ETH zürich

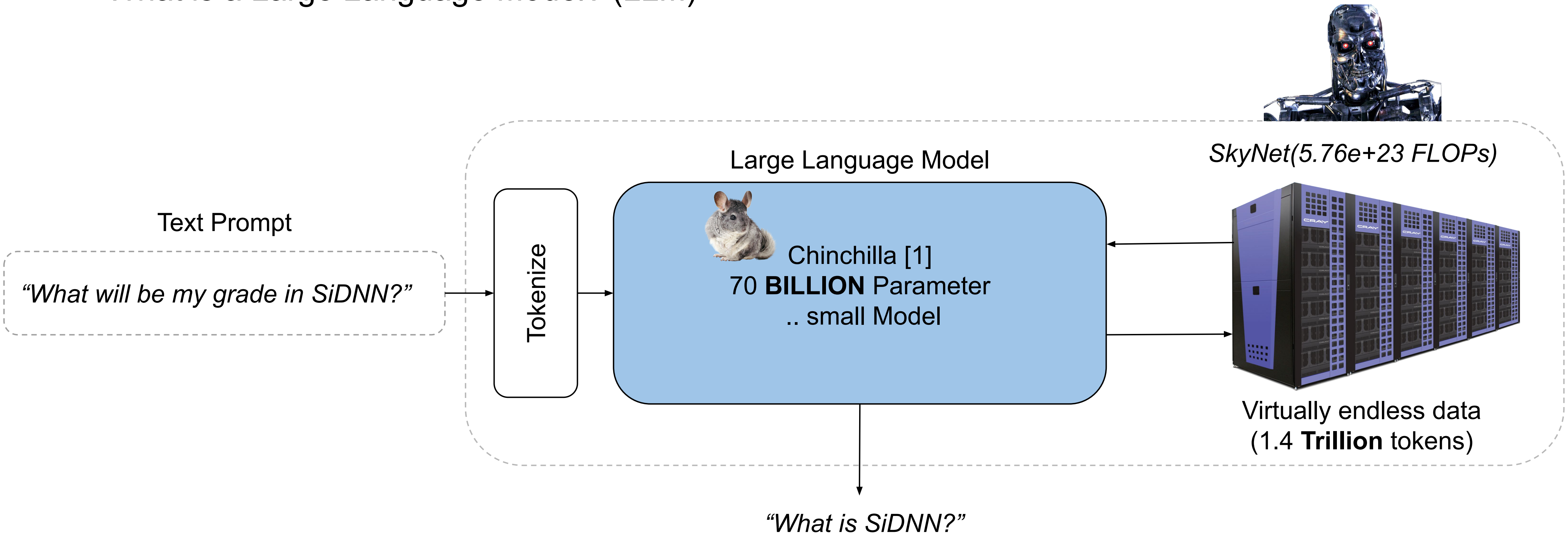
Flamingo: A Visual Language Model for Few-Shot Learning

Presenter : Turcan Tuna
Mentor : Ferjad Naeem

18.04.2023

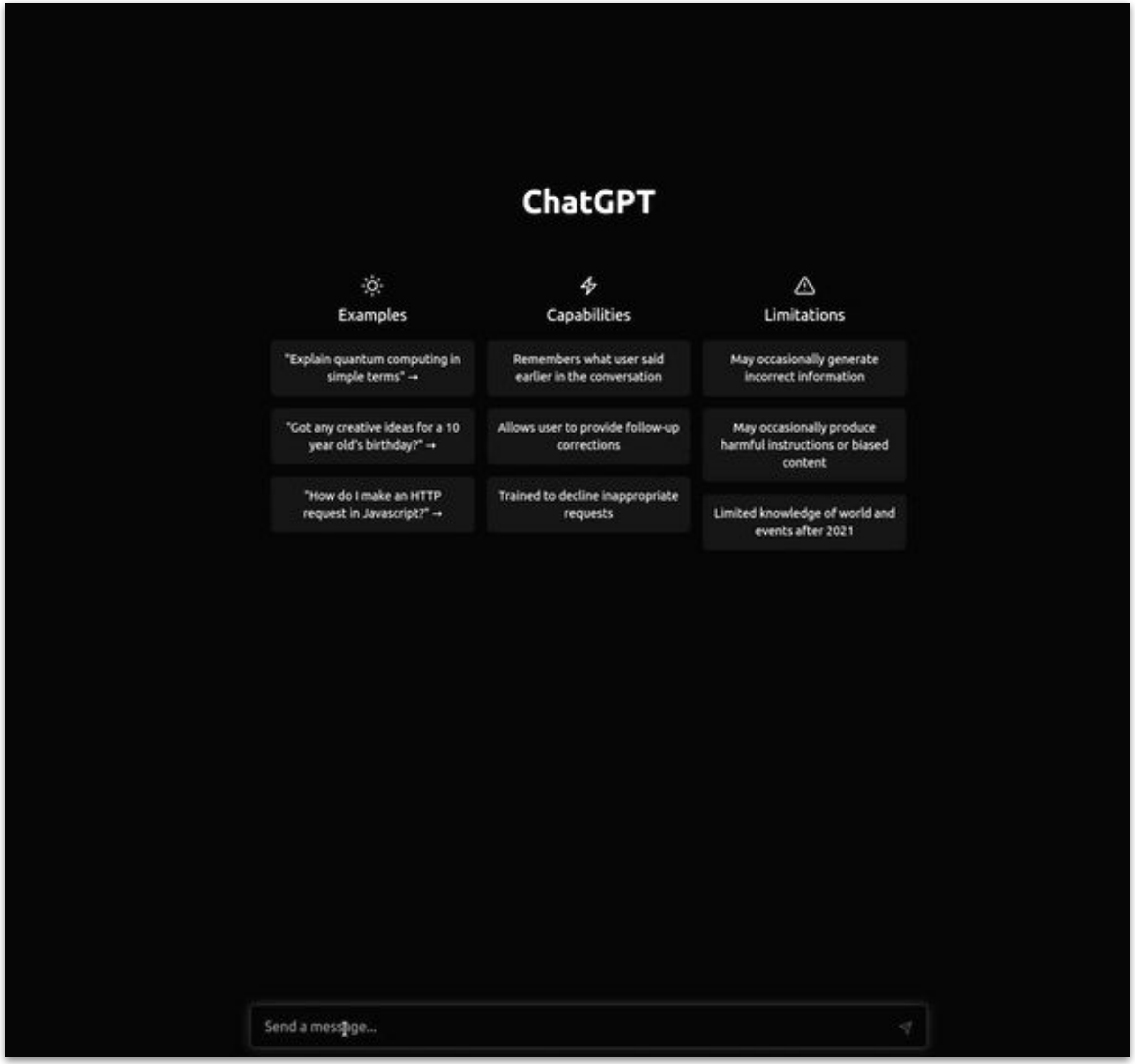
Premise - Large Language Models

- What is a Large Language Model? (LLM)



[1] Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., ... & Sifre, L. (2022). Training compute-optimal large language models. arXiv preprint arXiv:2203.15556.

