Few and zero shot learning

Flamingo & Kosmos-1

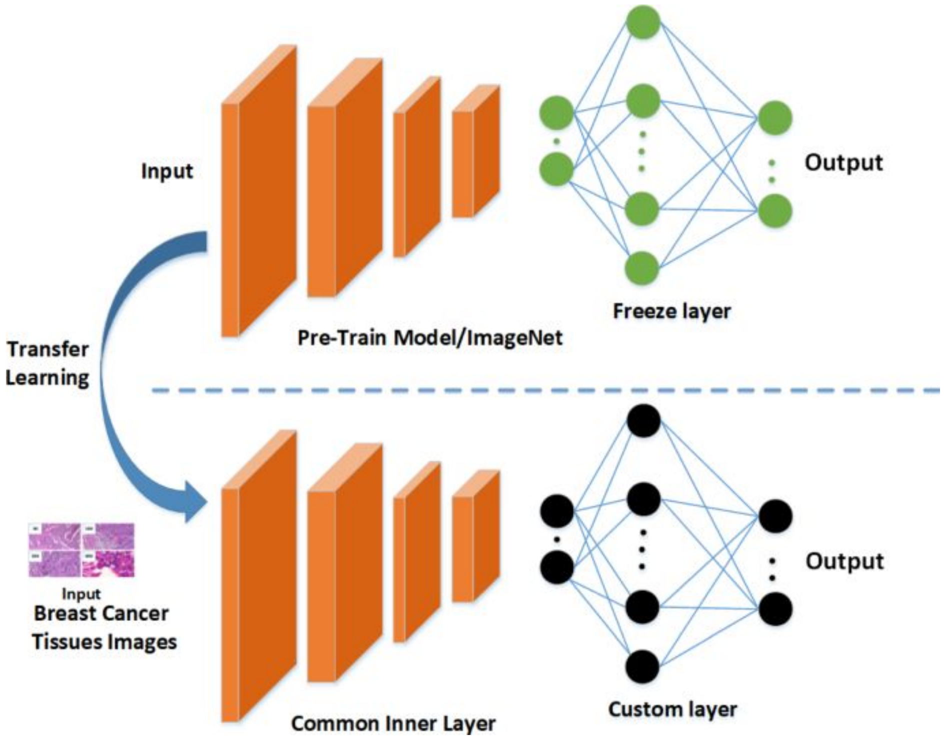
CS 6804: Multimodal Vision | Deval Srivastava

Background

- Few and zero shot learning can be seen as a measure of intelligence.

Background

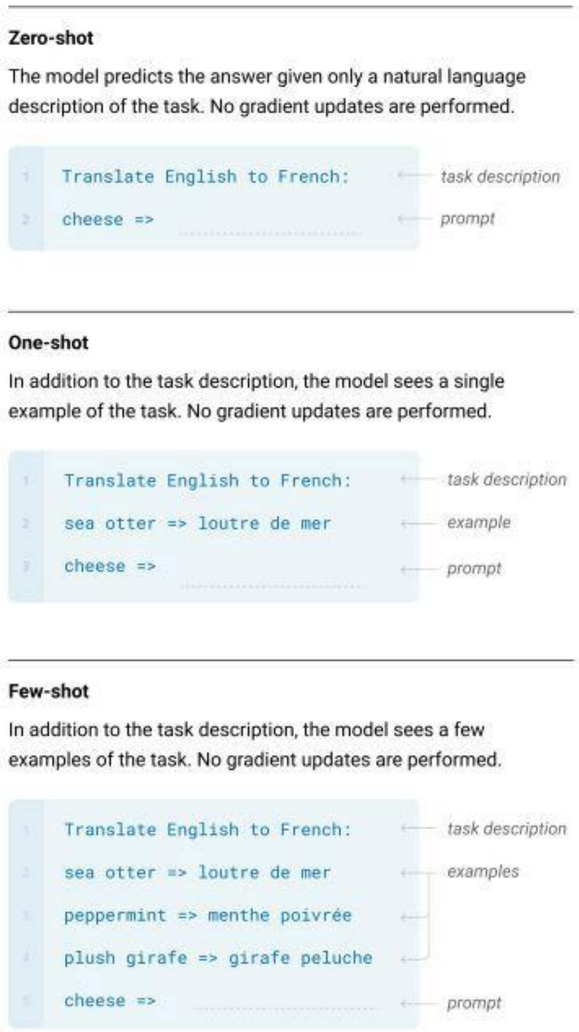
- Few and zero shot learning can be seen as a measure of intelligence.
- Different from how currently models learn.



https://link.springer.com/article/10.1007/s00371-021-02153-y

Background

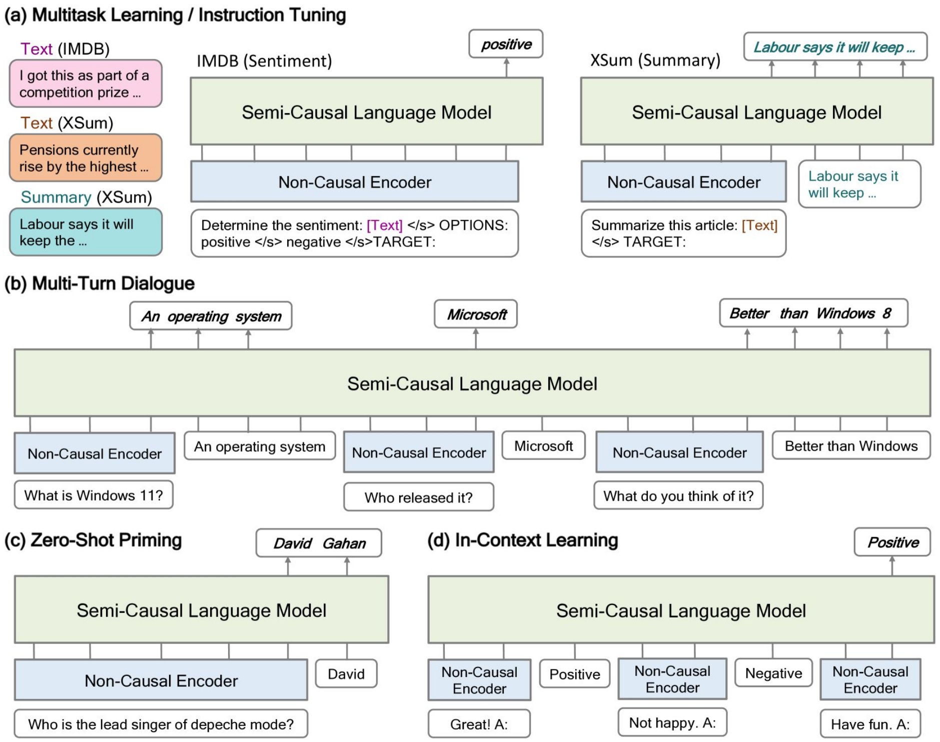
- Few and zero shot learning can be seen as a measure of intelligence.
- Different from how currently models learn.
- LLMs are able to do this to some extent[1]



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

Background

- Few and zero shot learning can be seen as a measure of intelligence.
- Different from how currently models learn.
- LLMs are able to do this to some extent[1].
- LLMs can do a number of tasks through their versatile text interface.



Hao, Yaru, et al. "Language models are general-purpose interfaces." arXiv preprint arXiv:2206.06336 (2022).

Motivation

GOAL:

- Design a multimodal LLM that can perform effective few shot and zero shot learning from prompts.
- This multimodal LLM would be trained on a variety of sources including interleaved images and text.

Two papers follow this methodology

- FLAMINGO[3]
- Kosmos-1[4]

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

Flamingo: Key Takeaways

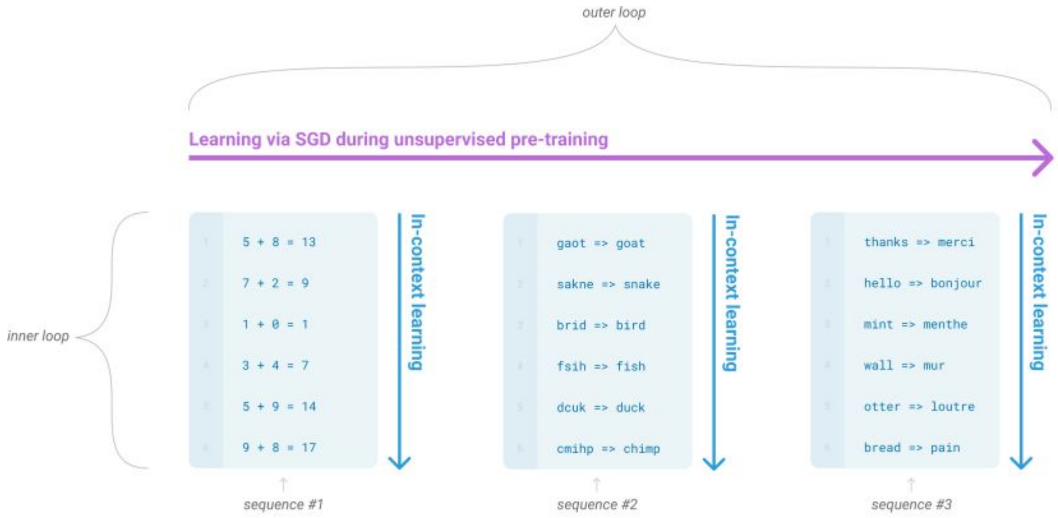
- A new large VLM that can ingest a sequence of text/image or interleaved tokens then output text.
- Sets a new state of the art on a variety of V+L tasks by being prompted with few input / output samples
- Introduces a novel architecture that bridges two frozen pretrained vision and language models.

Related Work

- **Chinchilla[5]**
  - Finetuning a LLM has become an effective strategy to use it for downstream tasks.

Related Work

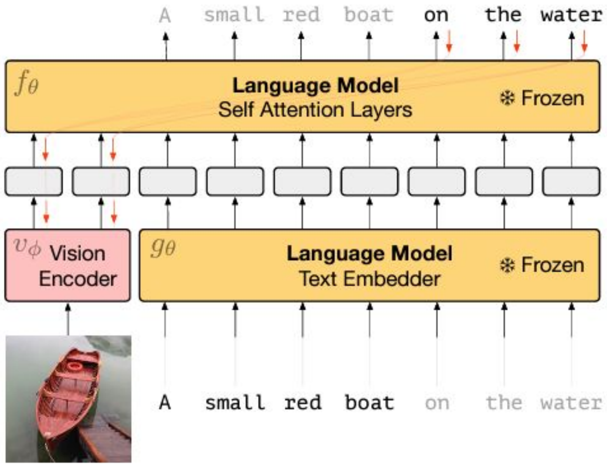
- **Chinchilla[5]**
- **GPT-3[1]**
    - Introduces a few shot in-context learning technique.



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

Related Work

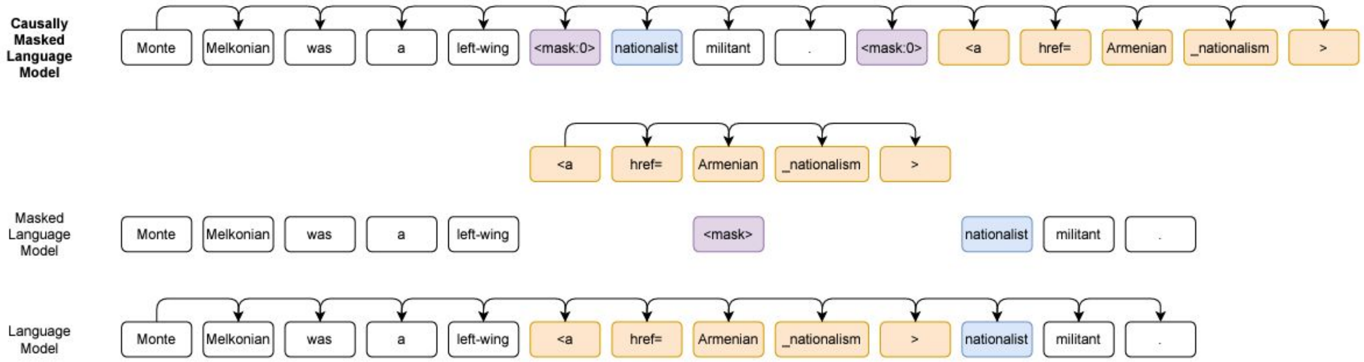
- **Chinchilla[5]**
- **GPT-3[1]**
- **Multimodal Few-Shot Learning with Frozen Language Models[6]**
    - Proposes to train frozen LLMs with few learnable layers on interleaved data for V+L tasks.



Tsimpoukelli, Maria, et al. "Multimodal few-shot learning with frozen language models." Advances in Neural Information Processing Systems 34 (2021): 200-212.

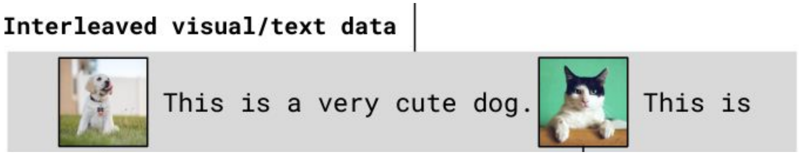
Related Work

- **Chinchilla[5]**
- **GPT-3[1]**
- **Multimodal Few-Shot Learning with Frozen Language Models[6]**
- **CM3[7]**
  - Proposes to train a masked LLM on extracted HTML data for language tasks



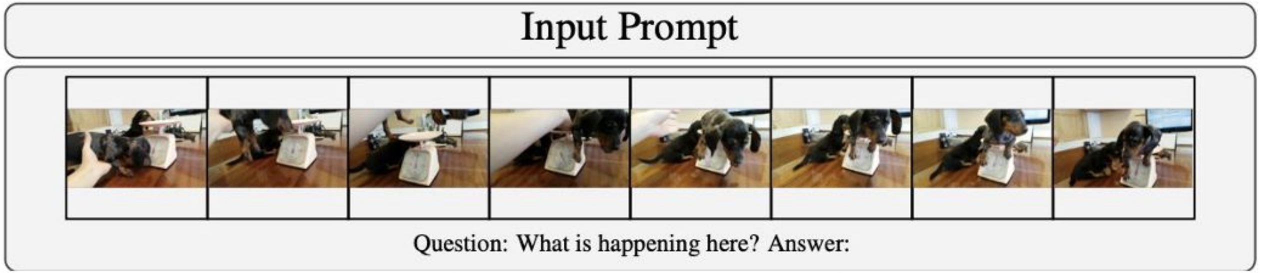
Aghajanyan, Armen, et al. "Cm3: A causal masked multimodal model of the internet." arXiv preprint arXiv:2201.07520 (2022).

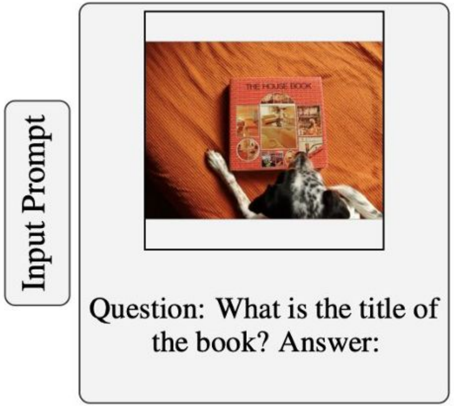
Flamingo:



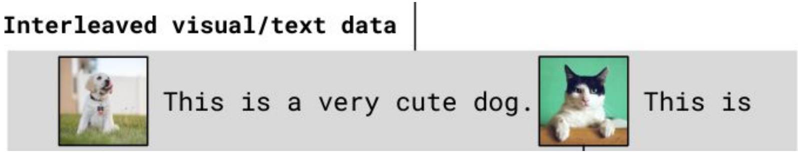
Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

Flamingo Overview



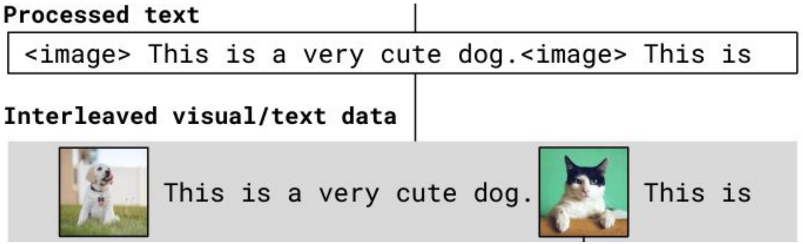


Input Styles



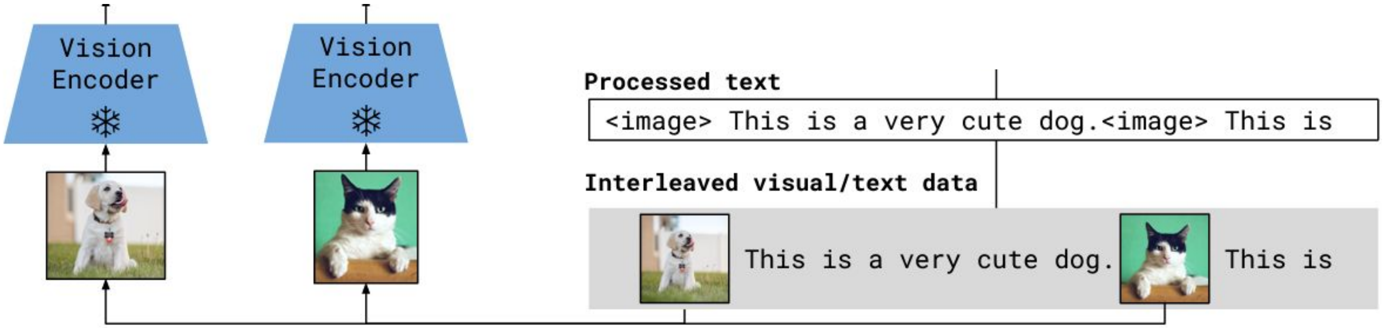
Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

