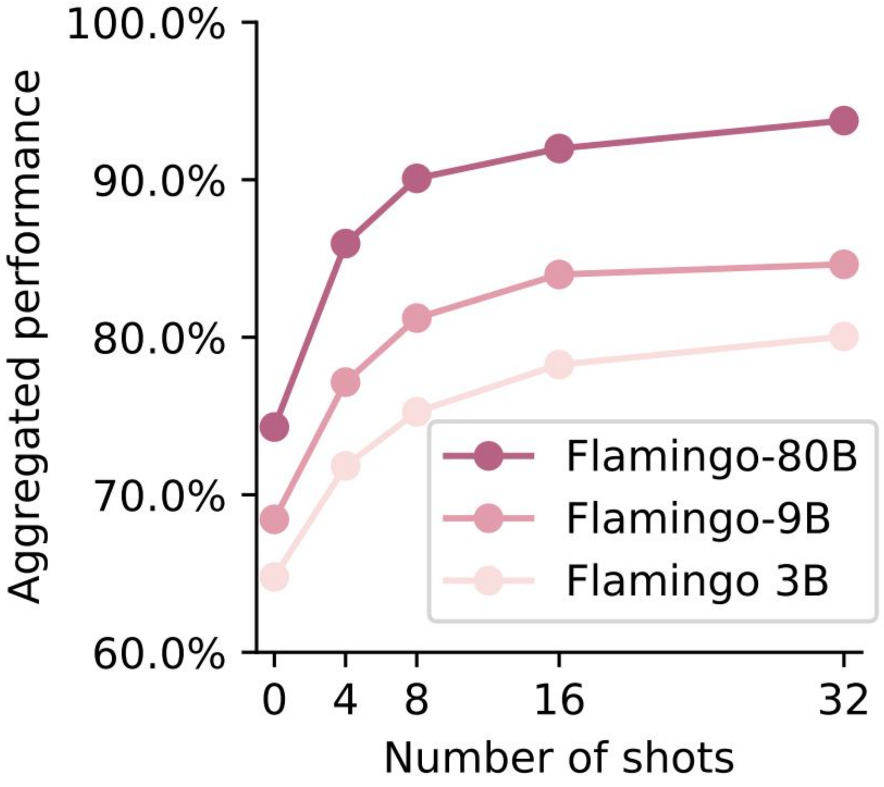
Fine Tuning Results

Method	VQAV2 test-dev	VQAV2 test-std	COCO test	VATEX test	VizWiz test-dev	VizWiz test-std	MSRVTTQA test	VisDial valid	VisDial test-std	YouCook2 valid	TextVQA valid	TextVQA test-std	HatefulMemes test seen
Flamingo - 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	**143.3**	-	-	-	-	-	-	-	-	-	-
OFA [119]	79.9	80.0	149.6	-	-	-	-	-	-	-	-	-	-
Florence [140]	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 [140]	80.4 [140]	**143.3** [124]	76.3 [153]	-	-	46.8 [51]	**75.2** [79]	**74.5** [79]	**138.7** [132]	54.7 [137]	**73.7** [84]	79.1 [62]
Unrestricted SotA	81.3 [133]	81.3 [133]	149.6 [119]	81.4 [153]	57.2 [65]	60.6 [65]	-	-	75.4 [123]	-	-	-	84.6 [152]

Model Scaling

	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**

Number of Shots



Ablation Studies

Ablation Studies

	Ablated setting	*Flamingo*-3B original value	Changed value	Overall score↑
		***Flamingo*-3B model**		**70.7**
(i)	Training data	All data	w/o Video-Text pairs	67.3
			w/o Image-Text pairs	60.9
			Image-Text pairs→ LAION	66.4
			w/o M3W	53.4
(ii)	Optimisation	Accumulation	Round Robin	62.9
(iii)	Tanh gating	✓	✗	66.5
(iv)	Cross-attention architecture	GATED XATTN-DENSE	VANILLA XATTN	66.9
			GRAFTING	63.1
(v)	Cross-attention frequency	Every	Single in middle	59.8
			Every 4th	68.8
			Every 2nd	68.2
(vi)	Resampler	Perceiver	MLP	66.6
			Transformer	66.7
(vii)	Vision encoder	NFNet-F6	CLIP ViT-L/14	64.9
			NFNet-F0	62.7
(viii)	Freezing LM	✓	✗ (random init)	57.8
			✗ (pretrained)	62.7

Pre-Training Dataset Ablation

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

Frozen Language Model

	Ablated setting	Flamingo 3B value	Changed value	Overall score↑
		Flamingo 3B model (short training)		**70.7**
(i)	Resampler size	Medium	Small Large	67.9 69.0
(ii)	Multi-Img att.	Only last	All previous	63.5
(iii)	p_{next}	0.5	0.0 1.0	69.6 70.4
(iv)	LM pretraining	MassiveText	C4	62.8
(v)	Freezing Vision	✓	✗ (random init) ✗ (pretrained)	61.4 68.1
(vi)	Co-train LM on MassiveText	✗	✓ (random init) ✓ (pretrained)	55.9 68.6
(vii)	Dataset and Vision encoder	M3W+ITP+VTP and NFNetF6	LAION400M and CLIP M3W+LAION400M+VTP and CLIP	54.7 64.9