Premise - Large Language Models



Premise - Flamingo

- What is Flamingo[2]? 🦩

Visual Features

A language model

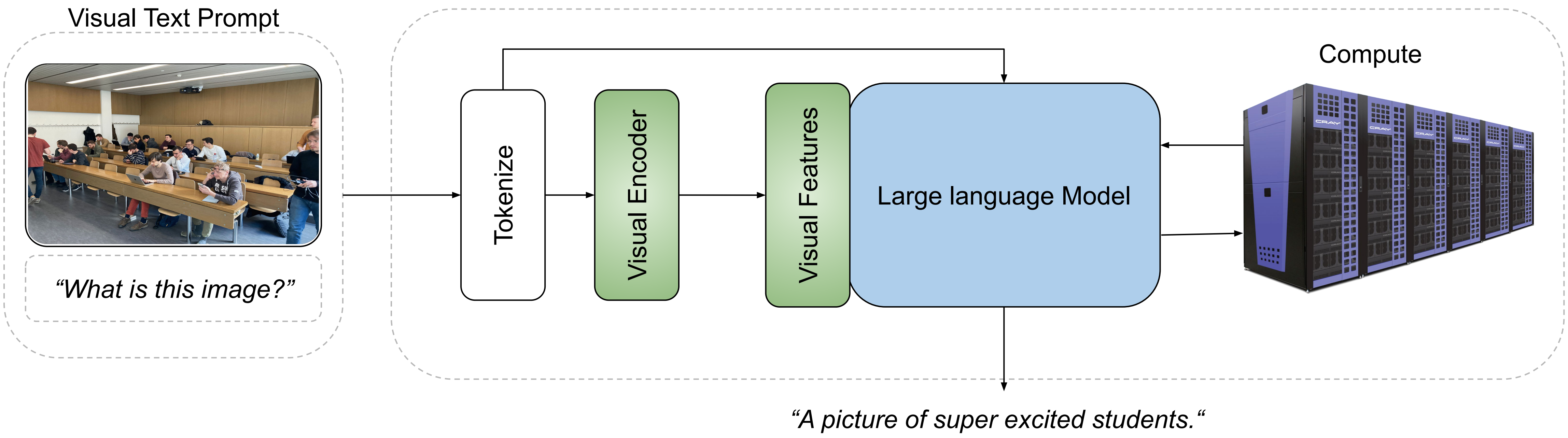
"...a visually-conditioned autoregressive text generation model able to ingest a sequence of text tokens interleaved with images and/or videos, and produce text as output".

Multi-Modal prompt

[2] Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... & Simonyan, K. (2022). Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems, 35, 23716-23736.

Premise - Visually Conditioned Large Language Models

- How LLMs can gain the **ability to see**?



Premise - Example

Visual Text Prompt

“Describe me this image.”

*Visually Conditioned LLM
(Flamingo)

“Students in the lecture hall”

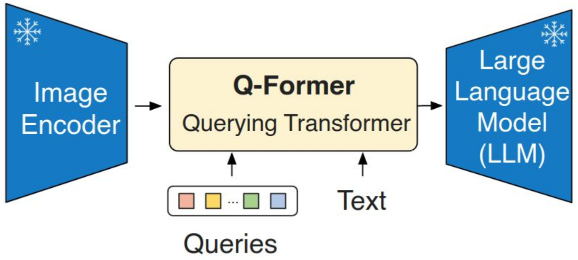
* Deployed the re-implementation from https://github.com/dhansmair/flamingo-mini.

Premise - Example

Visual Text Prompt

"Describe me this image."

Blip-2[3], 2023, SOTA.



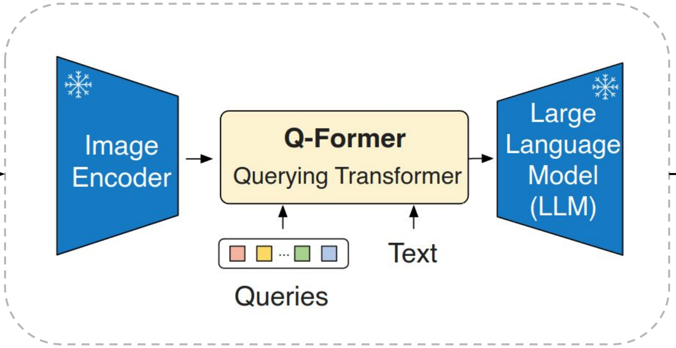
"A group of people sitting at desks in a classroom"

[3] Li, J., Li, D., Savarese, S., & Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597.

Premise - Advanced Example

Lets have some fun! , Visual - Question Answering (VQA).

Blip-2



“Describe me this image.”

“People in a room”

“How many people?”

“Too many”

Motivation

- LLMs are **very expensive** to train.
- No intuitive expansion support.
    - How to not forget?

- It **is** expensive to train. Really..

"....training BERT on GPU is roughly equivalent to a trans-American flight."[4]

Model	Hardware	Power (W)	Hours	kWh·PUE	Cloud compute cost
$\text{Transformer}_{base}$	P100x8	1415.78	12	27	\$41–\$140
Transformer_{big}	P100x8	1515.43	84	201	\$289–\$981
ELMo	P100x3	517.66	336	275	\$433–\$1472
BERT_{base}	V100x64	12,041.51	79	1507	\$3751–\$12,571
BERT_{base}	TPUv2x16	—	96	—	\$2074–\$6912
NAS	P100x8	1515.43	274,120	656,347	\$942,973–\$3,201,722
NAS	TPUv2x1	—	32,623	—	\$44,055–\$146,848
GPT-2	TPUv3x32	—	168	—	\$12,902–\$43,008

[4] Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. arXiv preprint arXiv:1906.02243.

Motivation

- LLMs require huge amount of data.
  - This data can be found as text (*MassiveText* [5])! But not for images.
  - 1.4 Trillion vs ~ 1.8 billion.

- ALIGN[6] dataset **1.8 billion images** paired with text but **noisy**!

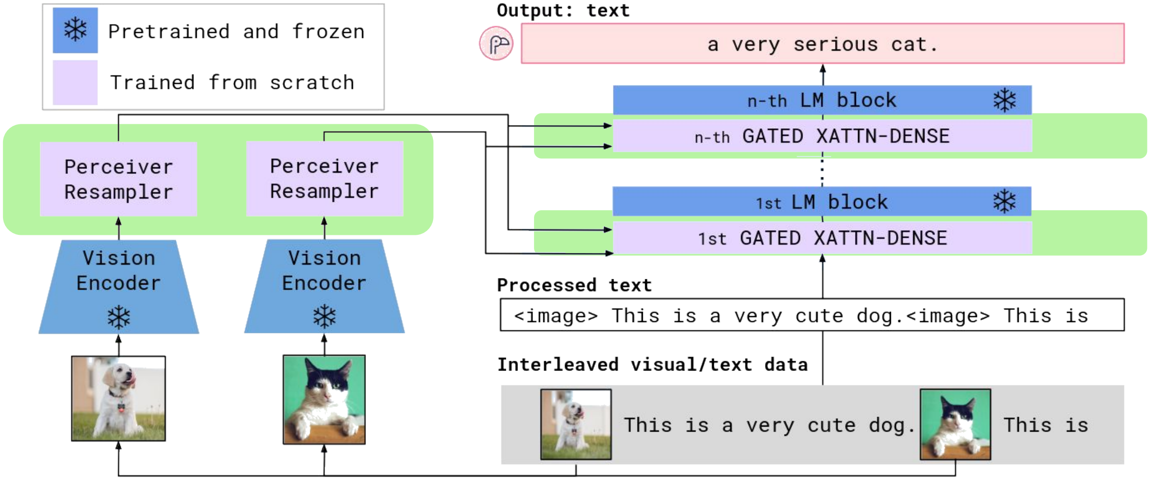
- A gap in the literature exists for **videos**!

[5] Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., ... & Irving, G. (2021). Scaling language models: Methods, analysis & insights from training gopher. arXiv preprint arXiv:2112.11446.
[6] Jia, C., Yang, Y., Xia, Y., Chen, Y. T., Parekh, Z., Pham, H., ... & Duerig, T. (2021, July). Scaling up visual and vision-language representation learning with noisy text supervision. In International Conference on Machine Learning (pp. 4904-4916). PMLR.