Flamingo Overview



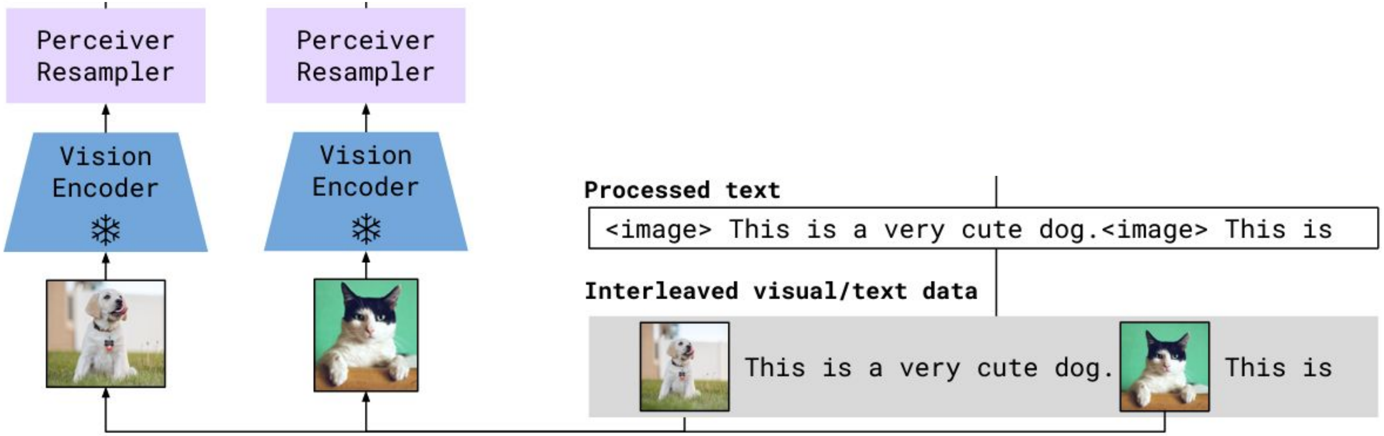
Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

Flamingo Overview



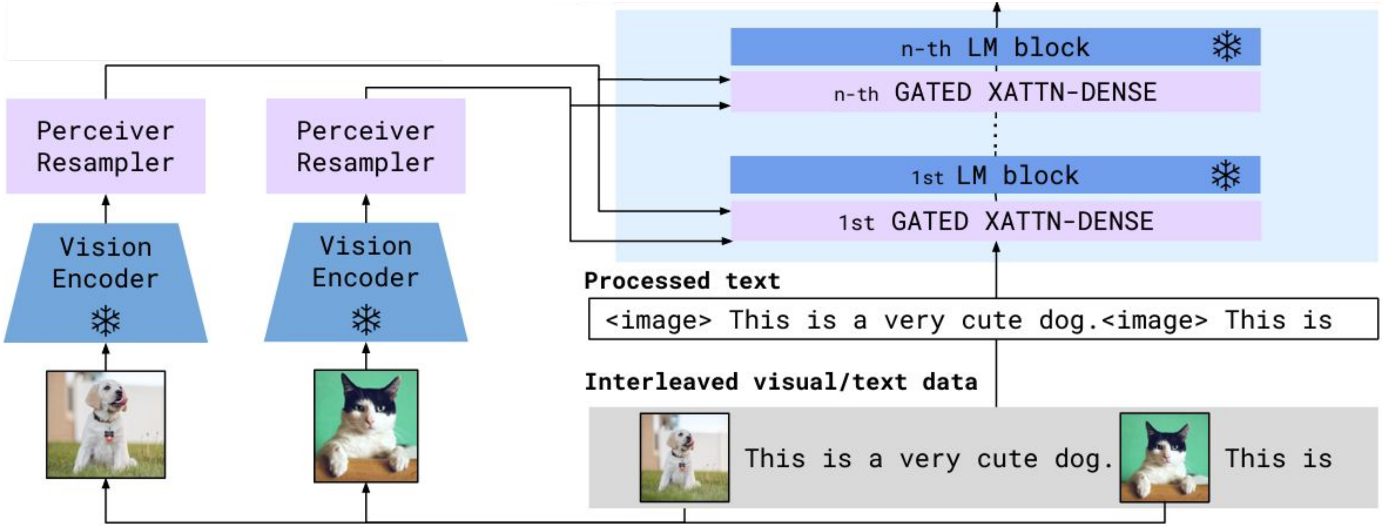
Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

Flamingo Overview



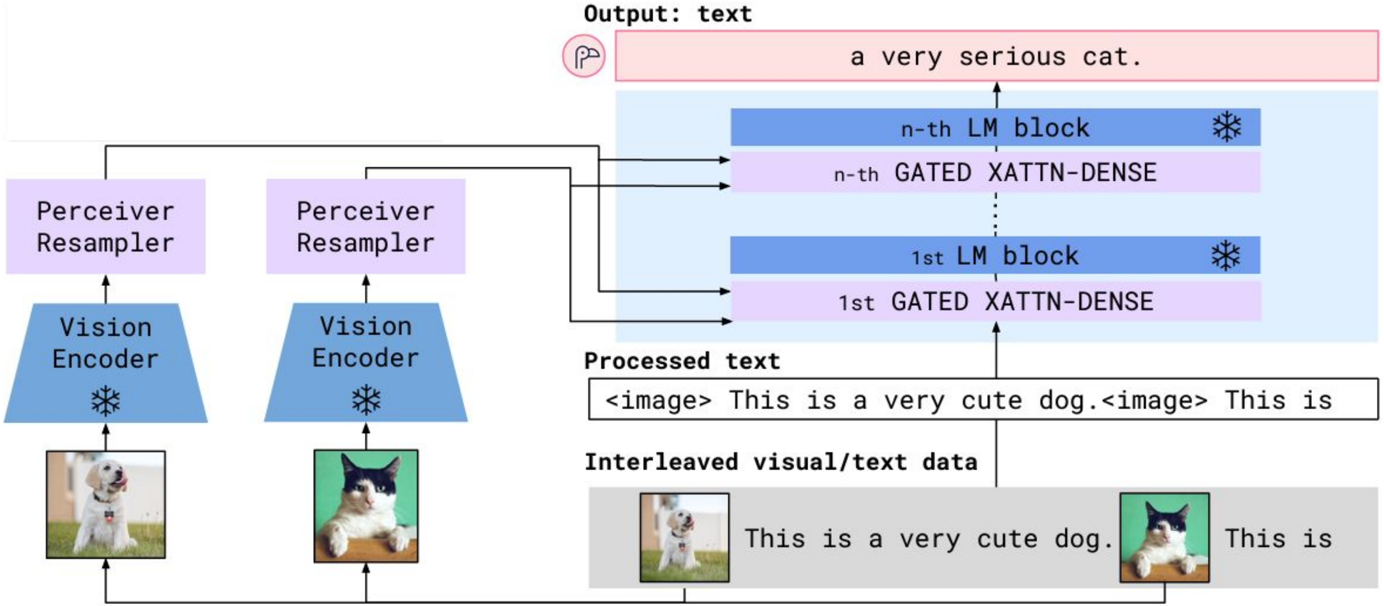
Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

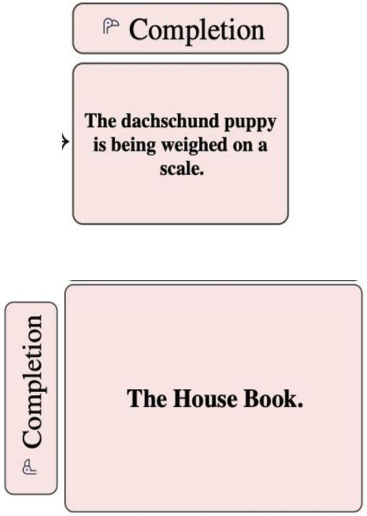
Flamingo Overview



Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

Flamingo Overview

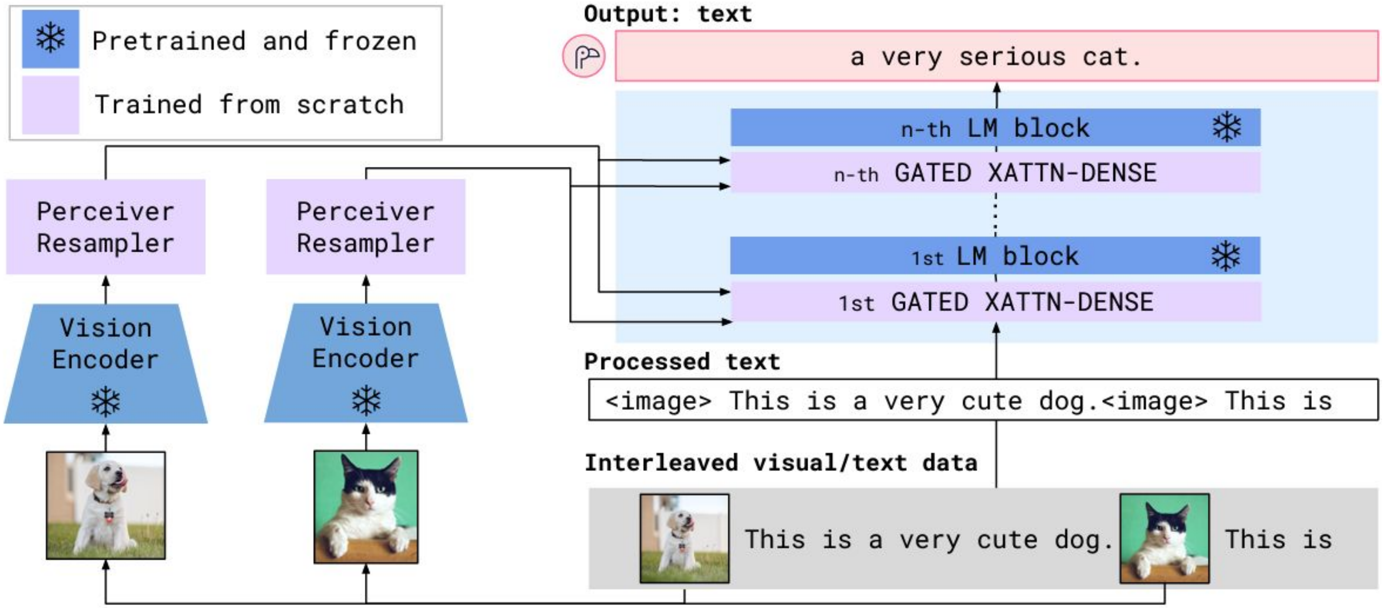




Generic Text output

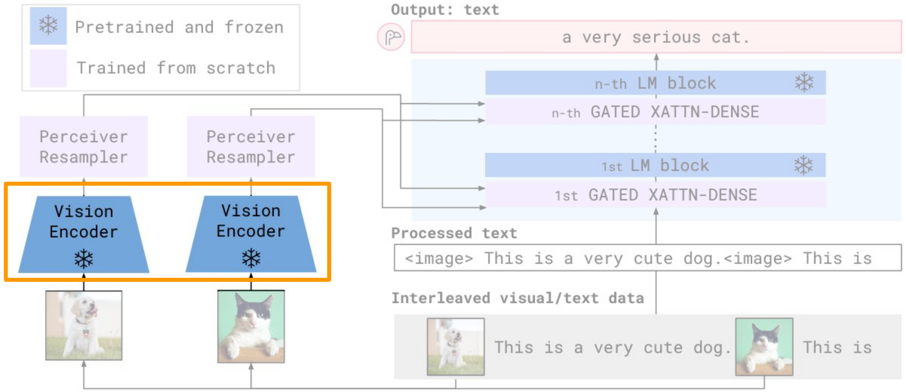
Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

Flamingo Overview



Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

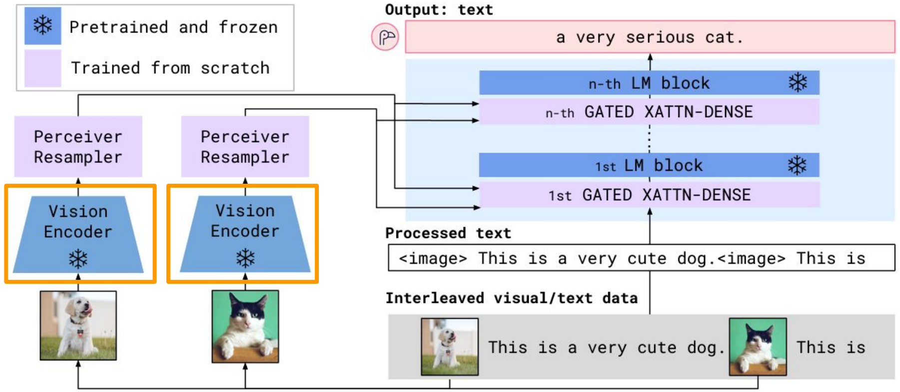
Vision Encoder



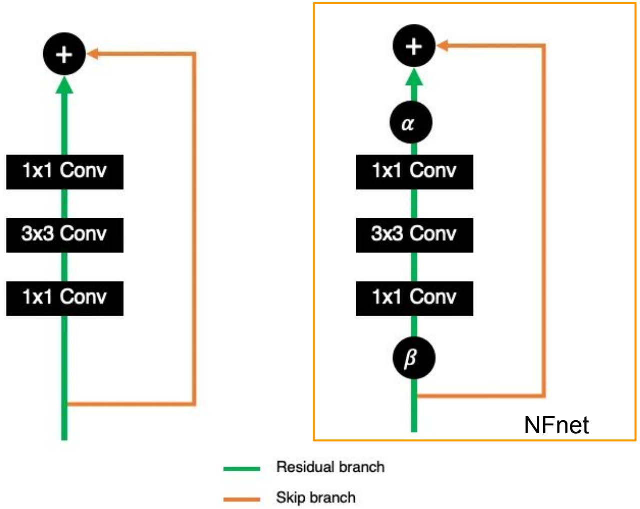
- **NFNet F6[8]** pretrained using the CLIP contrastive loss.

Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

Vision Encoder

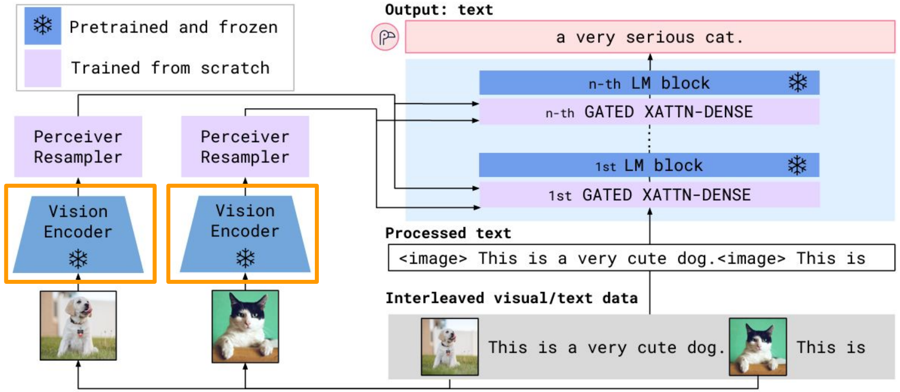


- **NFNet F6[8]** pretrained using the CLIP contrastive loss.



Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).
https://towardsdatascience.com/nfnets-explained-deepminds-new-state-of-the-art-image-classifier-10430c8599ee

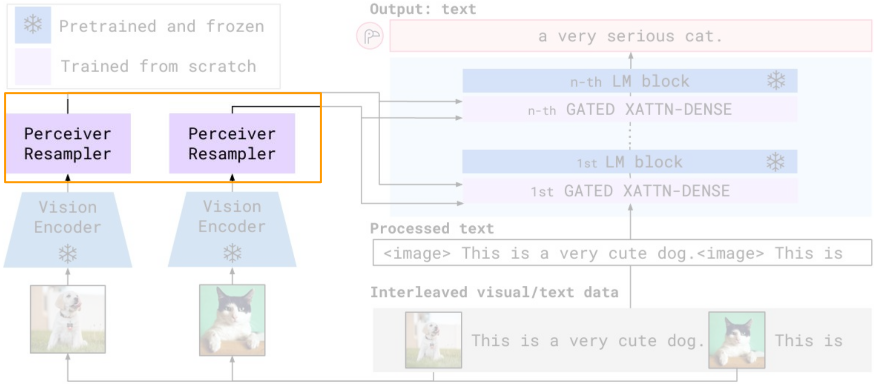
Vision Encoder



- **NFNet F6[8]** pretrained using the CLIP contrastive loss.
- Trained on ALIGN and LTIP
- **Input:** 288 x 288 image
- **Output:** 2D grid Flattened to 1D
- 1FPS sampling for Videos
- Model is Frozen After Pretraining

Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

Perceiver Resampler



Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

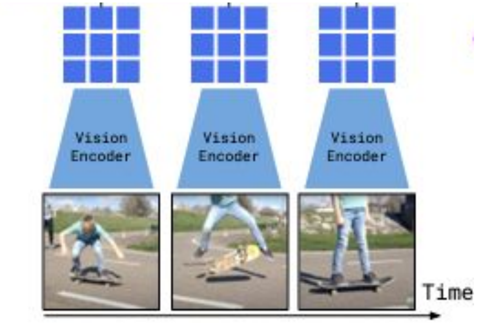
Perceiver Resampler

- Consumes **variable** number of input frames

Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

Perceiver Resampler

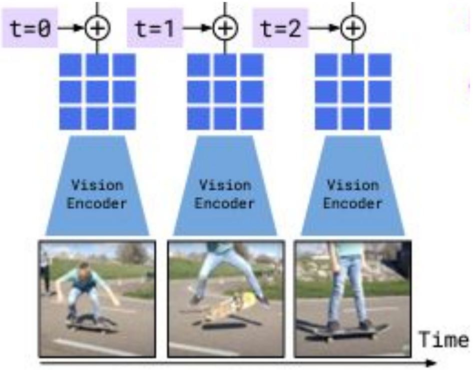
- Consumes **variable** number of input frames



Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

Perceiver Resampler

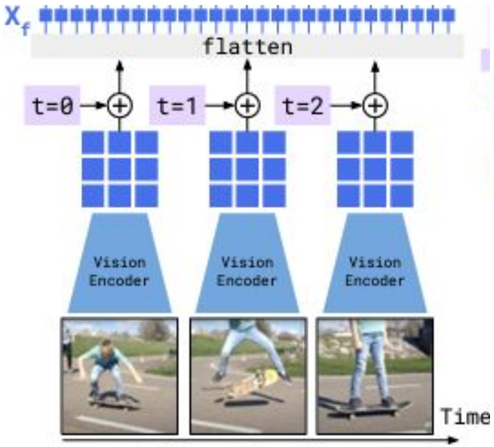
- Consumes **variable** number of input frames.
- Appends **temporal** encodings.



Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

Perceiver Resampler

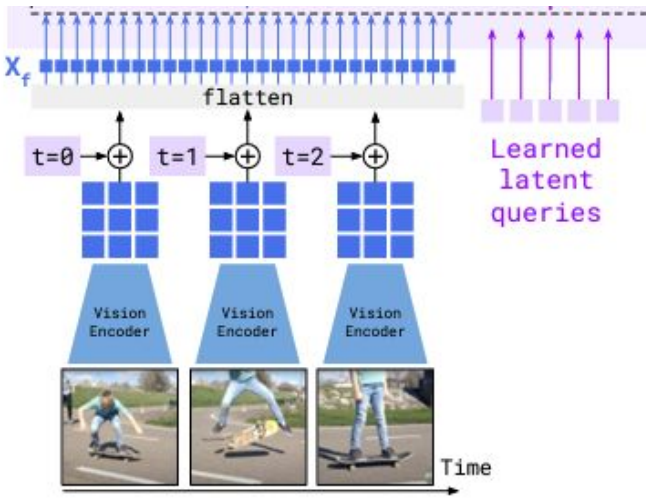
- Consumes **variable** number of input frames.
- Appends **temporal** encodings.
- Flattens the image grid.



Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

Perceiver Resampler

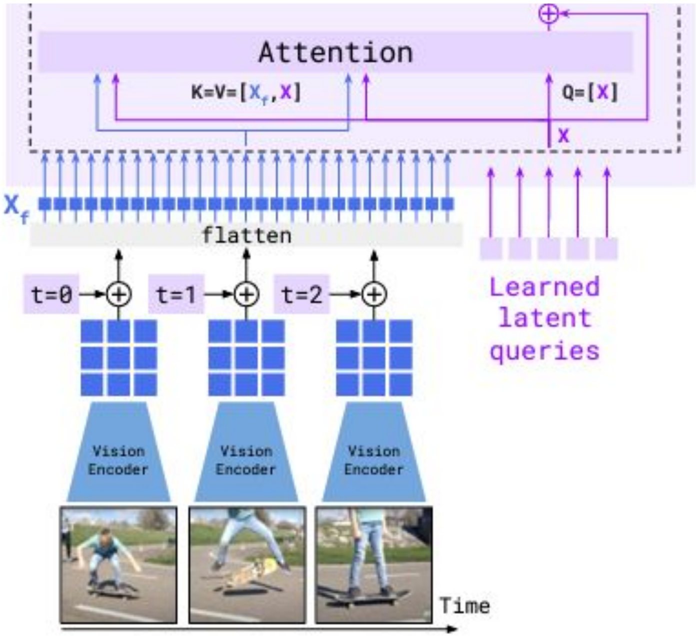
- Consumes **variable** number of input frames.
- Appends **temporal** encodings.
- Flattens the image grid.
- Combined with **fixed number** of latent queries



Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

Perceiver Resampler

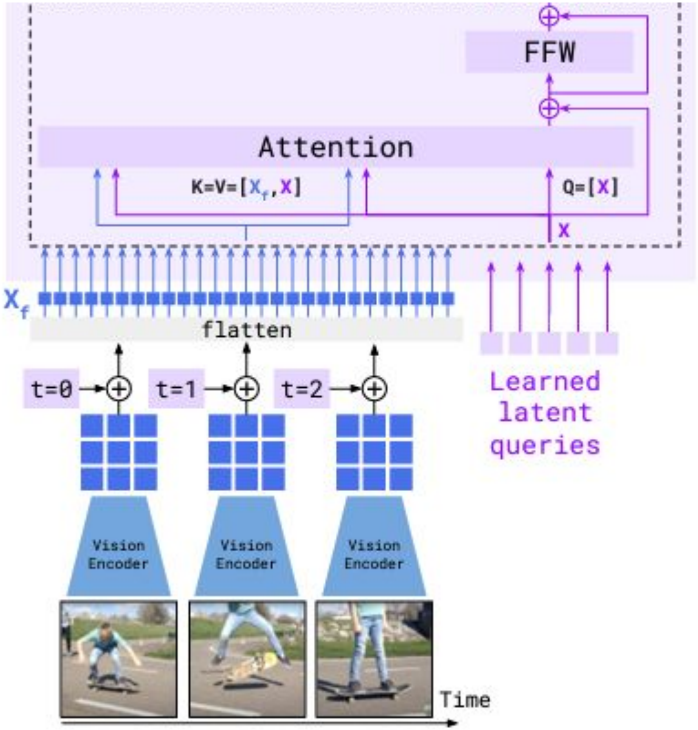
- Consumes **variable** number of input Frames.
- Appends **temporal** encodings.
- Flattens the image grid.
- Combined with **fixed number** of latent queries.
- Attention layer with Q = latent Queries, and K,V = [Image vector, latent Queries]



Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

Perceiver Resampler

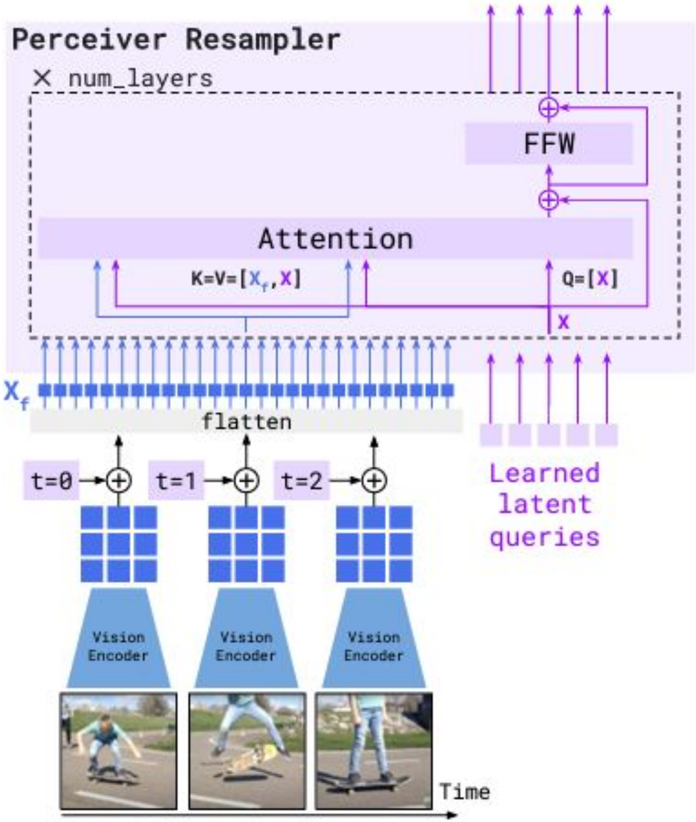
- Consumes **variable** number of input Frames.
- Appends **temporal** encodings.
- Flattens the image grid.
- Combined with **fixed number** of latent queries.
- Attention layer with Q = latent Queries, and K,V = [Image vector, latent Queries]



Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

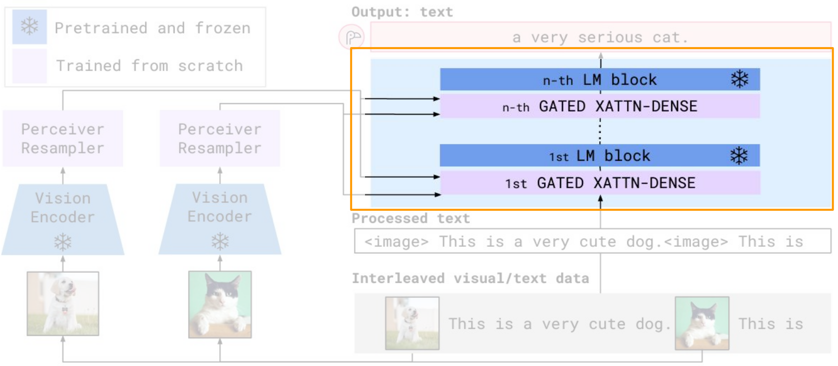
Perceiver Resampler

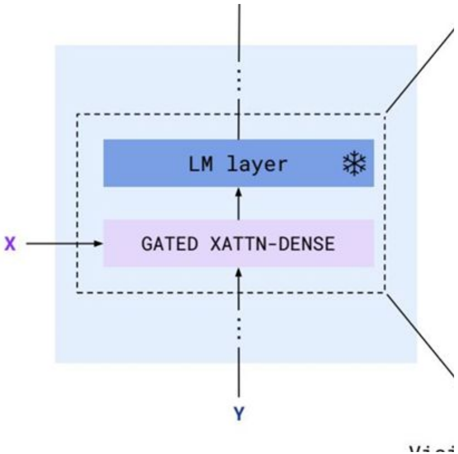
- Consumes **variable** number of input Frames.
- Appends **temporal** encodings.
- Flattens the image grid.
- Combined with **fixed number** of latent queries.
- Attention layer with Q = latent Queries, and K,V = [Image vector, latent Queries]
- Outputs a Fixed number of visual tokens



Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

Conditioning the Language model

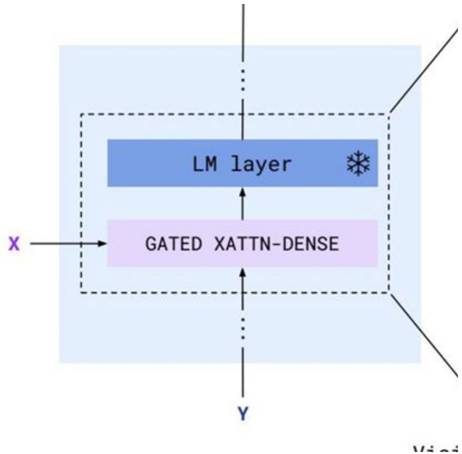




Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

Conditioning the Language model

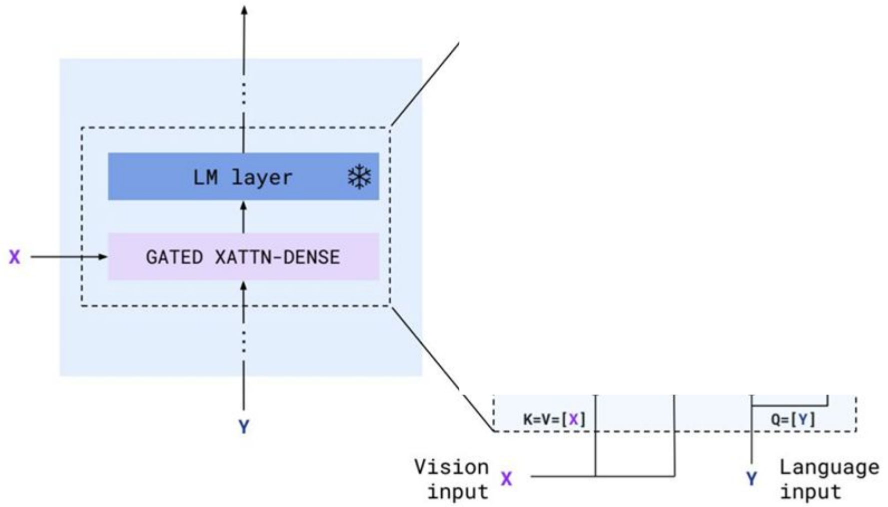
- Flamingo uses **Chinchilla** class of LLMs.



Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

Conditioning the Language model

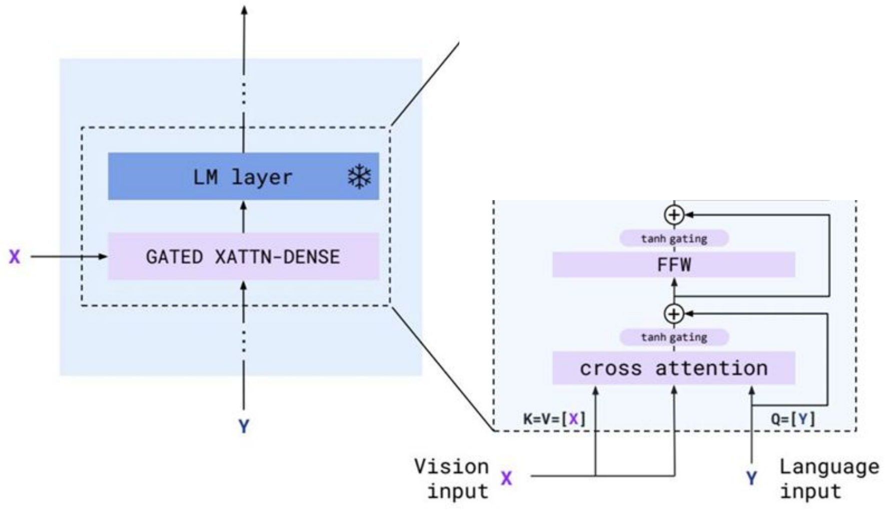
- Flamingo uses **Chinchilla** class of LLMs.
- Vision (X) and language (Y) input to a **XATTN** block



Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

Conditioning the Language model

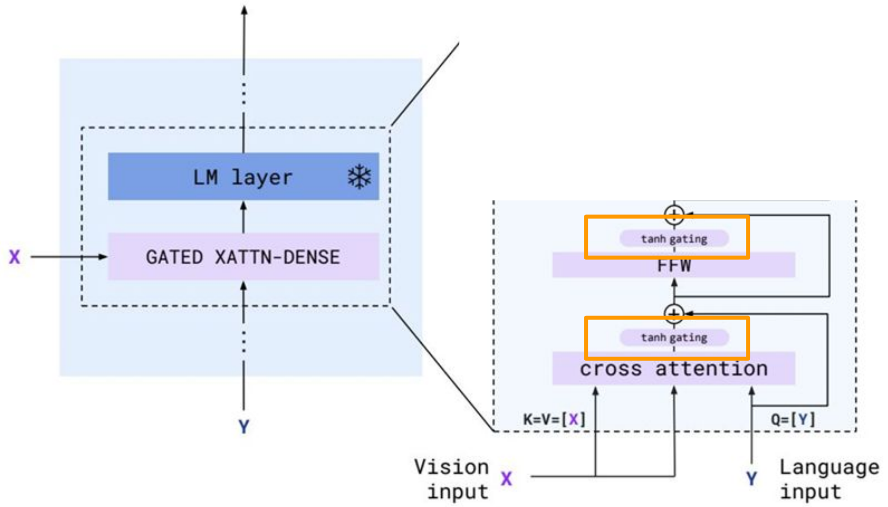
- Flamingo uses **Chinchilla** class of LLMs.
- Vision (X) and language (Y) input to a **XATTN** block



Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

Conditioning the Language model

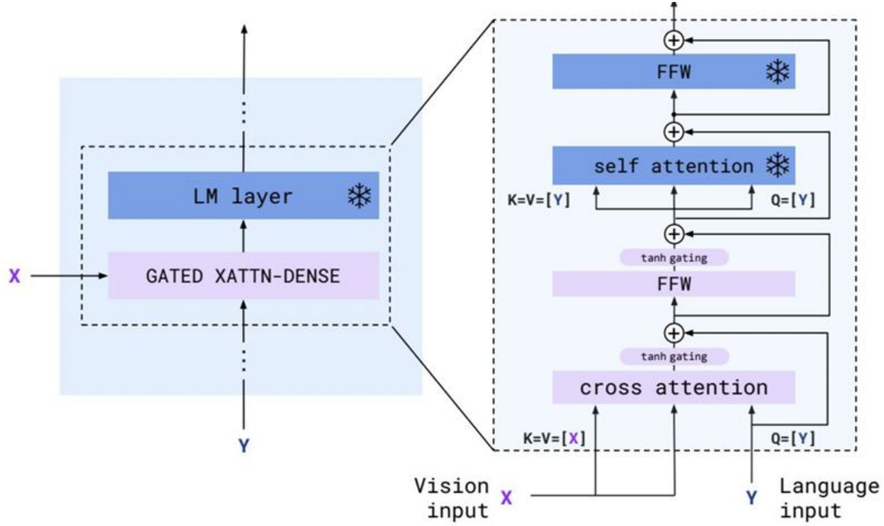
- Flamingo uses **Chinchilla** class of LLMs.
- Vision (X) and language (Y) input to a **XATTN** block
- Uses TanH gating with layer learnable **alpha**
- **Alpha** initialized to 0 for stability



Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

Conditioning the Language model

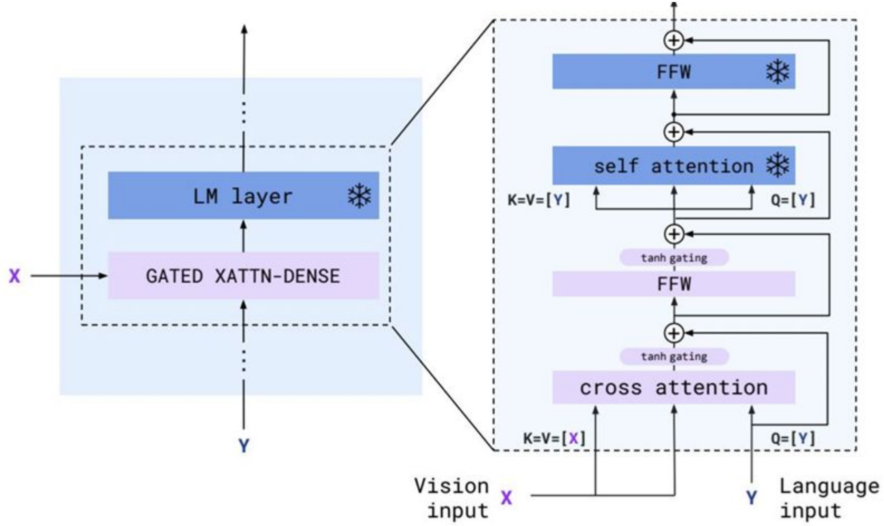
- Flamingo uses **Chinchilla** class of LLMs.
- Vision (X) and language (Y) input to a **XATTN** block
- Uses TanH gating with layer learnable **alpha**
- **Alpha** initialized to 0 for stability.



Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

Conditioning the Language model

- Flamingo uses **Chinchilla** class of LLMs.
- Vision (X) and language (Y) input to a **XATTN** block
- Uses TanH gating with layer learnable **alpha**
- **Alpha** initialized to 0 for stability.
- Flamingo model variations are introduced through XATTN layers only.



Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

Multi-visual input support

- Tags are added to input text.

Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

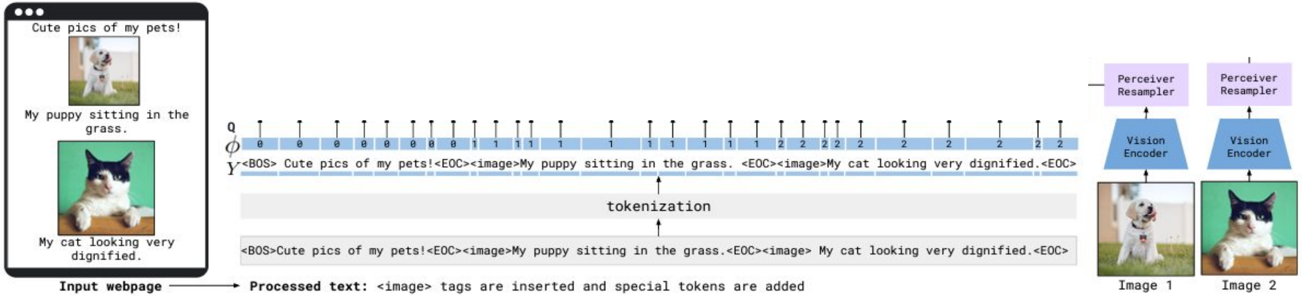
Multi-visual input support

- Tags are added to input text.
- Images are processed.

Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

Multi-visual input support

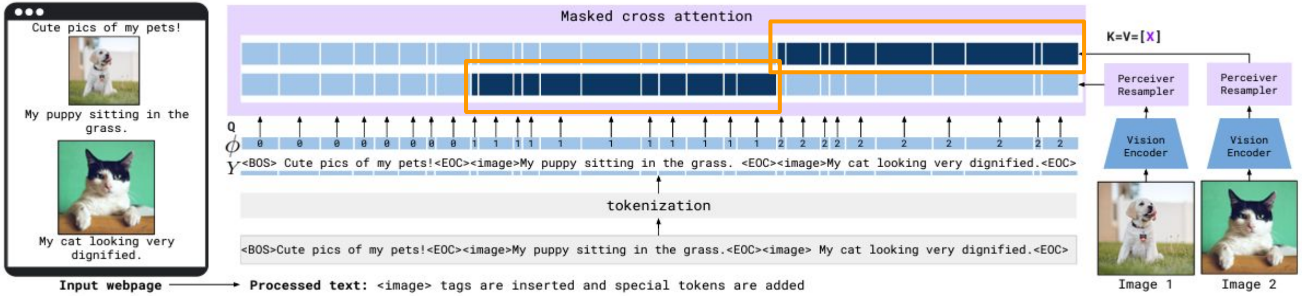
- Tags are added to input text.
- Images are processed.
- Function ϕ that maps each text token to the last image token.



Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

Multi-visual input support

- Tags are added to input text.
- Images are processed.
- Function ϕ that maps each text token to the last image token.
- Each token only attends to the last seen image token



Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

Training 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]

- **ALIGN:** 1.8B pairs with 12.4 tokens on average
- **LTIP:** 312M pairs with 20.5 tokens on average

VTP Dataset:

- 27M short Videos
- 22S duration on average

M3W Dataset:

- 185M Images
- 182GB of Text

Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).

Training

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

- Flamingo is trained by minimizing log likelihood of text given the previous input (text or image)
- The loss is weighted sum of all the datasets, where Dm and λm are the mth dataset and its weight.

Flamingo tasks

	Dataset	DEV	Gen.	Custom prompt	Task description	Eval set	Metric
Image	ImageNet-1k [94]	✓			Object classification	Val	Top-1 acc.
	MS-COCO [15]	✓	✓		Scene description	Test	CIDEr
	VQAv2 [3]	✓	✓		Scene understanding QA	Test-dev	VQA acc. [3]
	OKVQA [69]	✓	✓		External knowledge QA	Val	VQA acc. [3]
	Flickr30k [139]		✓		Scene description	Test (Karpathy)	CIDEr
	VizWiz [35]		✓		Scene understanding QA	Test-dev	VQA acc. [3]
	TextVQA [100]		✓		Text reading QA	Val	VQA acc. [3]
	VisDial [20]				Visual Dialogue	Val	NDCG
	HatefulMemes [54]			✓	Meme classification	Seen Test	ROC AUC
Video	Kinetics700 2020 [102]	✓			Action classification	Val	Top-1/5 avg
	VATEX [122]	✓	✓		Event description	Test	CIDEr
	MSVDQA [130]	✓	✓		Event understanding QA	Test	Top-1 acc.
	YouCook2 [149]		✓		Event description	Val	CIDEr
	MSRVTTQA [130]		✓		Event understanding QA	Test	Top-1 acc.
	iVQA [135]		✓		Event understanding QA	Test	iVQA acc. [135]
	RareAct [73]			✓	Composite action retrieval	Test	mWAP
	NextQA [129]		✓		Temporal/Causal QA	Test	WUPS
	STAR [128]				Multiple-choice QA	Test	Top-1 acc.

Zeroshot and few shot results

Method	FT	Shot	OKVQA (I)	VQAv2 (I)	COCO (I)	MSVDQA (V)	VATEX (V)	VizWiz (I)	Flick30K (I)	MSRVTTQA (V)	iVQA (V)	YouCook2 (V)	STAR (V)	VisDial (I)	TextVQA (I)	NextQA (I)	HatefulMemes (I)	RareAct (V)
Zero/Few shot SOTA	✗	(X)	[34] 43.3 (16)	[114] 38.2 (4)	[124] 32.2 (0)	[58] 35.2 (0)	-	-	-	[58] 19.2 (0)	[135] 12.2 (0)	-	[143] 39.4 (0)	[79] 11.6 (0)	-	-	[85] 66.1 (0)	[85] 40.7 (0)
Flamingo-3B	✗	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
	✗	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	-
	✗	32	45.9	57.1	99.0	42.6	59.2	45.5	71.2	25.6	37.7	76.7	41.6	47.3	30.6	26.1	56.3	-
Flamingo-9B	✗	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
	✗	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	-
	✗	32	51.0	60.4	106.3	47.2	57.4	44.0	72.8	29.4	40.7	77.3	41.2	50.4	32.6	28.4	63.5	-
Flamingo	✗	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**
	✗	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	-
	✗	32	**57.8**	**67.6**	**113.8**	**52.3**	**65.1**	**49.8**	**75.4**	**31.0**	**45.3**	**86.8**	42.2	**55.6**	**37.9**	**33.5**	**70.0**	-
Pretrained FT SOTA	✔	(X)	54.4 [34] (10K)	80.2 [140] (444K)	143.3 [124] (500K)	47.9 [28] (27K)	76.3 [153] (500K)	57.2 [65] (20K)	67.4 [150] (30K)	46.8 [51] (130K)	35.4 [135] (6K)	138.7 [132] (10K)	36.7 [128] (46K)	75.2 [79] (123K)	54.7 [137] (20K)	25.2 [129] (38K)	79.1 [62] (9K)	-

Zeroshot and few shot results

Method	FT	Shot	OKVQA (I)	VQAv2 (I)	COCO (I)	MSVDQA (V)	VATEX (V)	VizWiz (I)	Flick30K (I)	MSRVTTQA (V)	iVQA (V)	YouCook2 (V)	STAR (V)	VisDial (I)	TextVQA (I)	NextQA (I)	HatefulMemes (I)	RareAct (V)
Zero/Few shot SOTA	✗	(X)	[34] 43.3 (16)	[114] 38.2 (4)	[124] 32.2 (0)	[58] 35.2 (0)	-	-	-	[58] 19.2 (0)	[135] 12.2 (0)	-	[143] 39.4 (0)	[79] 11.6 (0)	-	-	[85] 66.1 (0)	[85] 40.7 (0)
Flamingo-3B	✗	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
	✗	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	-
	✗	32	45.9	57.1	99.0	42.6	59.2	45.5	71.2	25.6	37.7	76.7	41.6	47.3	30.6	26.1	56.3	-
Flamingo-9B	✗	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
	✗	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	-
	✗	32	51.0	60.4	106.3	47.2	57.4	44.0	72.8	29.4	40.7	77.3	41.2	50.4	32.6	28.4	63.5	-
Flamingo	✗	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**
	✗	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	-
	✗	32	**57.8**	**67.6**	**113.8**	**52.3**	**65.1**	**49.8**	**75.4**	**31.0**	**45.3**	**86.8**	42.2	**55.6**	**37.9**	**33.5**	**70.0**	-
Pretrained FT SOTA	✔	(X)	54.4 [34] (10K)	80.2 [140] (444K)	143.3 [124] (500K)	47.9 [28] (27K)	76.3 [153] (500K)	57.2 [65] (20K)	67.4 [150] (30K)	46.8 [51] (130K)	35.4 [135] (6K)	138.7 [132] (10K)	36.7 [128] (46K)	75.2 [79] (123K)	54.7 [137] (20K)	25.2 [129] (38K)	79.1 [62] (9K)	-

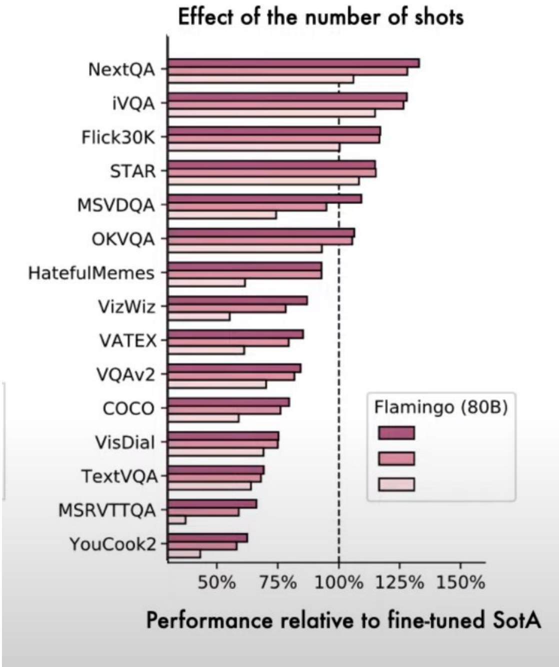
Flamingo is better than current SOTA few shot/zero shot

Zeroshot and few shot results

Method	FT	Shot	OKVQA (I)	VQAv2 (I)	COCO (I)	MSVDQA (V)	VATEX (V)	VizWiz (I)	Flick30K (I)	MSRVTTQA (V)	iVQA (V)	YouCook2 (V)	STAR (V)	VisDial (I)	TextVQA (I)	NextQA (I)	HatefulMemes (I)	RareAct (V)
Zero/Few shot SOTA	✗	(X)	[34] 43.3 (16)	[114] 38.2 (4)	[124] 32.2 (0)	[58] 35.2 (0)	-	-	-	[58] 19.2 (0)	[135] 12.2 (0)	-	[143] 39.4 (0)	[79] 11.6 (0)	-	-	[85] 66.1 (0)	[85] 40.7 (0)
Flamingo-3B	✗	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
	✗	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	-
	✗	32	45.9	57.1	99.0	42.6	59.2	45.5	71.2	25.6	37.7	76.7	41.6	47.3	30.6	26.1	56.3	-
Flamingo-9B	✗	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
	✗	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	-
	✗	32	51.0	60.4	106.3	47.2	57.4	44.0	72.8	29.4	40.7	77.3	41.2	50.4	32.6	28.4	63.5	-
Flamingo	✗	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**
	✗	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	-
	✗	32	**57.8**	**67.6**	**113.8**	**52.3**	**65.1**	**49.8**	**75.4**	**31.0**	**45.3**	**86.8**	42.2	**55.6**	**37.9**	**33.5**	**70.0**	-
Pretrained FT SOTA	✔	(X)	54.4 [34] (10K)	80.2 [140] (444K)	143.3 [124] (500K)	47.9 [28] (27K)	76.3 [153] (500K)	57.2 [65] (20K)	67.4 [150] (30K)	46.8 [51] (130K)	35.4 [135] (6K)	138.7 [132] (10K)	36.7 [128] (46K)	75.2 [79] (123K)	54.7 [137] (20K)	25.2 [129] (38K)	79.1 [62] (9K)	-

It achieves SOTA on 6 tasks

Zeroshot and few shot results



Performance increases generally when the number of shots are increased.

https://samuelalbanie.com/digests/2022-05-flamingo/

Finetuning 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**

Ablation studies

	Ablated setting	*Flamingo*-3B original value	Changed value	Param. count ↓	Step time ↓	COCO CIDEr↑	OKVQA top1↑	VQAv2 top1↑	MSVDQA top1↑	VATEX CIDEr↑	Overall score↑
		***Flamingo*-3B model**		3.2B	1.74s	86.5	42.1	55.8	36.3	53.4	**70.7**
(i)	Training data	All data	w/o Video-Text pairs	3.2B	1.42s	84.2	43.0	53.9	34.5	46.0	67.3
			w/o Image-Text pairs	3.2B	0.95s	66.3	39.2	51.6	32.0	41.6	60.9
			Image-Text pairs→ LAION	3.2B	1.74s	79.5	41.4	53.5	33.9	47.6	66.4
			w/o M3W	3.2B	1.02s	54.1	36.5	52.7	31.4	23.5	53.4
(ii)	Optimisation	Accumulation	Round Robin	3.2B	1.68s	76.1	39.8	52.1	33.2	40.8	62.9
(iii)	Tanh gating	✓	✗	3.2B	1.74s	78.4	40.5	52.9	35.9	47.5	66.5
(iv)	Cross-attention architecture	GATED XATTN-DENSE	VANILLA XATTN	2.4B	1.16s	80.6	41.5	53.4	32.9	50.7	66.9
			GRAFTING	3.3B	1.74s	79.2	36.1	50.8	32.2	47.8	63.1
(v)	Cross-attention frequency	Every	Single in middle	2.0B	0.87s	71.5	38.1	50.2	29.1	42.3	59.8
			Every 4th	2.3B	1.02s	82.3	42.7	55.1	34.6	50.8	68.8
			Every 2nd	2.6B	1.24s	83.7	41.0	55.8	34.5	49.7	68.2
(vi)	Resampler	Perceiver	MLP	3.2B	1.85s	78.6	42.2	54.7	35.2	44.7	66.6
			Transformer	3.2B	1.81s	83.2	41.7	55.6	31.5	48.3	66.7
(vii)	Vision encoder	NFNet-F6	CLIP ViT-L/14	3.1B	1.58s	76.5	41.6	53.4	33.2	44.5	64.9
			NFNet-F0	2.9B	1.45s	73.8	40.5	52.8	31.1	42.9	62.7
(viii)	Freezing LM	✓	✗ (random init)	3.2B	2.42s	74.8	31.5	45.6	26.9	50.1	57.8
			✗ (pretrained)	3.2B	2.42s	81.2	33.7	47.4	31.0	53.9	62.7