0-initialized tanh gating

	Ablated setting	*Flamingo*-3B original value	Changed value	Overall score↑
		***Flamingo*-3B model**		**70.7**
(i)	Training data	All data	w/o Video-Text pairs	67.3
			w/o Image-Text pairs	60.9
			Image-Text pairs→ LAIO	66.4
			w/o M3W	53.4
(ii)	Optimisation	Accumulation	Round Robin	62.9
(iii)	Tanh gating	✓	✗	66.5
(iv)	Cross-attention architecture	GATED XATTN-DENSE	VANILLA XATTN	66.9
			GRAFTING	63.1
(v)	Cross-attention frequency	Every	Single in middle	59.8
			Every 4th	68.8
			Every 2nd	68.2
(vi)	Resampler	Perceiver	MLP	66.6
			Transformer	66.7

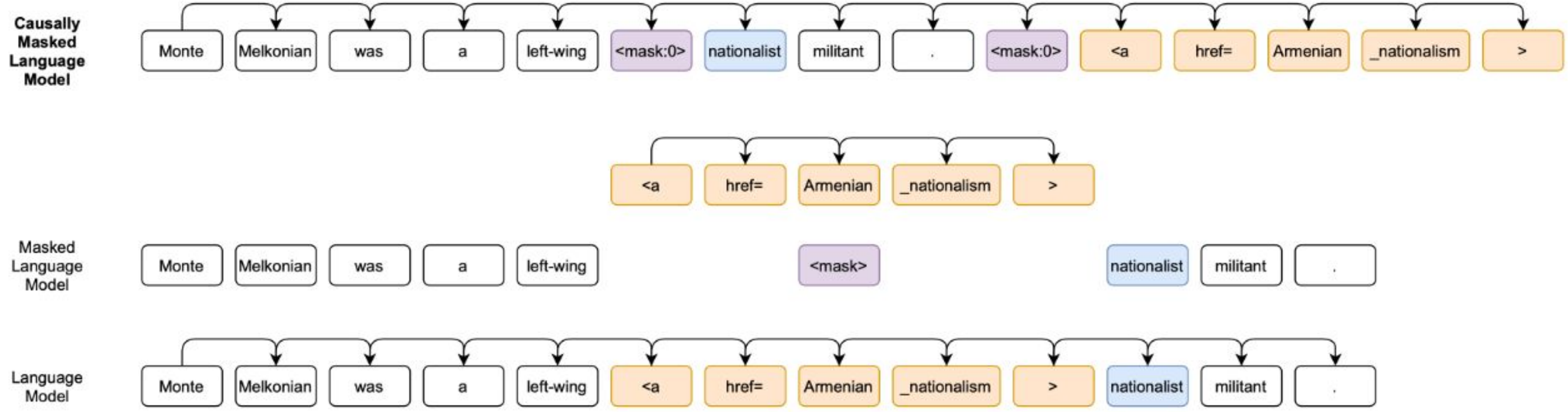
Failures: Hallucinations

Input Prompt	 Question: What is on the phone screen? Answer:	 Question: What can you see out the window? Answer:	 Question: Whom is the person texting? Answer:
Output	**A text message from a friend.**	**A parking lot.**	**The driver.**

Survey of Visual LMs

CM3

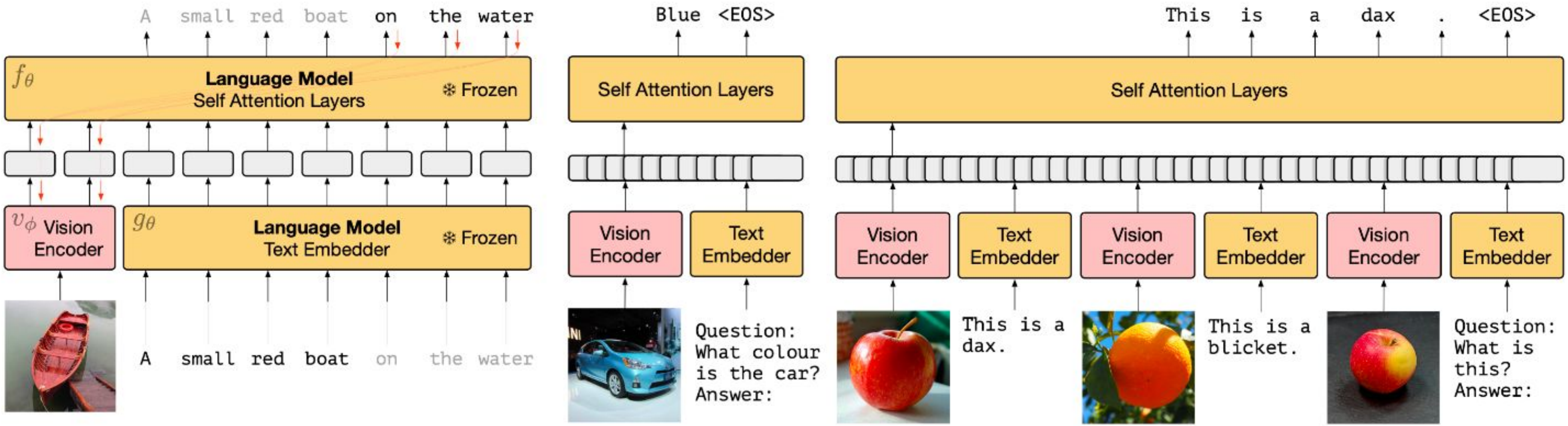
- Causally Masked Multimodal Modeling
- Images tokenized by VQVAE-GAN (source: https://arxiv.org/abs/2012.09841)



Paper: https://arxiv.org/abs/2201.07520

Learning Image Embeddings on Frozen LM Prefix

- Multimodal few shot learning for interleaved vision and text



Paper: https://arxiv.org/abs/2106.13884

Discussion

If you are going to build a visual LM for few-shot learning, what are the other ways of fusing visual and textual data? What pre-training data would you consider?