Contributions

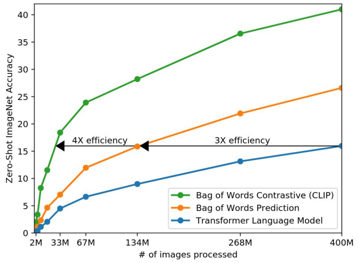
- A way to combined interleaved images and text
  - Gated cross-attention module
- A unique perceiver architecture with fixed output.
  - Perceiver sampler
- Evaluation and ablation

The **Flamingo** Model



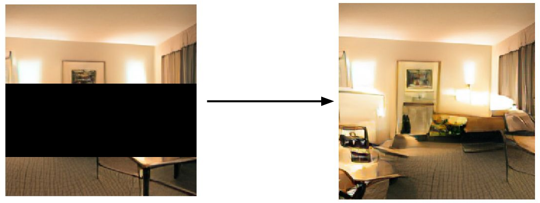
Related Work

CLIP[7]
- Close-ended task superiority.
- Trained from scratch.
- Very good zero-shot, few-shot performance.

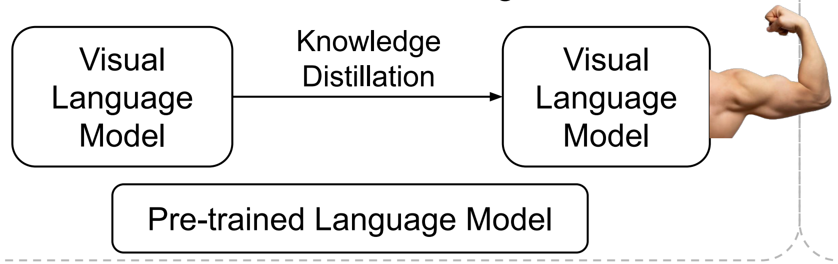


CM3[8]
- Big model 24 days on 384 A100s
- Multimodal & Unimodal tasks.
- Directly works on HTML.
- Can output images.

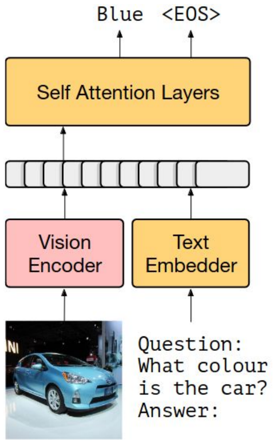
“Complete”



VLKD[9]
- Visual-Language Knowledge Dist.
- Efficient and compact.
- Similar results to Flamingo



Frozen[10] (DeepMind, 2021)
- Image - conditioned prompt learning.
- A good few-shot learner.
- Goal: generalizability



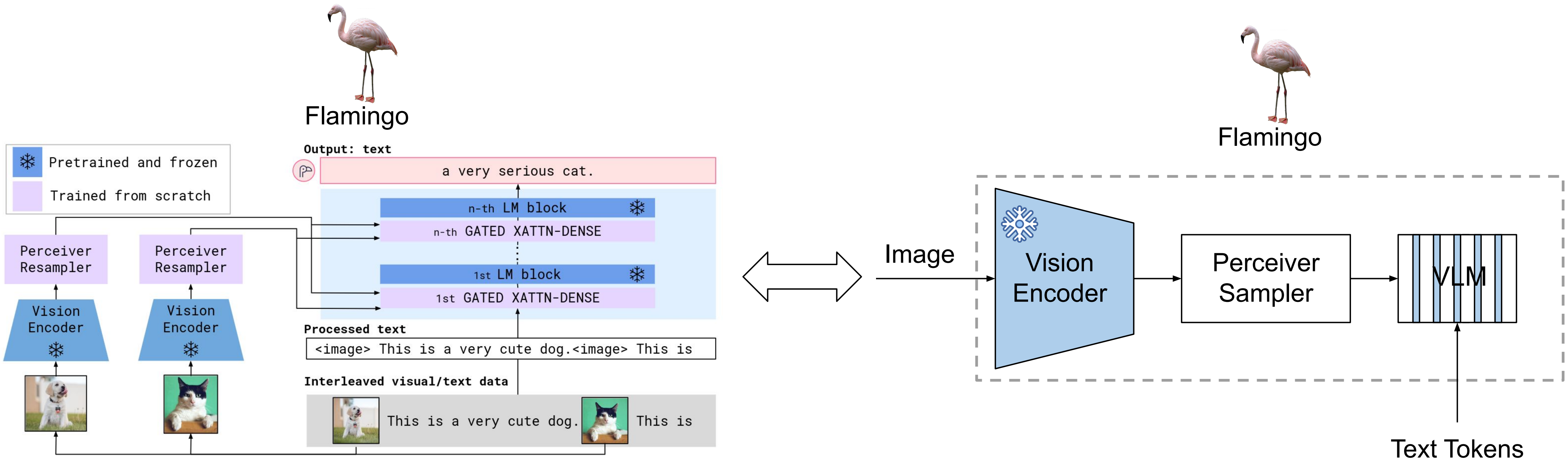
[7] (CLIP) A. Radford et al., "Learning transferable visual models from natural language supervision", ICML (2021)
[8] (CM3) A. Aghajanyan et al., "CM3: A Causal Masked Multimodal Model of the Internet", arxiv (2022)
[9] (VLKD) Dai, W., Hou, L., Shang, L., Jiang, X., Liu, Q., & Fung, P. (2022). Enabling multimodal generation on CLIP via vision-language knowledge distillation. arXiv preprint arXiv:2203.06386.
[10] (Frozen) M. Tsimpoukelli et al., "Multimodal few-shot learning with frozen language models", NeurIPS (2021)

Method - Introduction

- Goal: Provide LLMs ability to see.
  - Convert LLM to VLM.

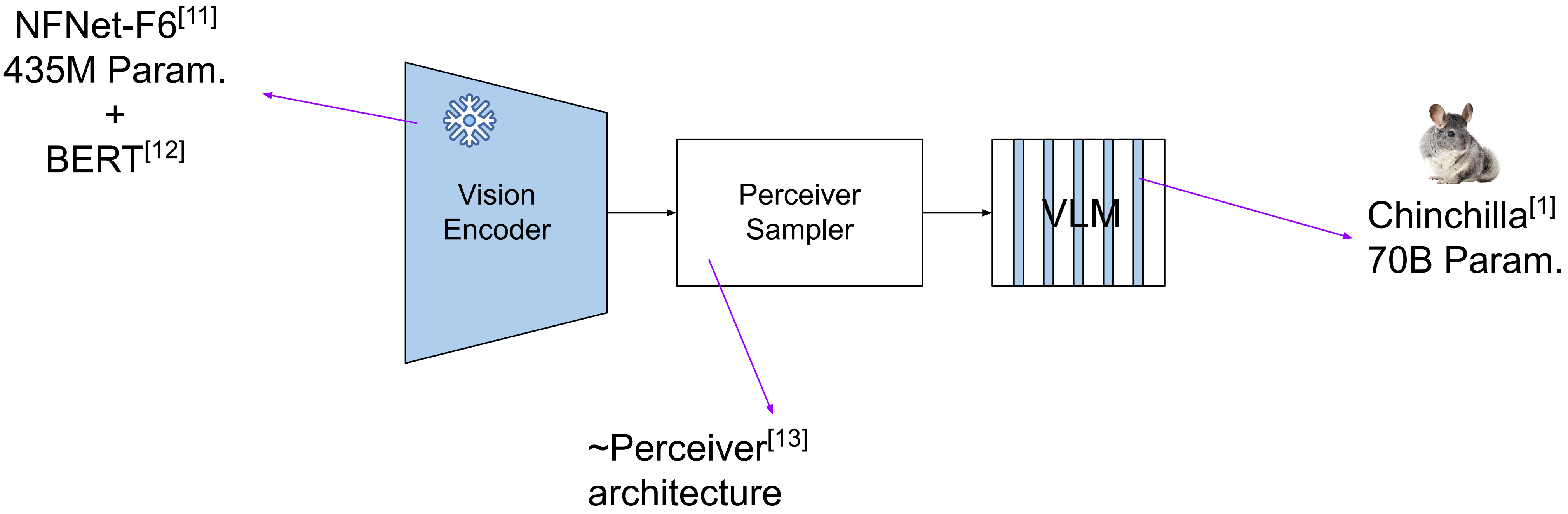
- Key ideas:
  - Extend a **frozen** Pre-trained Language Model.
  - Reducing visual input to a **fixed** number of **tokens** with **Perceiver Sampler**.
  - Cross attention layers to visually **condition LLM**.
  - Training on a different types of data.