Ablation studies

	Ablated setting	*Flamingo*-3B original value	Changed value	Param. count ↓	Step time ↓	COCO CIDEr↑	OKVQA top1↑	VQAv2 top1↑	MSVDQA top1↑	VATEX CIDEr↑	Overall score↑
	***Flamingo*-3B model**			3.2B	1.74s	86.5	42.1	55.8	36.3	53.4	**70.7**
(i)	Training data	All data	w/o Video-Text pairs	3.2B	1.42s	84.2	43.0	53.9	34.5	46.0	67.3
			w/o Image-Text pairs	3.2B	0.95s	66.3	39.2	51.6	32.0	41.6	60.9
			Image-Text pairs→ LAION	3.2B	1.74s	79.5	41.4	53.5	33.9	47.6	66.4
			w/o M3W	3.2B	1.02s	54.1	36.5	52.7	31.4	23.5	53.4
(ii)	Optimisation	Accumulation	Round Robin	3.2B	1.68s	76.1	39.8	52.1	33.2	40.8	62.9
(iii)	Tanh gating	✓	✗	3.2B	1.74s	78.4	40.5	52.9	35.9	47.5	66.5
(iv)	Cross-attention architecture	GATED XATTN-DENSE	VANILLA XATTN	2.4B	1.16s	80.6	41.5	53.4	32.9	50.7	66.9
			GRAFTING	3.3B	1.74s	79.2	36.1	50.8	32.2	47.8	63.1
(v)	Cross-attention frequency	Every	Single in middle	2.0B	0.87s	71.5	38.1	50.2	29.1	42.3	59.8
			Every 4th	2.3B	1.02s	82.3	42.7	55.1	34.6	50.8	68.8
			Every 2nd	2.6B	1.24s	83.7	41.0	55.8	34.5	49.7	68.2
(vi)	Resampler	Perceiver	MLP	3.2B	1.85s	78.6	42.2	54.7	35.2	44.7	66.6
			Transformer	3.2B	1.81s	83.2	41.7	55.6	31.5	48.3	66.7
(vii)	Vision encoder	NFNet-F6	CLIP ViT-L/14	3.1B	1.58s	76.5	41.6	53.4	33.2	44.5	64.9
			NFNet-F0	2.9B	1.45s	73.8	40.5	52.8	31.1	42.9	62.7
(viii)	Freezing LM	✓	✗ (random init)	3.2B	2.42s	74.8	31.5	45.6	26.9	50.1	57.8
			✗ (pretrained)	3.2B	2.42s	81.2	33.7	47.4	31.0	53.9	62.7

Ablation studies

	Ablated setting	*Flamingo*-3B original value	Changed value	Param. count ↓	Step time ↓	COCO CIDEr↑	OKVQA top1↑	VQAv2 top1↑	MSVDQA top1↑	VATEX CIDEr↑	Overall score↑
		***Flamingo*-3B model**		3.2B	1.74s	86.5	42.1	55.8	36.3	53.4	**70.7**
(i)	Training data	All data	w/o Video-Text pairs	3.2B	1.42s	84.2	43.0	53.9	34.5	46.0	67.3
			w/o Image-Text pairs	3.2B	0.95s	66.3	39.2	51.6	32.0	41.6	60.9
			Image-Text pairs→ LAION	3.2B	1.74s	79.5	41.4	53.5	33.9	47.6	66.4
			w/o M3W	3.2B	1.02s	54.1	36.5	52.7	31.4	23.5	53.4
(ii)	Optimisation	Accumulation	Round Robin	3.2B	1.68s	76.1	39.8	52.1	33.2	40.8	62.9
(iii)	Tanh gating	✓	✗	3.2B	1.74s	78.4	40.5	52.9	35.9	47.5	66.5
(iv)	Cross-attention architecture	GATED XATTN-DENSE	VANILLA XATTN	2.4B	1.16s	80.6	41.5	53.4	32.9	50.7	66.9
			GRAFTING	3.3B	1.74s	79.2	36.1	50.8	32.2	47.8	63.1
(v)	Cross-attention frequency	Every	Single in middle	2.0B	0.87s	71.5	38.1	50.2	29.1	42.3	59.8
			Every 4th	2.3B	1.02s	82.3	42.7	55.1	34.6	50.8	68.8
			Every 2nd	2.6B	1.24s	83.7	41.0	55.8	34.5	49.7	68.2
(vi)	Resampler	Perceiver	MLP	3.2B	1.85s	78.6	42.2	54.7	35.2	44.7	66.6
			Transformer	3.2B	1.81s	83.2	41.7	55.6	31.5	48.3	66.7
(vii)	Vision encoder	NFNet-F6	CLIP ViT-L/14	3.1B	1.58s	76.5	41.6	53.4	33.2	44.5	64.9
			NFNet-F0	2.9B	1.45s	73.8	40.5	52.8	31.1	42.9	62.7
(viii)	Freezing LM	✓	✗ (random init)	3.2B	2.42s	74.8	31.5	45.6	26.9	50.1	57.8
			✗ (pretrained)	3.2B	2.42s	81.2	33.7	47.4	31.0	53.9	62.7

Ablation studies

	Ablated setting	*Flamingo*-3B original value	Changed value	Param. count ↓	Step time ↓	COCO CIDEr↑	OKVQA top1↑	VQAv2 top1↑	MSVDQA top1↑	VATEX CIDEr↑	Overall score↑
		***Flamingo*-3B model**		3.2B	1.74s	86.5	42.1	55.8	36.3	53.4	**70.7**
(i)	Training data	All data	w/o Video-Text pairs	3.2B	1.42s	84.2	43.0	53.9	34.5	46.0	67.3
			w/o Image-Text pairs	3.2B	0.95s	66.3	39.2	51.6	32.0	41.6	60.9
			Image-Text pairs→ LAION	3.2B	1.74s	79.5	41.4	53.5	33.9	47.6	66.4
			w/o M3W	3.2B	1.02s	54.1	36.5	52.7	31.4	23.5	53.4
(ii)	Optimisation	Accumulation	Round Robin	3.2B	1.68s	76.1	39.8	52.1	33.2	40.8	62.9
(iii)	Tanh gating	✓	✗	3.2B	1.74s	78.4	40.5	52.9	35.9	47.5	66.5
(iv)	Cross-attention architecture	GATED XATTN-DENSE	VANILLA XATTN	2.4B	1.16s	80.6	41.5	53.4	32.9	50.7	66.9
			GRAFTING	3.3B	1.74s	79.2	36.1	50.8	32.2	47.8	63.1
(v)	Cross-attention frequency	Every	Single in middle	2.0B	0.87s	71.5	38.1	50.2	29.1	42.3	59.8
			Every 4th	2.3B	1.02s	82.3	42.7	55.1	34.6	50.8	68.8
			Every 2nd	2.6B	1.24s	83.7	41.0	55.8	34.5	49.7	68.2
(vi)	Resampler	Perceiver	MLP	3.2B	1.85s	78.6	42.2	54.7	35.2	44.7	66.6
			Transformer	3.2B	1.81s	83.2	41.7	55.6	31.5	48.3	66.7
(vii)	Vision encoder	NFNet-F6	CLIP ViT-L/14	3.1B	1.58s	76.5	41.6	53.4	33.2	44.5	64.9
			NFNet-F0	2.9B	1.45s	73.8	40.5	52.8	31.1	42.9	62.7
(viii)	Freezing LM	✓	✗ (random init)	3.2B	2.42s	74.8	31.5	45.6	26.9	50.1	57.8
			✗ (pretrained)	3.2B	2.42s	81.2	33.7	47.4	31.0	53.9	62.7

Classification 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

Strengths

- The addition of extra layers while keeping the rest of the model frozen preserves knowledge of both models and is novel.
- The method is able to get very impressive results just using few input samples as demonstrations.
- The paper and appendix include a huge number of studies, justifying most of their model decisions, data decisions, parameter choices etc.

Weaknesses

- The spotlight flamingo model that gets the best results is exceptionally big at 80B parameters and making it quite cumbersome to use.
- The authors havent released their model and data, this is not a technical weakness but sets a bad precedent within the research community.
- Flamingo performs worse than its vision encoder on image classification.

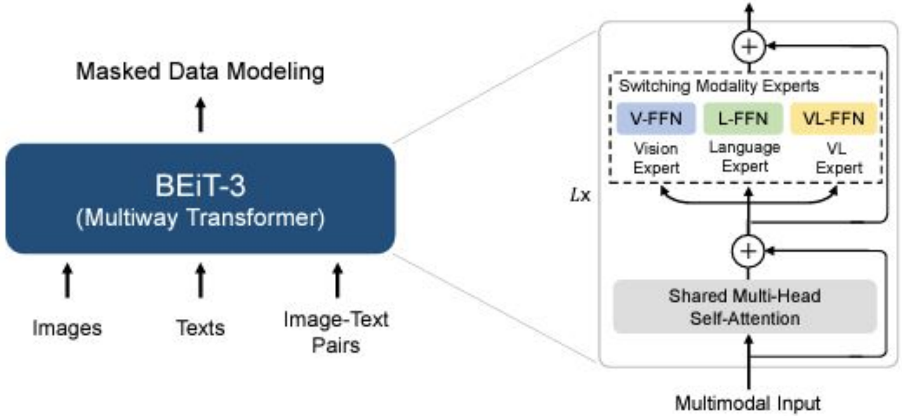
Language Is Not All You Need: Aligning Perception with Language Models

Kosmos-1: Key Takeaways

- A large multimodal LLM that can perceive general modalities, perform zero shot and few shot learning.
- Trained on a web scale multimodal corpora containing interleaved text and images.
- Kosmos-1 demonstrates impressive capabilities across, vision, language and perception language tasks.
- They evaluate on unique tasks like multimodal chain of thought reasoning, OCR free NLP and a novel nonverbal reasoning test.

Related Works

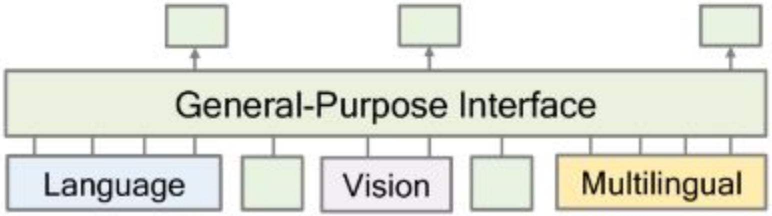
- **MetaLM[2]: LLMs are general purpose interfaces**
  - Any input and output format that can be converted to text token can be a LLM usecase.
- **Extending LLMs to multimodal tasks**
  - Flamingo[3]: Large MLLM for few shot learning
  - BeIT[9]: masked language modelling on images, text, and pairs in a unified manner



Wang, Wenhui, et al. "Image as a foreign language: Beit pretraining for all vision and vision-language tasks." arXiv preprint arXiv:2208.10442 (2022).

Kosmos - 1: Overview

- Kosmos - 1 follows the same philosophy as the MetaLM and treats language models as a universal task layer.
- It builds on MetaLM, trains on more multimodal data, uses interleaved inputs.

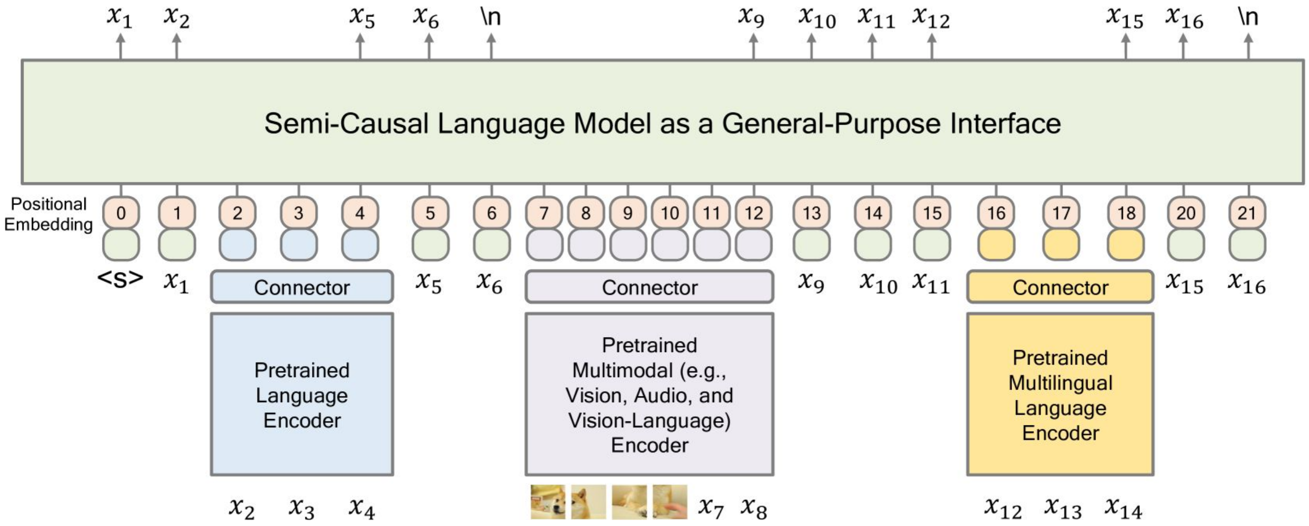


MetaLM

Hao, Yaru, et al. "Language models are general-purpose interfaces." arXiv preprint arXiv:2206.06336 (2022).

Meta LM: Language models are a general purpose Interface

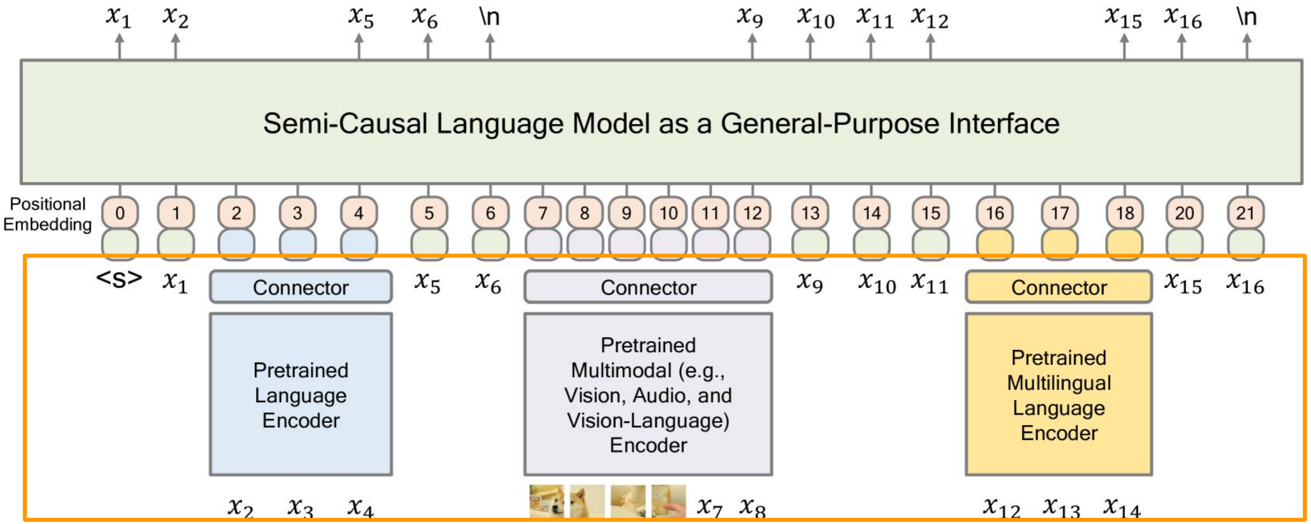
- MetaLM proposes to use LMs as a general interface for all kinds of input like video, images, multilingual etc.



Hao, Yaru, et al. "Language models are general-purpose interfaces." arXiv preprint arXiv:2206.06336 (2022).

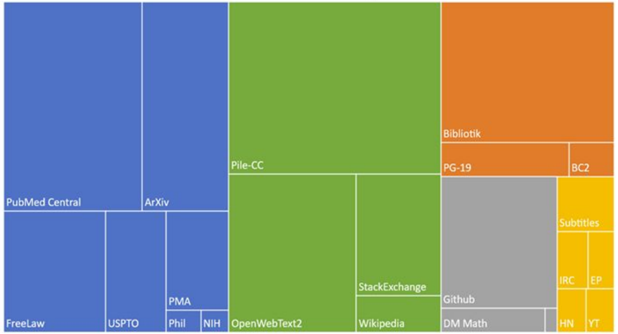
Meta LM: Language models are a general purpose Interface

- MetaLM proposes to use LMs as a general interface for all kinds of input like video, images, multilingual etc.



Hao, Yaru, et al. "Language models are general-purpose interfaces." arXiv preprint arXiv:2206.06336 (2022).

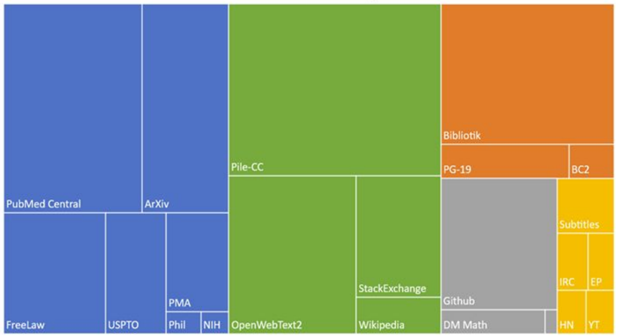
Meta LM: Pretraining and Evaluation



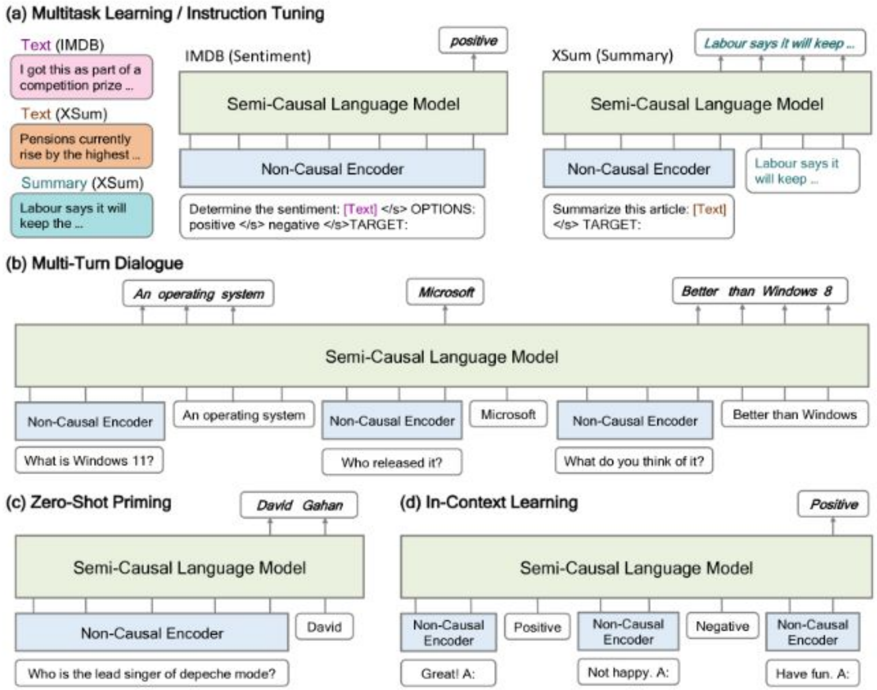
Train on PILE for language only Tasks

Hao, Yaru, et al. "Language models are general-purpose interfaces." arXiv preprint arXiv:2206.06336 (2022).