Method - Introduction



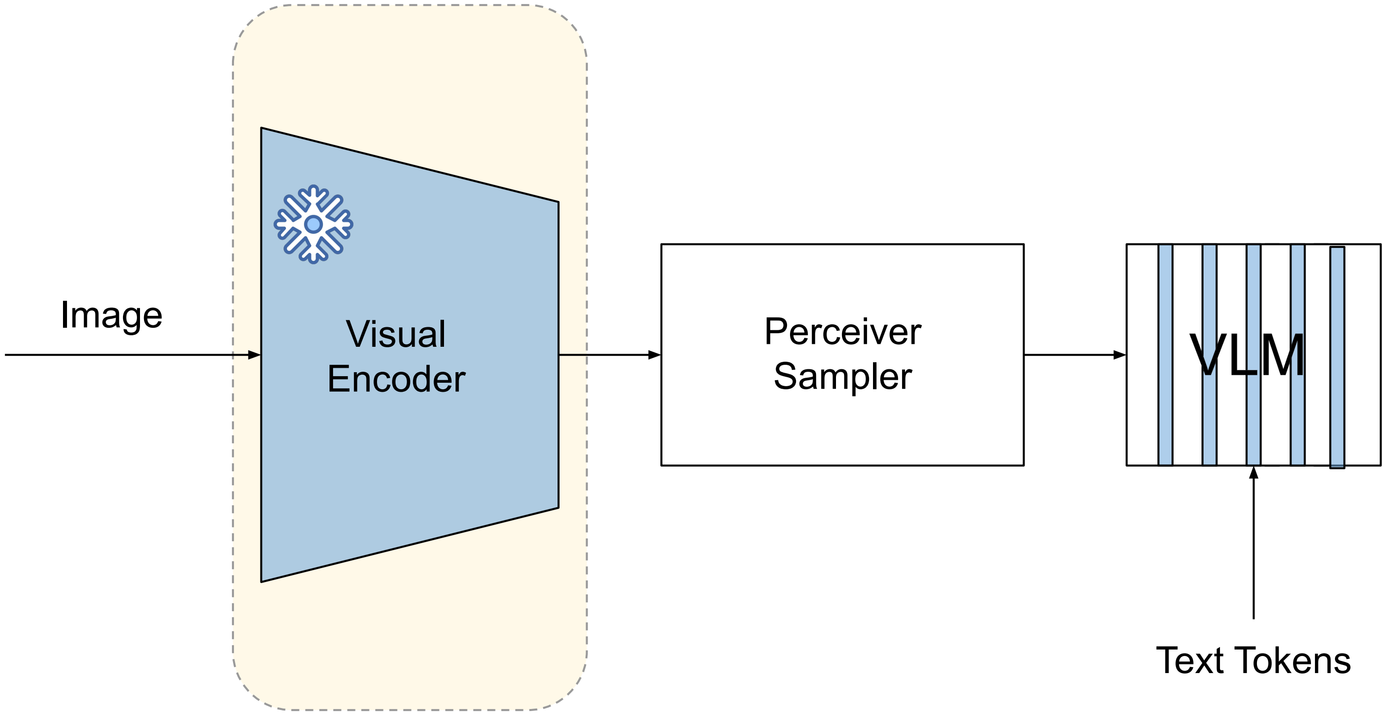
Method - Introduction

- Multiple Models are Recycled.

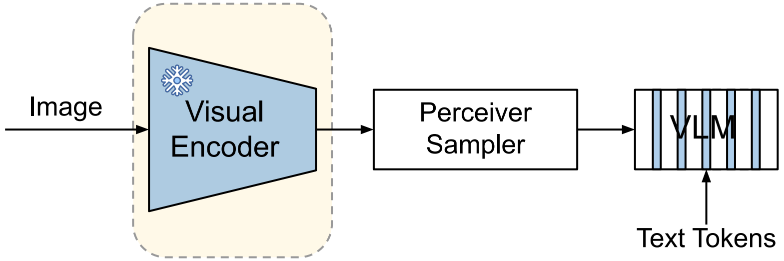


[11] Brock, A., De, S., Smith, S. L., & Simonyan, K. (2021, July). High-performance large-scale image recognition without normalization. In International Conference on Machine Learning (pp. 1059-1071). PMLR.
[12] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
[13] Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., & Carreira, J. (2021, July). Perceiver: General perception with iterative attention. In International conference on machine learning (pp. 4651-4664). PMLR.

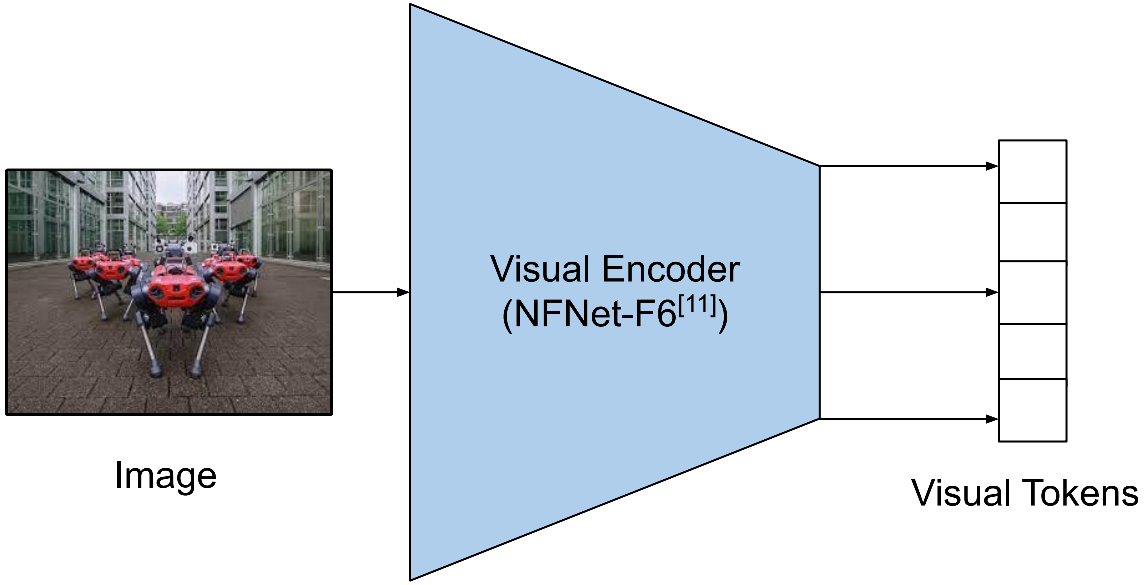
Method

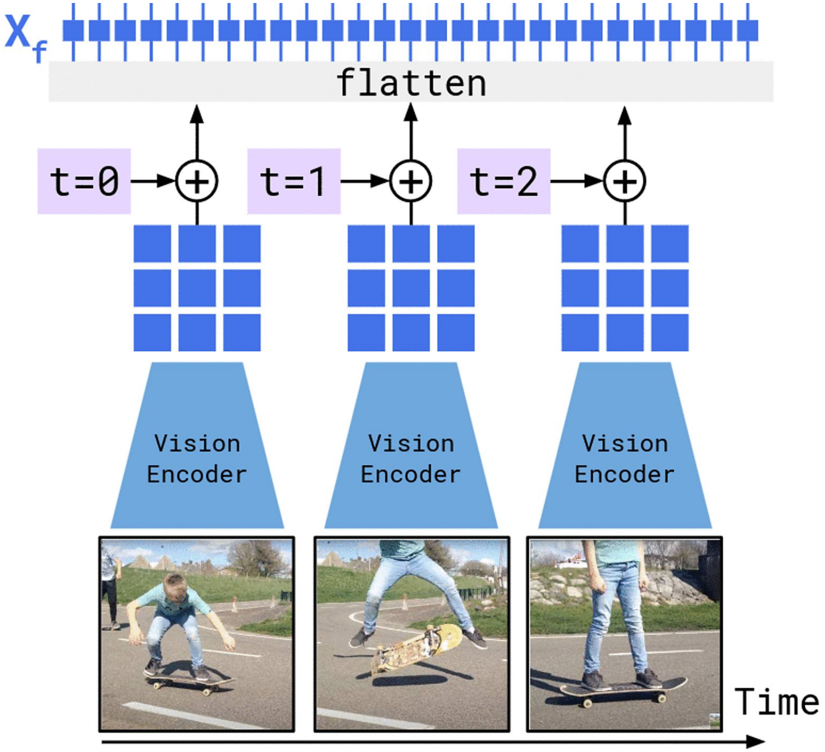


Method - Visual Encoder



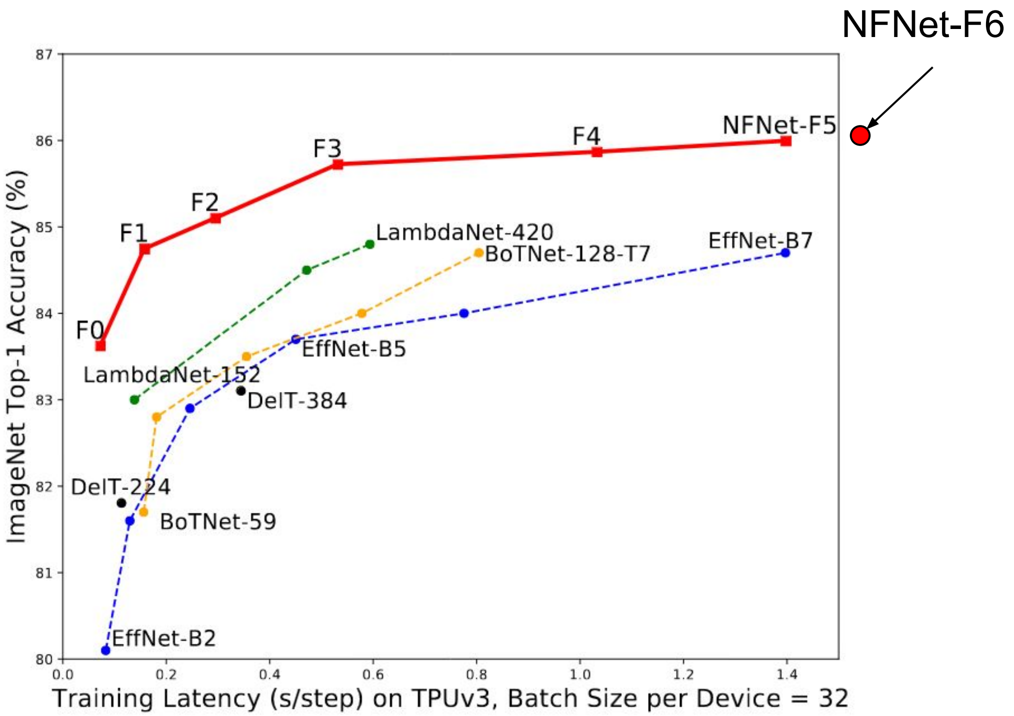
- VLM needs ***text-conditioned*** Visual tokens.



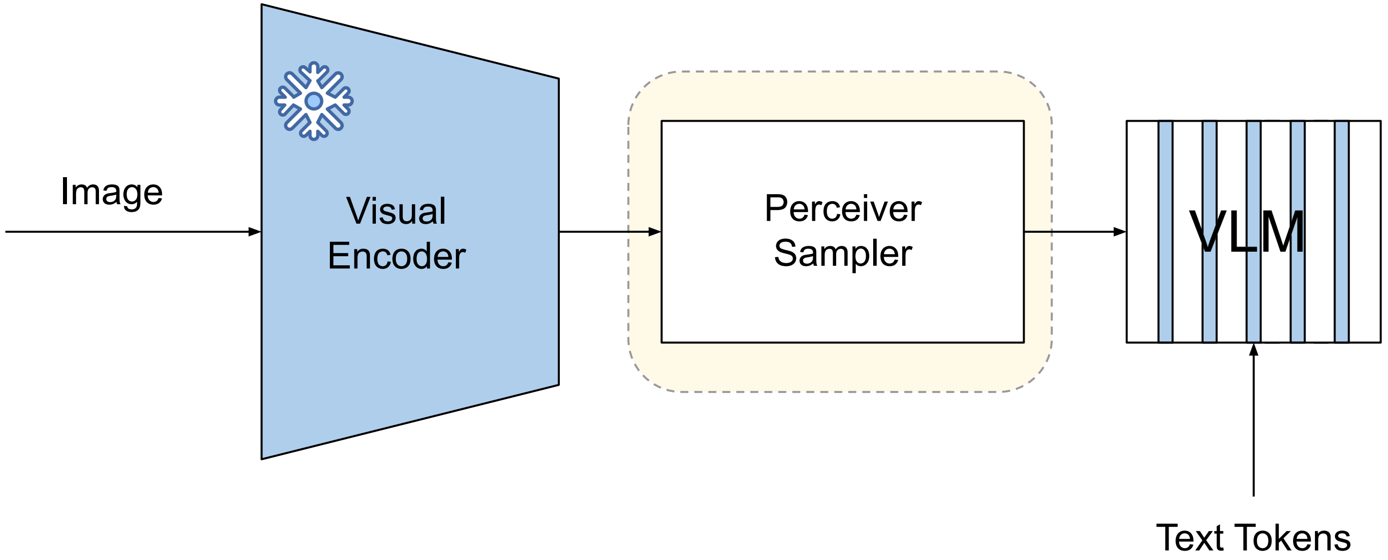


Method - Visual Encoder

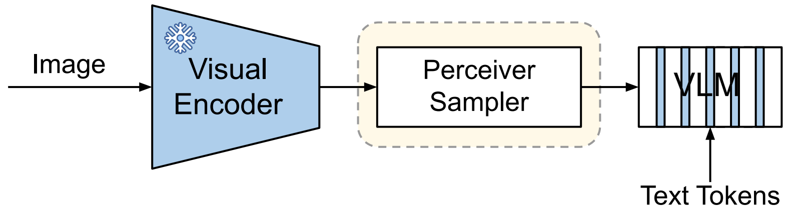
- The NFNet-F6 architecture is from Normalizer-Free ResNet.