Meta LM: Pretraining and Evaluation



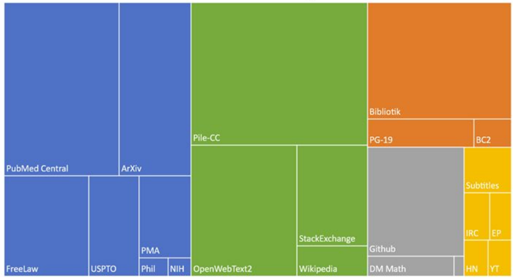
Train on PILE for language only Tasks



Evaluate on a Number of Tasks

Hao, Yaru, et al. "Language models are general-purpose interfaces." arXiv preprint arXiv:2206.06336 (2022).

Meta LM: Pretraining and Evaluation



+

Visual Genome

COCO caption

Conceptual Captions

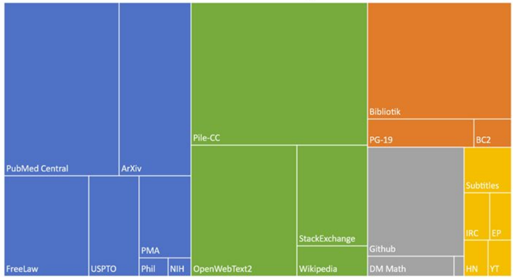
SBU Caption

4M images + 10M Pairs

Train on PILE with Image-caption datasets

Hao, Yaru, et al. "Language models are general-purpose interfaces." arXiv preprint arXiv:2206.06336 (2022).

Meta LM: Pretraining and Evaluation



Visual Genome

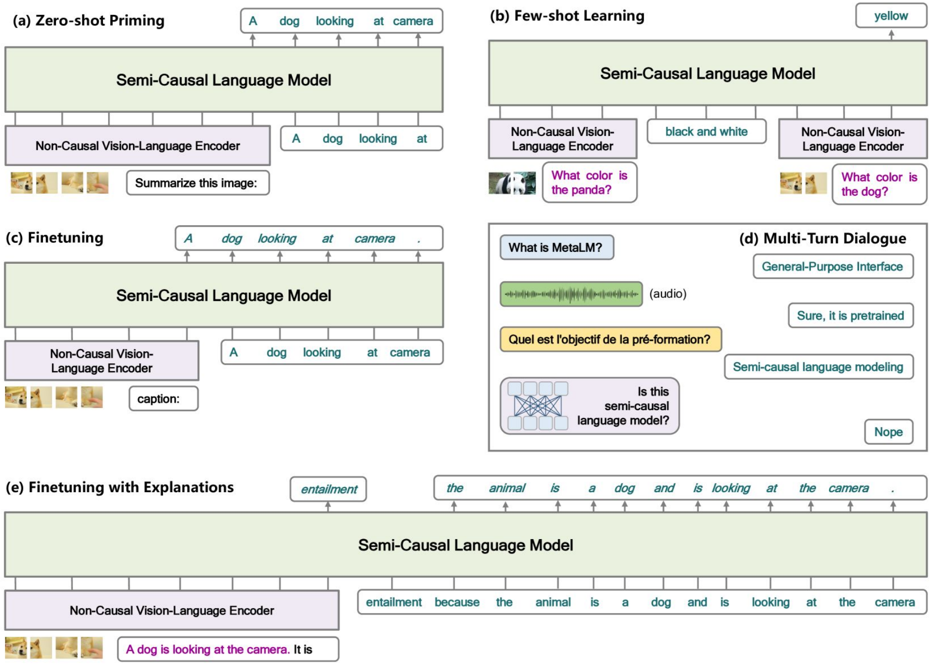
COCO caption

Conceptual Captions

SBU Caption

4M images + 10M Pairs

Train on PILE with Image-caption datasets

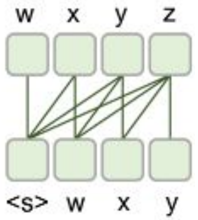


Evaluate on a number of tasks

Hao, Yaru, et al. "Language models are general-purpose interfaces." arXiv preprint arXiv:2206.06336 (2022).

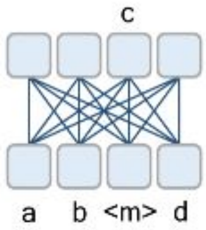
Meta LM: Model Architecture

- Introduces a new semi-causal architecture that jointly learns with a combination of pretrained encoders each focusing on a modality.



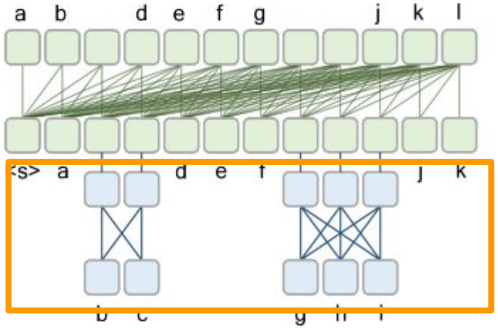
(a) Causal LM
(Unidirectional)

GPT 3 style decoder



(c) Non-Causal LM
(Bidirectional)

BERT style encoder



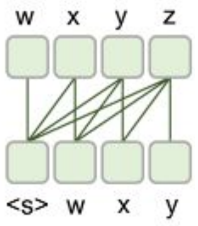
(d) Semi-Causal LM

MetaLM

Hao, Yaru, et al. "Language models are general-purpose interfaces." arXiv preprint arXiv:2206.06336 (2022).

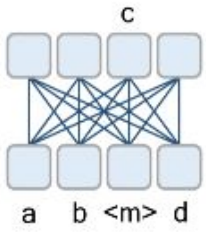
Meta LM: Model Architecture

- Introduces a new semi-causal architecture that jointly learns with a combination of pretrained encoders each focusing on a modality.



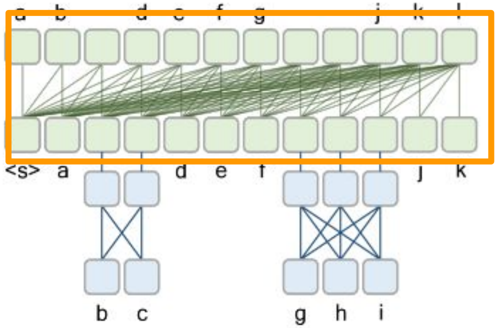
(a) Causal LM
(Unidirectional)

GPT 3 style decoder



(c) Non-Causal LM
(Bidirectional)

BERT style encoder



(d) Semi-Causal LM

MetaLM

Hao, Yaru, et al. "Language models are general-purpose interfaces." arXiv preprint arXiv:2206.06336 (2022).

Meta LM: Loss Function

- Trains on a semi causal modelling objective, where the model predicts the text token given the representations of previous tokens from bidirectional encoders.

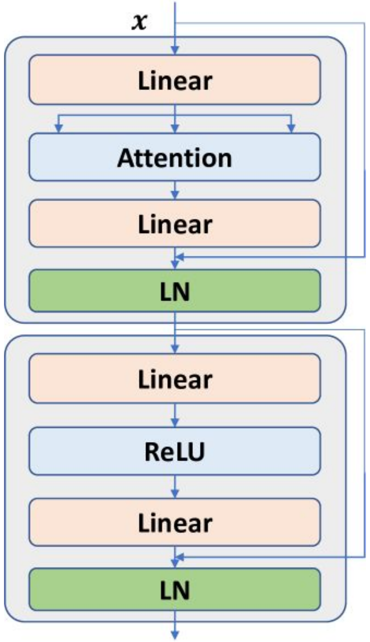
$$\max \sum_{i=0}^{k} \sum_{t=e_i}^{s_{(i+1)}} \log P(x_t \mid \boldsymbol{x}_{<t}, \{\boldsymbol{h}(\boldsymbol{x}_{s_j}^{e_j})\}_{j<i})$$

Here H(X) is the encoder function for each span which can be text or image.

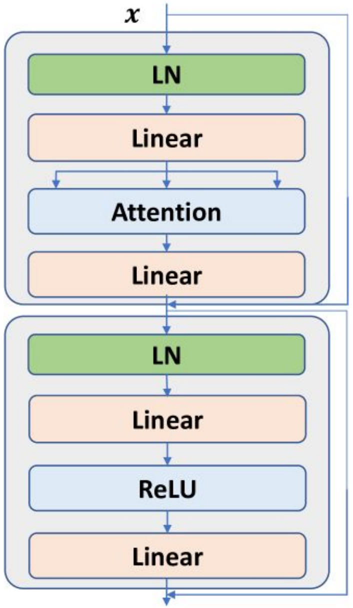
Going from MetaLM to Kosmos-1

- Kosmos-1 trains a model with 24 layers with a hidden dimension size of 2048 and 32 attention heads, totalling to 1.3B parameters similar to MetaLM.
- They change the default transformer module to the **magento[10]** module.

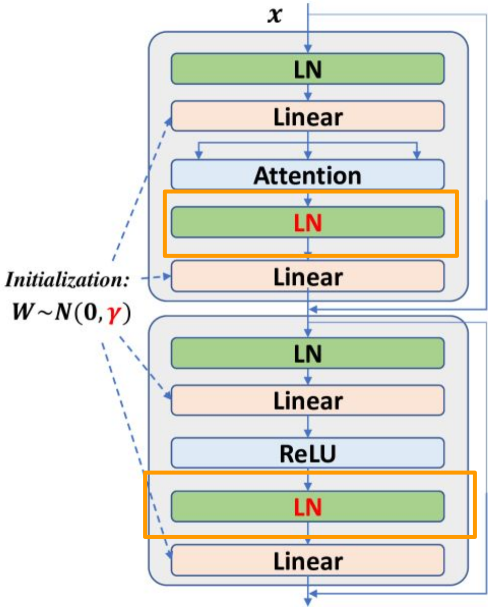
Background: Magento Module vs Other methods



BERT, NMT, Transformer



ViT, GPT



Magneto

Wang, Hongyu, et al. "Foundation transformers." arXiv preprint arXiv:2210.06423 (2022).

Going from MetaLM to Kosmos-1

- Kosmos-1 trains a model with 24 layers with a hidden dimension size of 2048 and 32 attention heads, totalling to 1.3B parameters similar to MetaLM.
- They change the default transformer to the **magento**[10] module
- They use Extrapolatable position embedding (**xPos**)[11] which generalizes better at long term dependencies.

Length	**256**	**512**	**1024**	**2048**	**4096**
	Interpolation			Extrapolation	
Transformer	46.34	36.39	29.94	132.63	1283.79
Alibi	37.66	29.92	24.99	23.14	24.26
Roformer	38.09	30.38	25.52	73.6	294.45
LEX Transformer (Ours)	**34.3**	**27.55**	**23.31**	**21.6**	**20.73**

Going from MetaLM to Kosmos-1

- Kosmos-1 trains a model with 24 layers with a hidden dimension size of 2048 and 32 attention heads, totalling to 1.3B parameters similar to MetaLM.
- They change the default transformer to the **magento**[10] module.
- They use Extrapolatable position embedding (**xPos**)[11] which generalizes better at long term dependencies.
- Trains with the semi causal next token prediction task, minimizing log likelihood.

Kosmos-1: Input Format

Text:

<s> Kosmos-1 can perceive multimodal input, learn in context, and generate output. </s>

Image-Caption:

<s> <image> Image Embedding </image> WALL-E giving potted plant to EVE. </s>

Multimodal:

<s> <image> Image Embedding </image> This is WALL-E. <image> Image Embedding </image> This is EVE. </s>

Kosmos - 1: Other Details

- **Vision Encoder:** CLIP ViT-L/14 that has been frozen except for the last layer.
- **Image preprocessing**: resized to 224 x 224
- **Tokenizer:** SentencePiece
- **Optimizer:** AdamW
- **Batch Size:** 1.2M tokens (0.5M tokens from text corpora, 0.5M tokens from image-caption pairs, and 0.2M tokens from interleaved data)

Kosmos - 1 : Training Data

- **TextData:** Subset of the PILE dataset and common Crawl.

Datasets	Tokens (billion)	Weight (%)	Epochs
OpenWebText2	14.8	21.8%	1.47
CC-2021-04	82.6	17.7%	0.21
Books3	25.7	16.2%	0.63
CC-2020-50	68.7	14.7%	0.21
Pile-CC	49.8	10.6%	0.21
Realnews	21.9	10.2%	0.46
Wikipedia	4.2	5.4%	1.29
BookCorpus2	1.5	1.1%	0.75
Gutenberg (PG-19)	2.7	1.0%	0.38
CC-Stories	5.3	1.0%	0.19
NIH ExPorter	0.3	0.2%	0.75

277.5B Tokens

Kosmos - 1 : Training Data

- **TextData:** Subset of the PILE dataset and common Crawl.
- **Image-caption pairs:** Collection of Several image caption datasets

Datasets	Tokens (billion)	Weight (%)	Epochs
OpenWebText2	14.8	21.8%	1.47
CC-2021-04	82.6	17.7%	0.21
Books3	25.7	16.2%	0.63
CC-2020-50	68.7	14.7%	0.21
Pile-CC	49.8	10.6%	0.21
Realnews	21.9	10.2%	0.46
Wikipedia	4.2	5.4%	1.29
BookCorpus2	1.5	1.1%	0.75
Gutenberg (PG-19)	2.7	1.0%	0.38
CC-Stories	5.3	1.0%	0.19
NIH ExPorter	0.3	0.2%	0.75

277.5B Tokens

+

LAION 2B

LAION 400M

COYO 700M

Conceptual Captions

3.1B pairs

Kosmos - 1 : Training Data

- **TextData:** Subset of the PILE dataset and common Crawl.
- **Image-caption pairs:** Collection of Several image caption datasets
- **Interleaved Data:** documents containing images with text.

Datasets	Tokens (billion)	Weight (%)	Epochs
OpenWebText2	14.8	21.8%	1.47
CC-2021-04	82.6	17.7%	0.21
Books3	25.7	16.2%	0.63
CC-2020-50	68.7	14.7%	0.21
Pile-CC	49.8	10.6%	0.21
Realnews	21.9	10.2%	0.46
Wikipedia	4.2	5.4%	1.29
BookCorpus2	1.5	1.1%	0.75
Gutenberg (PG-19)	2.7	1.0%	0.38
CC-Stories	5.3	1.0%	0.19
NIH ExPorter	0.3	0.2%	0.75

277.5B Tokens

+

LAION 2B

LAION 400M

COYO 700M

Conceptual Captions

3.1B pairs

Interleaved data

71M documents

Kosmos - 1 : Training Data

- **TextData:** Subset of the PILE dataset and common Crawl.
- **Image-caption pairs:** Collection of Several image caption datasets
- **Interleaved Data:** documents containing images with text.
- **Language only instruction data**: Unnatural Instructions and FLANv2

Datasets	Tokens (billion)	Weight (%)	Epochs
OpenWebText2	14.8	21.8%	1.47
CC-2021-04	82.6	17.7%	0.21
Books3	25.7	16.2%	0.63
CC-2020-50	68.7	14.7%	0.21
Pile-CC	49.8	10.6%	0.21
Realnews	21.9	10.2%	0.46
Wikipedia	4.2	5.4%	1.29
BookCorpus2	1.5	1.1%	0.75
Gutenberg (PG-19)	2.7	1.0%	0.38
CC-Stories	5.3	1.0%	0.19
NIH ExPorter	0.3	0.2%	0.75

277.5B Tokens

+

LAION 2B

LAION 400M

COYO 700M

Conceptual Captions

3.1B pairs

Interleaved data

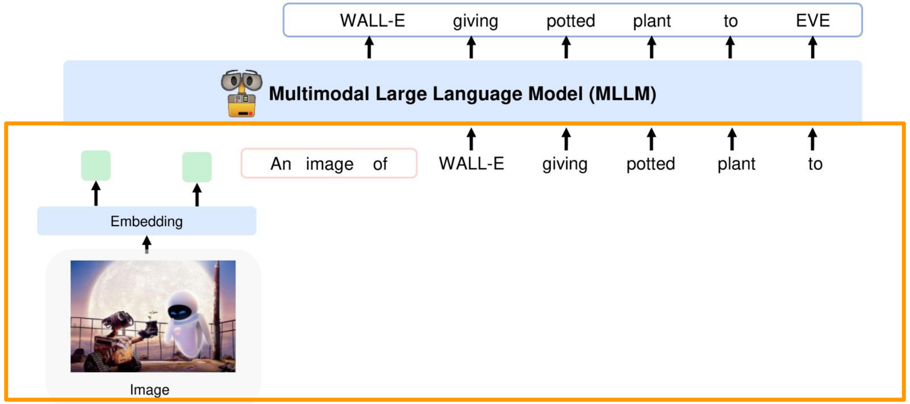
71M documents

Unnatural Instructions

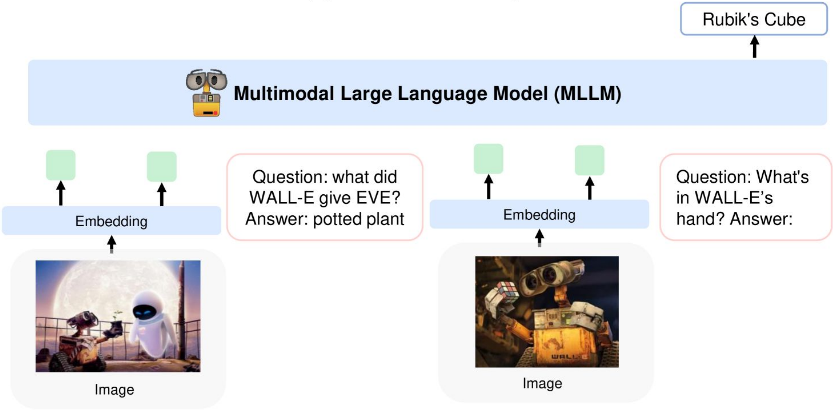
FLANv2

122K pairs

Kosmos-1: Evaluation Format



(a) Zero-shot learning



(b) Few-shot learning

Huang, Shaohan, et al. "Language Is Not All You Need: Aligning Perception with Language Models." arXiv preprint arXiv:2302.14045 (2023).