Method



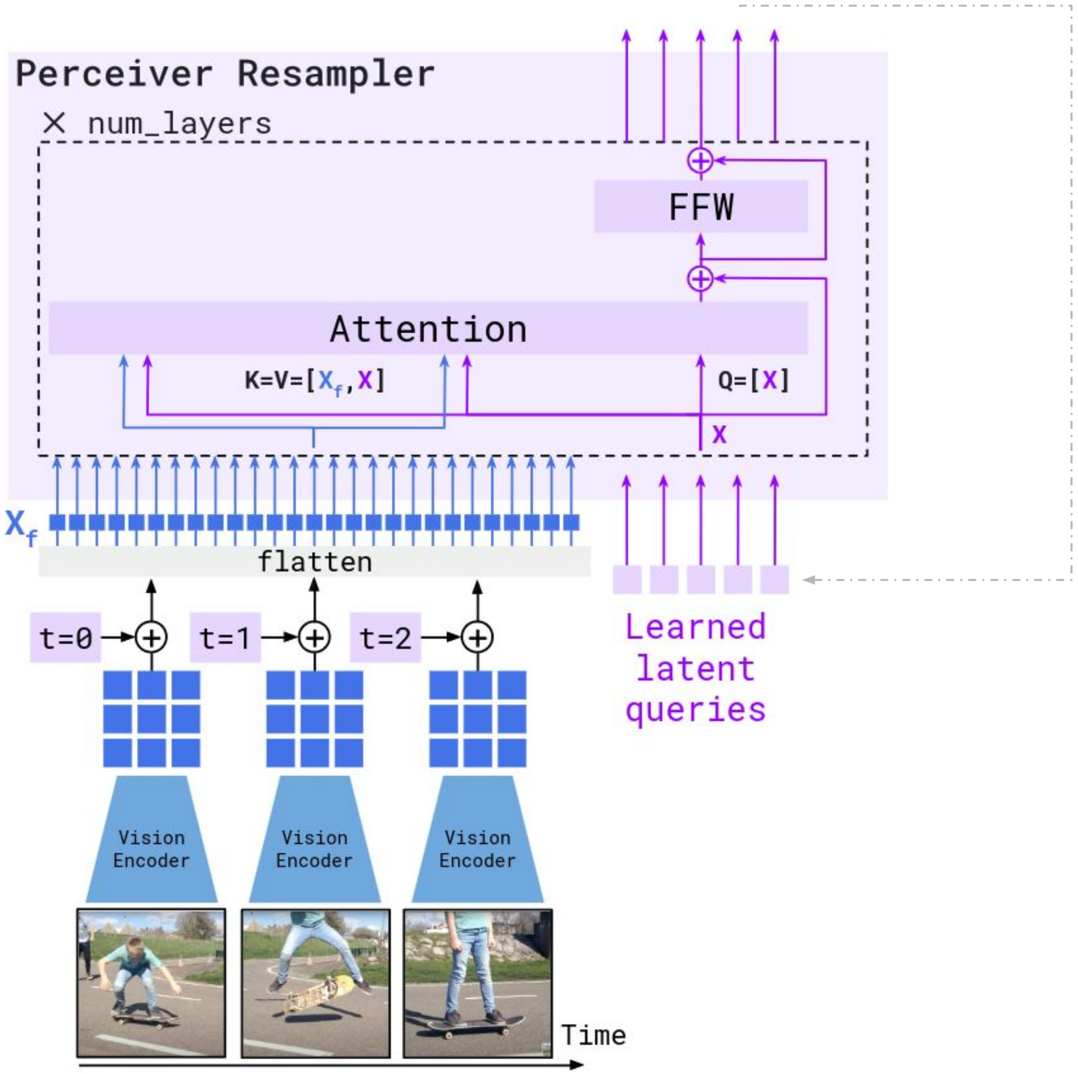
Method - Perceived Sampler



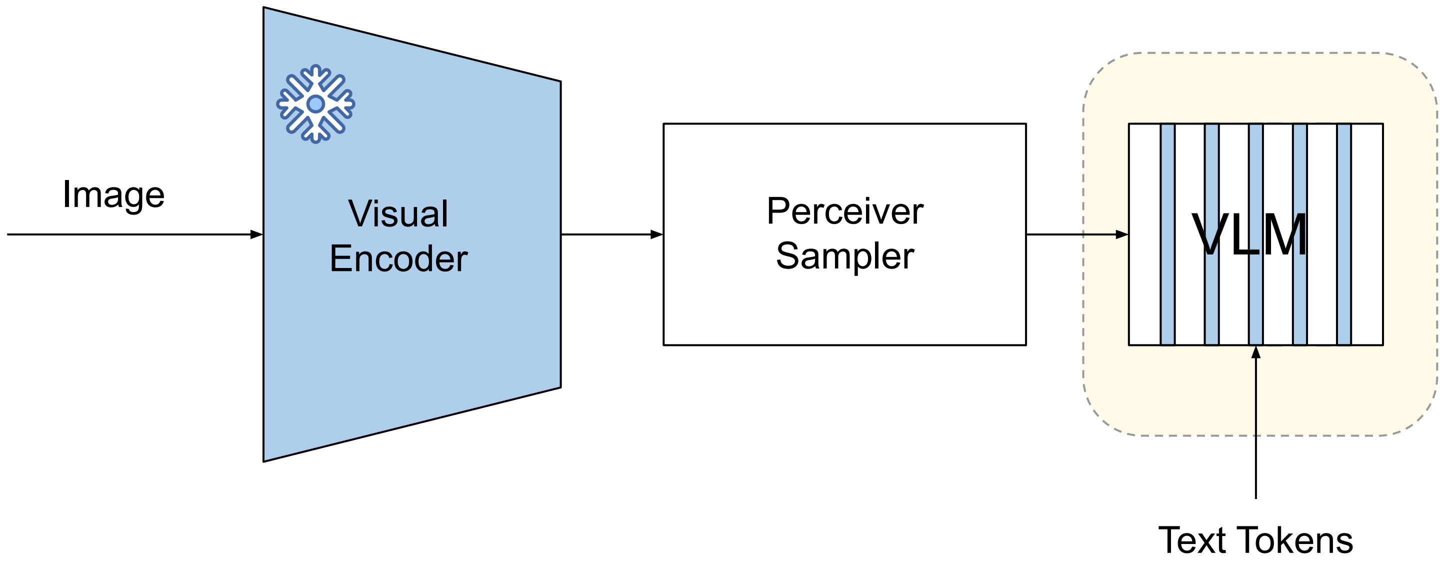
- Latent queries:
  - Most important part of the data.
  - Model **learns what to extract**.

- No explicit spatial grid position encodings.

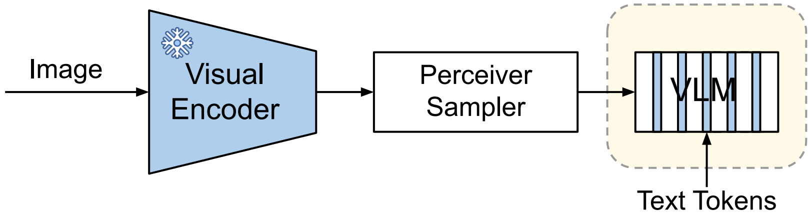
- Nb of **Output tokens** = Nb of **latent queries.**



Method

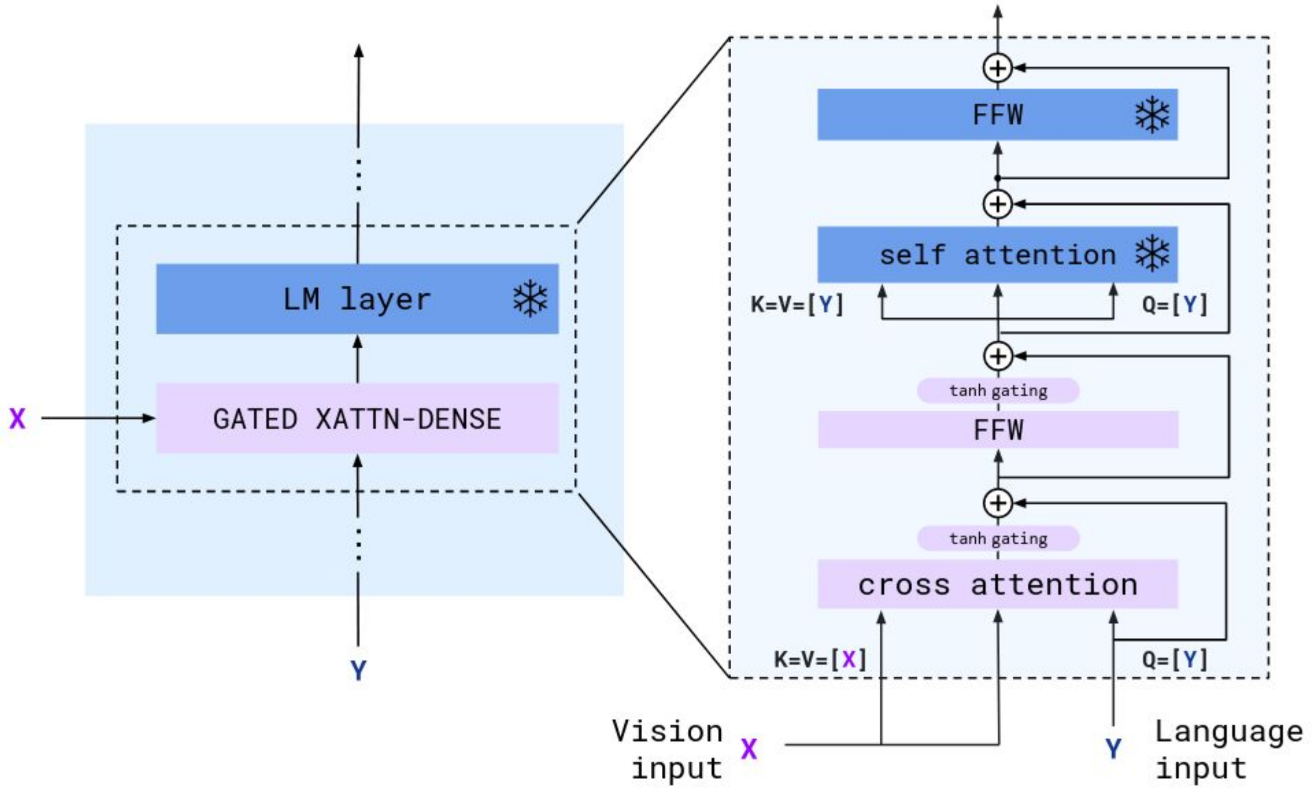


Method - Gated Cross attention Layers

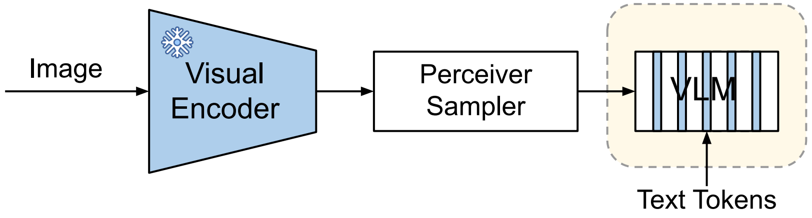


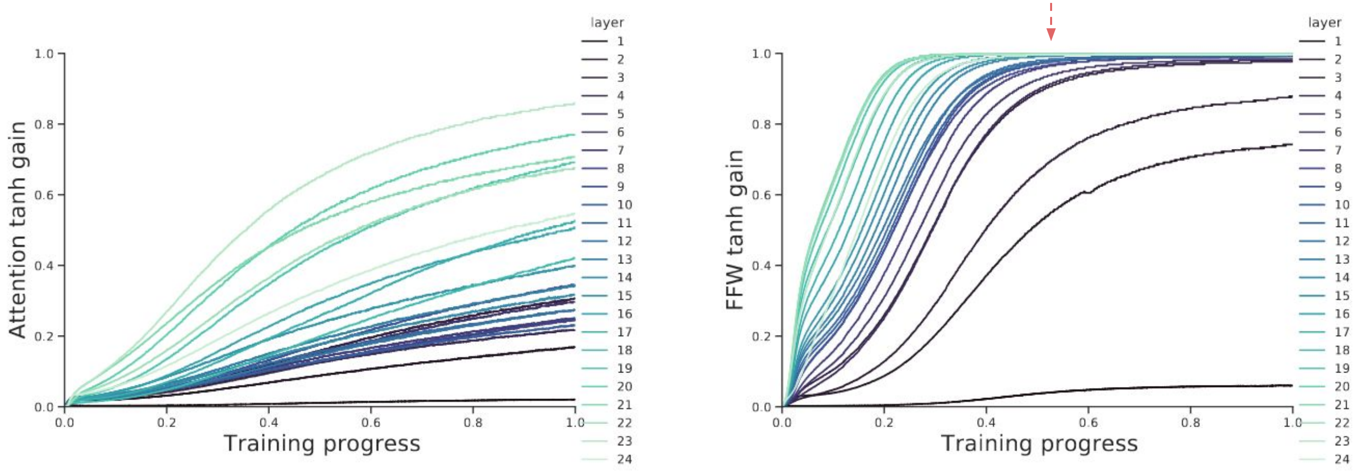
- **Gated cross-attention blocks**
  - Text conditioning on visual representations.

- Integrate new skills to LLMs without forgetting.
  - **Tanh gating**.