Evaluation Tasks

Dataset	Task description	Metric	Zero-shot	Few-shot
Language tasks				
StoryCloze [34]	Commonsense reasoning	Accuracy	✓	✓
HellaSwag [61]	Commonsense NLI	Accuracy	✓	✓
Winograd [28]	Word ambiguity	Accuracy	✓	✓
Winogrande [40]	Word ambiguity	Accuracy	✓	✓
PIQA [8]	Physical commonsense	Accuracy	✓	✓
BoolQ [11]	Question answering	Accuracy	✓	✓
CB [16]	Textual entailment	Accuracy	✓	✓
COPA [37]	Causal reasoning	Accuracy	✓	✓
Rendered SST-2 [38]	OCR-free sentiment classification	Accuracy	✓	
HatefulMemes [25]	OCR-free meme classification	ROC AUC	✓	
Cross-modal transfer				
RelativeSize [5]	Commonsense reasoning (object size)	Accuracy	✓	
MemoryColor [36]	Commonsense reasoning (object color)	Accuracy	✓	
ColorTerms [4]	Commonsense reasoning (object color)	Accuracy	✓	
Nonverbal reasoning tasks				
IQ Test	Raven's Progressive Matrices	Accuracy	✓	
Perception-language tasks				
COCO Caption [32]	Image captioning	CIDEr, etc.	✓	✓
Flicker30k [60]	Image captioning	CIDEr, etc.	✓	✓
VQAv2 [18]	Visual question answering	VQA acc.	✓	✓
VizWiz [19]	Visual question answering	VQA acc.	✓	✓
WebSRC [14]	Web page question answering	F1 score	✓	
Vision tasks				
ImageNet [15]	Zero-shot image classification	Top-1 acc.	✓	
CUB [51]	Zero-shot image classification with descriptions	Accuracy	✓	

Image Captioning

Kosmos-1 is able to outperform both 3B and 9B Flamingo models while being only 1.6B

Model	COCO		Flickr30k	
	CIDEr	SPICE	CIDEr	SPICE
ZeroCap	14.6	5.5	-	-
VLKD	58.3	13.4	-	-
FewVLM	-	-	31.0	10.0
MetaLM	82.2	15.7	43.4	11.7
Flamingo-3B*	73.0	-	60.6	-
Flamingo-9B*	79.4	-	61.5	-
Kosmos-1 (1.6B)	**84.7**	**16.8**	**67.1**	**14.5**
Flamingo (80B)	84.3	-	**67. 2**	-

Zero Shot results on COCO caption karpathy split

Image Captioning

Kosmos-1 Performs degrades when number of shots are increased from 4

Model	COCO			Flickr30k		
	$k=2$	$k=4$	$k=8$	$k=2$	$k=4$	$k=8$
Flamingo-3B	-	85.0	90.6	-	72.0	71.7
Flamingo-9B	-	93.1	**99.0**	-	72.6	**73.4**
Kosmos-1 (1.6B)	**99.6**	**101.7**	96.7	**70.0**	**75.3**	68.0

	$K = 32$	$k = 32$
Flamingo - 3B	99.0	71.2
Flamingo - 9B	106.3	72.8

Zero Shot results on COCO caption karpathy split

Visual Question answering

Kosmos-1 outperforms Flamingo on VizWiz

Model	VQAv2	VizWiz
Frozen	29.5	-
VLKDViT-B/16	38.6	-
METALM	41.1	-
Flamingo-3B*	49.2	28.9
Flamingo-9B*	**51.8**	28.8
KOSMOS-1 (1.6B)	51.0	**29.2**
Flamingo (80B)	**56.3**	**31.6**

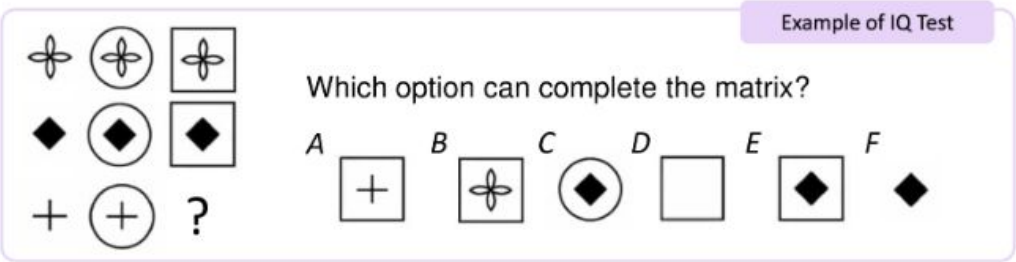
Visual Question Answering

Kosmos - 1 does better in k=2,4 but does poorly with higher K

Model	VQAv2			VizWiz		
	$k=2$	$k=4$	$k=8$	$k=2$	$k=4$	$k=8$
Frozen	-	38.2	-	-	-	-
MetaLM	-	45.3	-	-	-	-
Flamingo-3B	-	53.2	55.4	-	34.4	38.4
Flamingo-9B	-	**56.3**	**58.0**	-	34.9	**39.4**
Kosmos-1 (1.6B)	**51.4**	51.8	51.4	**31.4**	**35.3**	39.0

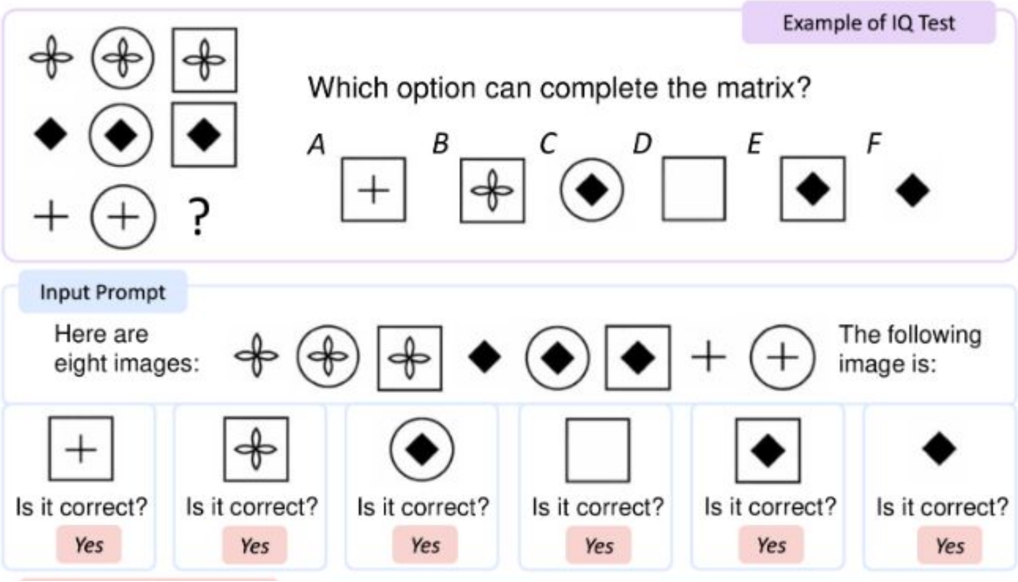
	$k=32$	$k=32$
Flamingo - 3B	57.1	**45.5**
Flamingo - 9B	**60.4**	44.0

IQ- Test



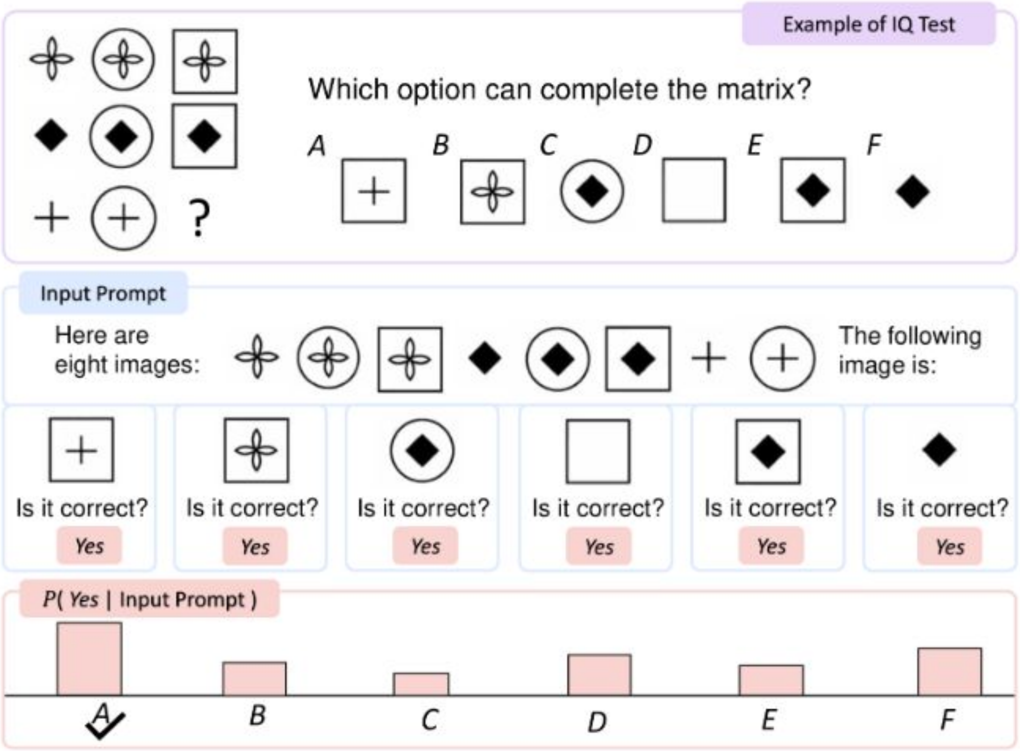
Huang, Shaohan, et al. "Language Is Not All You Need: Aligning Perception with Language Models." arXiv preprint arXiv:2302.14045 (2023).

IQ- Test



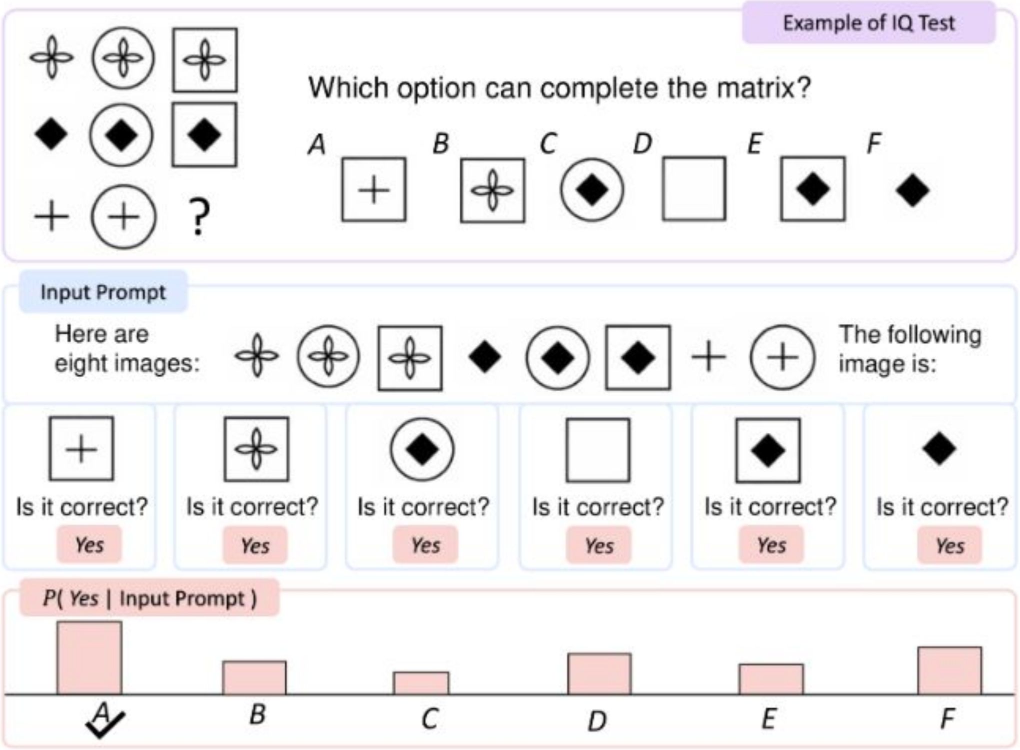
Huang, Shaohan, et al. "Language Is Not All You Need: Aligning Perception with Language Models." arXiv preprint arXiv:2302.14045 (2023).

IQ- Test



Huang, Shaohan, et al. "Language Is Not All You Need: Aligning Perception with Language Models." arXiv preprint arXiv:2302.14045 (2023).

IQ- Test



Method	Accuracy
Random Choice	17%
Kosmos-1	22%
w/o language-only instruction tuning	**26%**

Results

Huang, Shaohan, et al. "Language Is Not All You Need: Aligning Perception with Language Models." arXiv preprint arXiv:2302.14045 (2023).

OCR Free Language understanding

It's clear the filmmakers weren't sure where they wanted their story to go, and even more clear that they lack the skills to get us to this undetermined destination.

"Question: does this picture contain real hate speech? Answer: {answer}"

"Question: what is the sentiment of the opinion? Answer: {answer}"

Kiela, Douwe, et al. "The hateful memes challenge: Detecting hate speech in multimodal memes." Advances in Neural Information Processing Systems 33 (2020): 2611-2624.

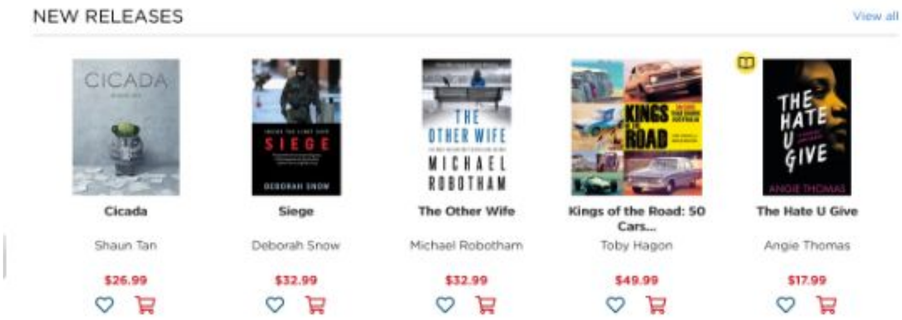
OCR Free Language understanding

Kosmos -1 models dont uses OCR text whereas flamingo models do use them

Model	HatefulMemes	Rendered SST-2
CLIP ViT-B/32	57.6	59.6
CLIP ViT-B/16	61.7	59.8
CLIP ViT-L/14	63.3	64.0
Flamingo-3B	53.7	-
Flamingo-9B	57.0	-
Kosmos-1 (1.6B)	**63.9**	**67.1**
Flamingo (80B)	46.4	-

Huang, Shaohan, et al. "Language Is Not All You Need: Aligning Perception with Language Models." arXiv preprint arXiv:2302.14045 (2023).

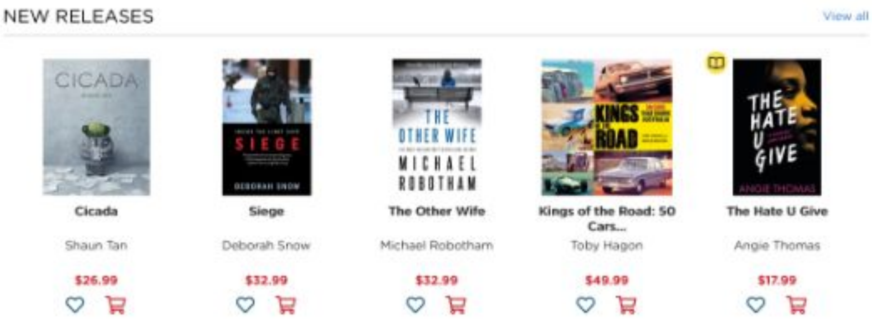
WEb page Question answering



“Given the context below from web page, extract the answer from the given text like this: Question: Who is the publisher of this book? Answer: Penguin Books Ltd. Context: {WebText} Q: {question} A: {answer} ”

Huang, Shaohan, et al. "Language Is Not All You Need: Aligning Perception with Language Models." arXiv preprint arXiv:2302.14045 (2023).

WEb page Question answering

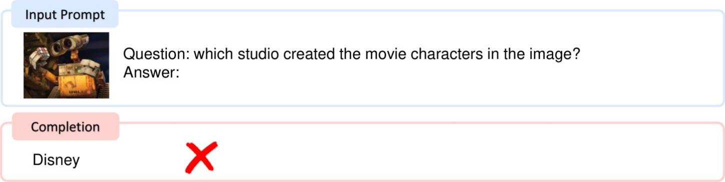


Models	EM	F1
Using extracted text		
LLM	7.6	17.9
Kosmos-1	**15.8**	**31.3**
Without using extracted text		
Kosmos-1	3.8	10.6

“Given the context below from web page, extract the answer from the given text like this:
Question: Who is the publisher of this book?
Answer: Penguin Books Ltd. Context: {WebText}
Q: {question} A: {answer} ”

Huang, Shaohan, et al. "Language Is Not All You Need: Aligning Perception with Language Models." arXiv preprint arXiv:2302.14045 (2023).

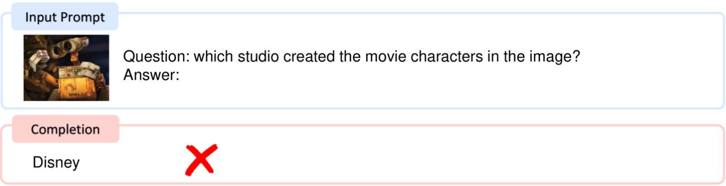
Multimodal chain of thought prompting



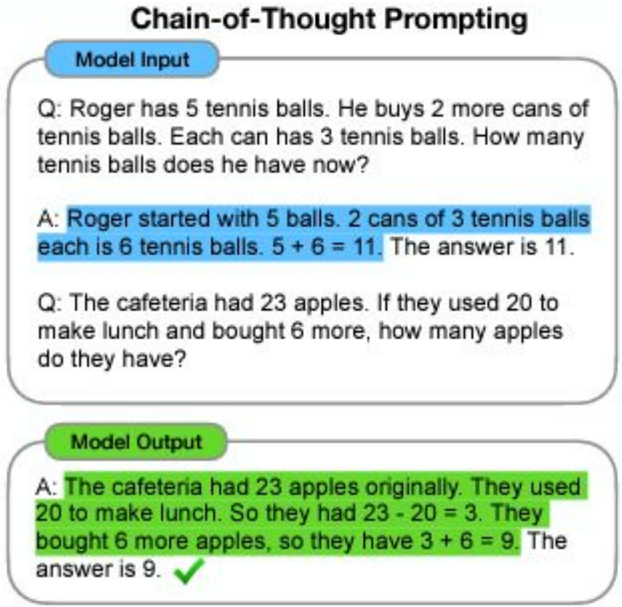
Standard Prompting

Huang, Shaohan, et al. "Language Is Not All You Need: Aligning Perception with Language Models." arXiv preprint arXiv:2302.14045 (2023).

Multimodal chain of thought prompting



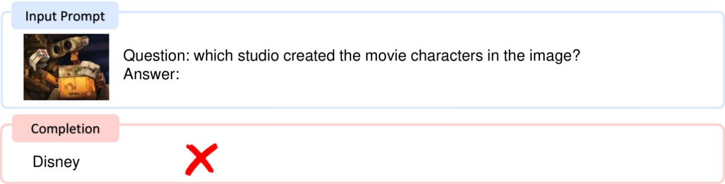
Standard Prompting



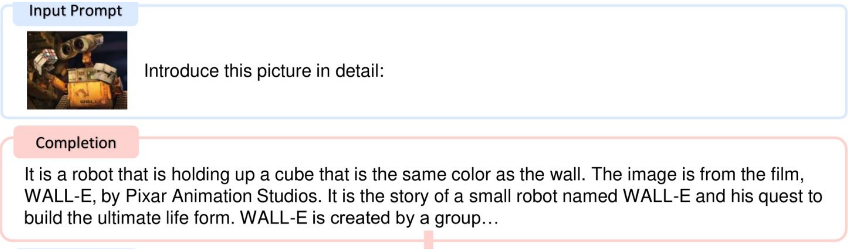
Chain of thought prompting for language

Wei, Jason, et al. "Chain of thought prompting elicits reasoning in large language models." arXiv preprint arXiv:2201.11903 (2022).
Huang, Shaohan, et al. "Language Is Not All You Need: Aligning Perception with Language Models." arXiv preprint arXiv:2302.14045 (2023).

Multimodal chain of thought prompting

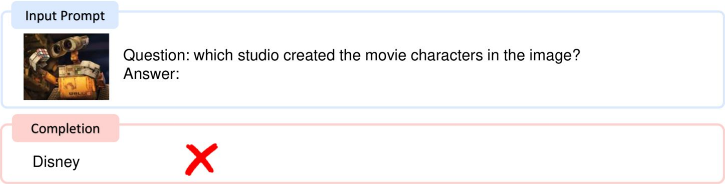


Standard Prompting

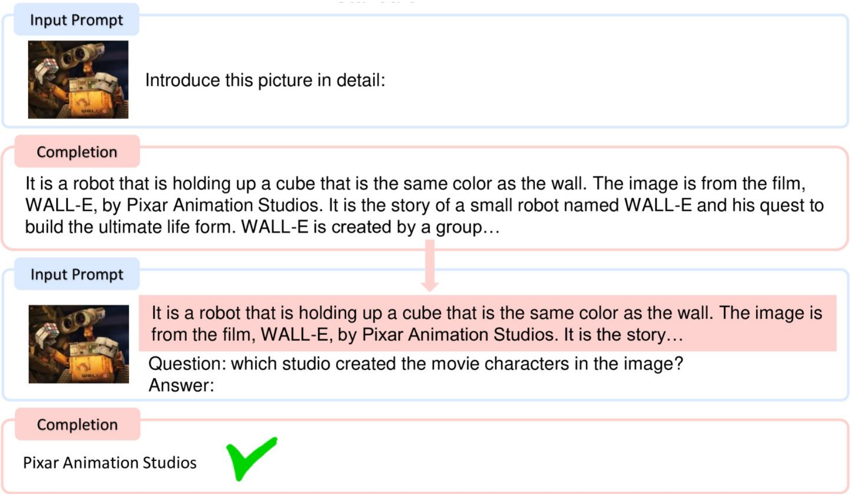


Huang, Shaohan, et al. "Language Is Not All You Need: Aligning Perception with Language Models." arXiv preprint arXiv:2302.14045 (2023).

Multimodal chain of thought prompting



Standard Prompting

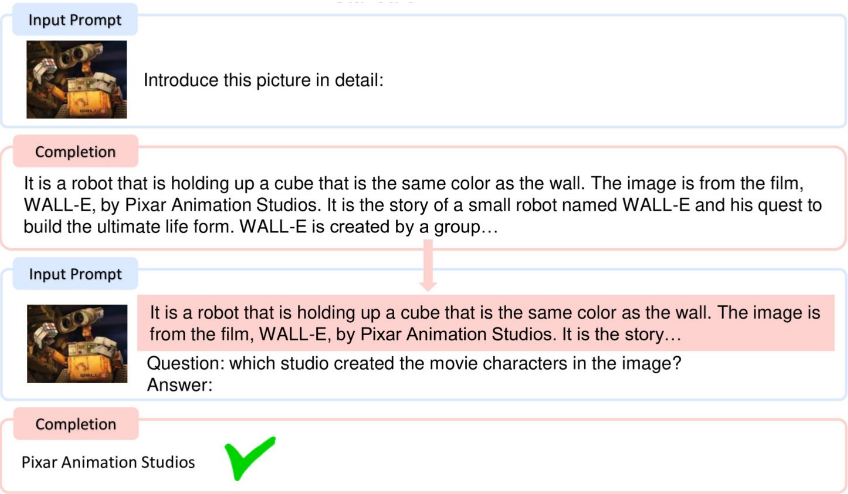


Multimodal chain of thought

Huang, Shaohan, et al. "Language Is Not All You Need: Aligning Perception with Language Models." arXiv preprint arXiv:2302.14045 (2023).

Multimodal chain of thought prompting

Models	Accuracy
CLIP ViT-B/32	59.6
CLIP ViT-B/16	59.8
CLIP ViT-L/14	64.0
Kosmos-1	67.1
w/ multimodal CoT prompting	**72.9**



Multimodal chain of thought

Huang, Shaohan, et al. "Language Is Not All You Need: Aligning Perception with Language Models." arXiv preprint arXiv:2302.14045 (2023).

Zero shot image classification

Input is “The photo of the” and
output is constrained to 1K classes

Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009.

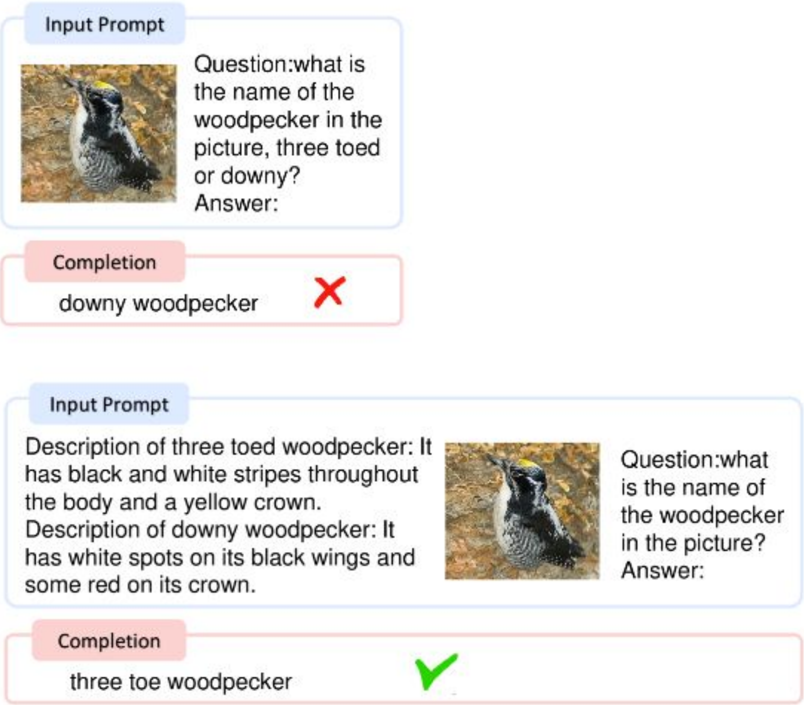
Zero shot image classification

Input is “The photo of the” and
output is constrained to 1K classes

Model	Without Constraints	With Constraints
GIT [58]	1.9	33.5
Kosmos-1	**4.0**	**38.1**

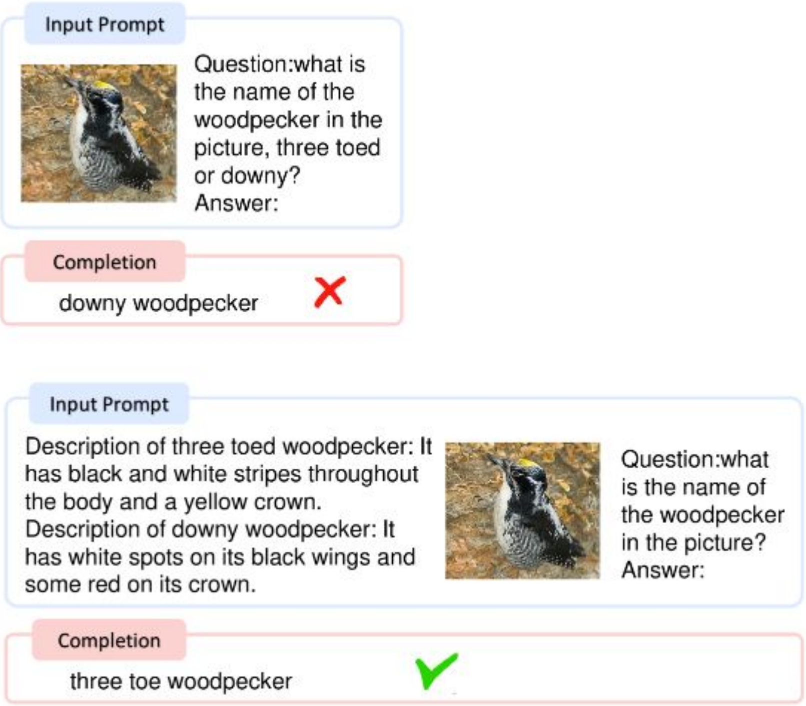
Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009.

Zero shot Image classification with Descriptions



Huang, Shaohan, et al. "Language Is Not All You Need: Aligning Perception with Language Models." arXiv preprint arXiv:2302.14045 (2023).

Zero shot Image classification with Descriptions



Settings	Accuracy
Without Descriptions	61.7
With Descriptions	**90.0**

Huang, Shaohan, et al. "Language Is Not All You Need: Aligning Perception with Language Models." arXiv preprint arXiv:2302.14045 (2023).

Language Tasks

MLLM trained on the text corpora as an LLM is the baseline

Task	Zero-shot		One-shot		Few-shot ($k = 4$)	
	LLM	KOSMOS-1	LLM	KOSMOS-1	LLM	KOSMOS-1
StoryCloze	**72.9**	72.1	**72.9**	72.2	**73.1**	72.3
HellaSwag	**50.4**	50.0	**50.2**	50.0	**50.4**	50.3
Winograd	**71.6**	69.8	**71.2**	68.4	**70.9**	69.8
Winogrande	**56.7**	54.8	**56.7**	54.5	**57.0**	55.7
PIQA	**73.2**	72.9	**73.0**	72.5	**72.6**	72.3
BoolQ	**56.4**	**56.4**	55.1	**57.2**	58.7	**59.2**
CB	39.3	**44.6**	41.1	**48.2**	42.9	**53.6**
COPA	**68.0**	63.0	**69.0**	64.0	**69.0**	64.0
Average	61.1	60.5	61.2	60.9	61.8	62.2

Cross Modal Transfer

Comparing Kosmos-1 trained with and without language-only instruction tuning on V+L tasks.

Model	COCO	Flickr30k	VQAv2	VizWiz
Kosmos-1	84.7	**67.1**	**51.0**	**29.2**
w/o language-only instruction tuning	**87.6**	65.2	46.7	27.9

Cross Modal Transfer

Comparing Kosmos-1 trained with and without language-only instruction tuning on V+L tasks.

Model	COCO	Flickr30k	VQAv2	VizWiz
Kosmos-1	84.7	**67.1**	**51.0**	**29.2**
w/o language-only instruction tuning	**87.6**	65.2	46.7	27.9

Comparing Kosmos-1 and LLM baseline on common sense reasoning tasks.

Model	Size Reasoning	Color Reasoning	
	RelativeSize	MemoryColor	ColorTerms
Using retrieved images			
VaLM [53]	85.0	58.6	52.7
Language-only zero-shot evaluation			
LLM	92.7	61.4	63.4
Kosmos-1	**94.2**	**76.1**	**73.1**

Strengths

- Kosmos - 1 uses fewer parameters than Flamingo models but is competitive on results and often outperforms.
- Unlike Flamingo their zero shot evaluation methods dont use two mock demonstrations.
- They train their models on open source datasets.

Weaknesses

- Kosmos - 1 seems to be scaling poorly to higher K, also it should be noted that they haven't done experiments for K = 32 a setting at which Flamingo does best and can outperform kosmos-1.
- Kosmos - 1 seems to be doing poorly on classification tests, when looking at imagenet results and it performs well only with instructions, which may not be available at all times.
- Even though kosmos - 1 can ingest interleaved multimodal input, they have not performed any experiments around video reasoning tasks.

Future Work

- Authors can try to scale the models so that their sizes can be comparable to Flamingo for an apples to apples comparison.
- Authors can try to incorporate more modalities into their model, such as video etc.
- Authors can look into distillation methods to create smaller models when they scale up.

Discussion

- Do you think not releasing models and datasets to the public hurts future research ?
- IQ tests seem like a effective method to learn how close a model is to human intelligence. Can we train models that do well on theses tests, have logic and pattern recognition skills?

References

1. Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901.
2. Hao, Yaru, et al. "Language models are general-purpose interfaces." arXiv preprint arXiv:2206.06336 (2022).
3. Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." arXiv preprint arXiv:2204.14198 (2022).
4. Huang, Shaohan, et al. "Language Is Not All You Need: Aligning Perception with Language Models." arXiv preprint arXiv:2302.14045 (2023).
5. Hoffmann, Jordan, et al. "Training compute-optimal large language models." arXiv preprint arXiv:2203.15556 (2022).
6. Tsimpoukelli, Maria, et al. "Multimodal few-shot learning with frozen language models." Advances in Neural Information Processing Systems 34 (2021): 200-212.
7. Aghajanyan, Armen, et al. "Cm3: A causal masked multimodal model of the internet." arXiv preprint arXiv:2201.07520 (2022).
8. Brock, Andy, et al. "High-performance large-scale image recognition without normalization." International Conference on Machine Learning. PMLR, 2021.
9. Wang, Wenhui, et al. "Image as a foreign language: Beit pretraining for all vision and vision-language tasks." arXiv preprint arXiv:2208.10442 (2022).
10. Wang, Hongyu, et al. "Foundation transformers." arXiv preprint arXiv:2210.06423 (2022).
11. Sun, Yutao, et al. "A Length-Extrapolatable Transformer." arXiv preprint arXiv:2212.10554 (2022).