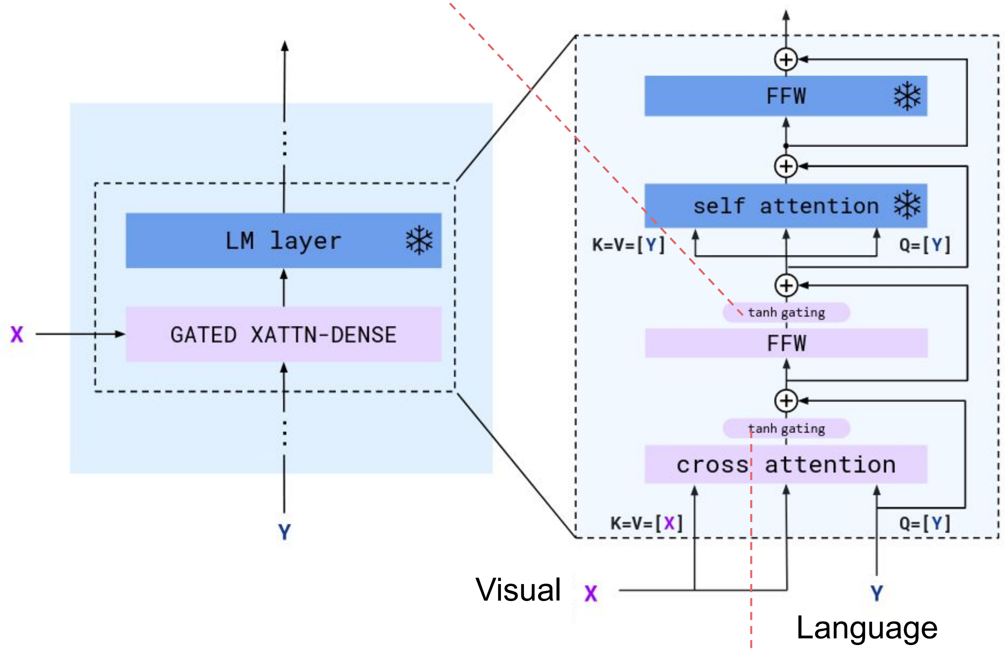


Method - Gated Cross attention Layers

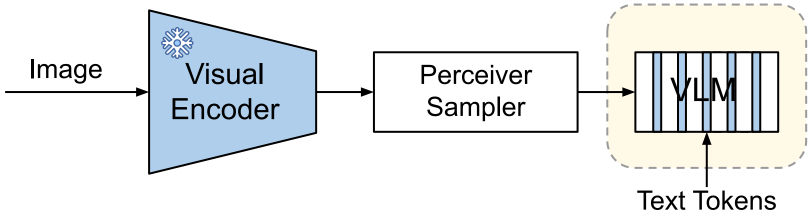




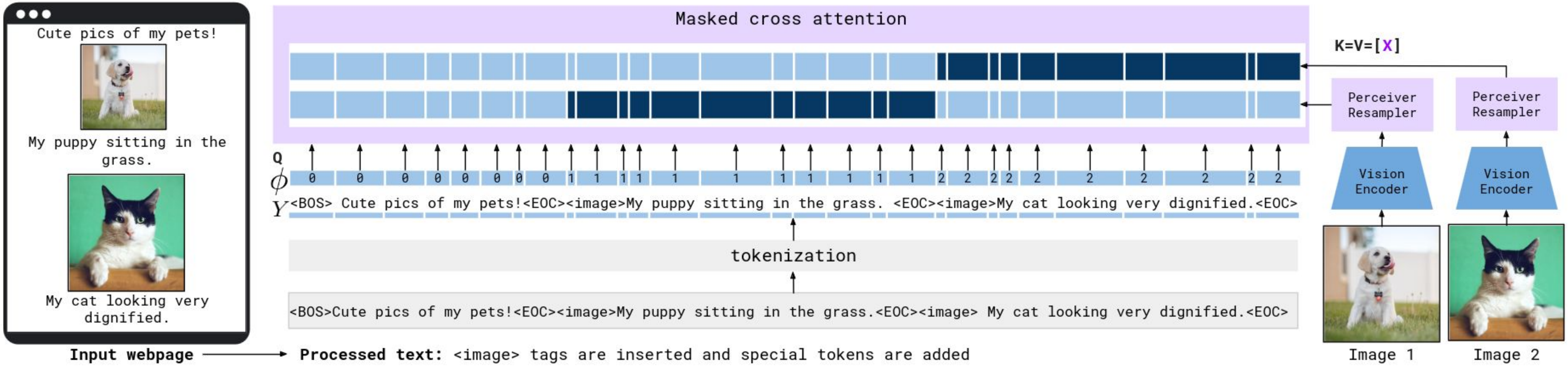
(a) Attention tanh gating (b) FFW tanh gating.



Method - Interleaved Visual / Text data support

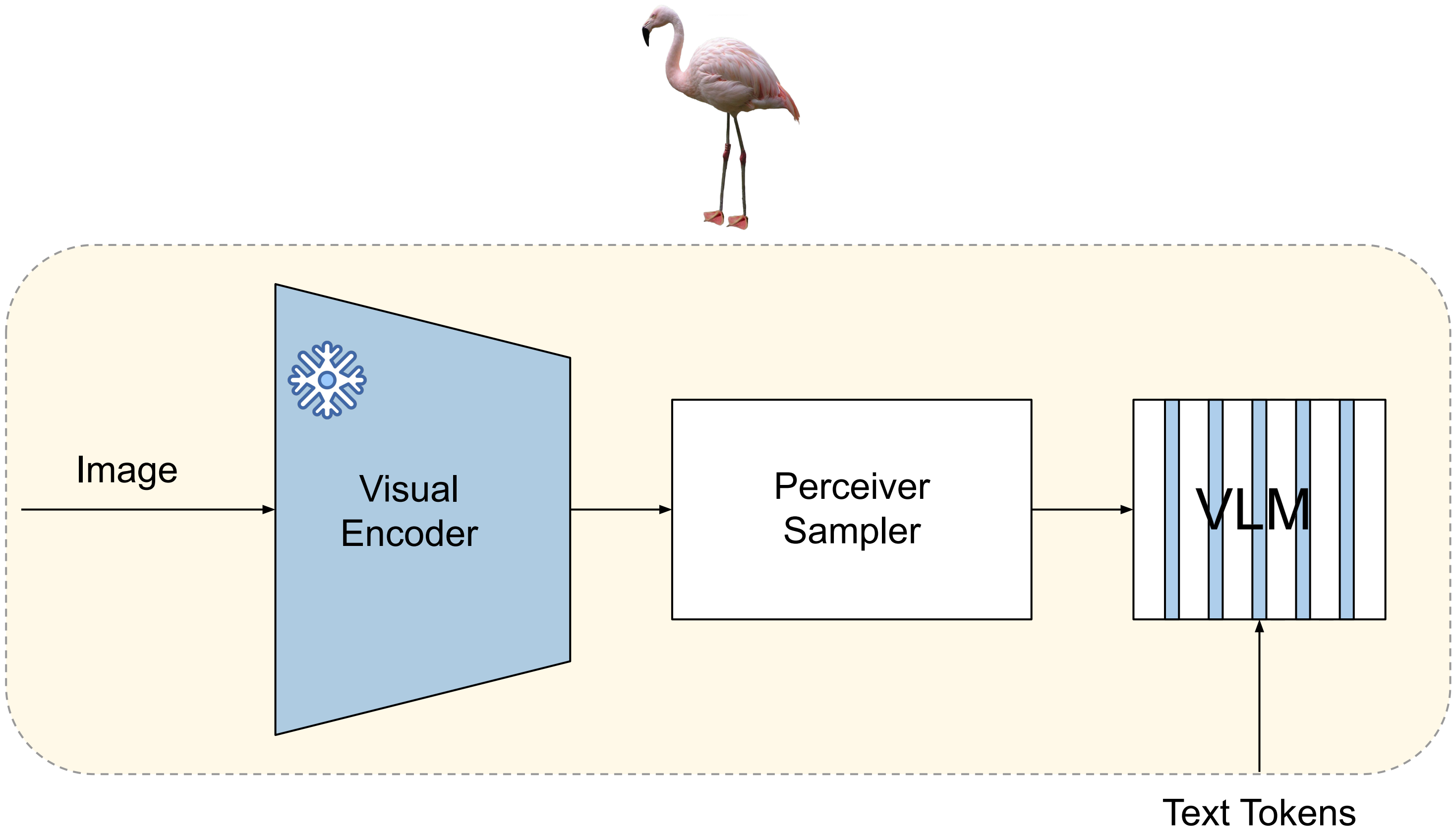


- How to link **interleaved** multi-modal prompts?



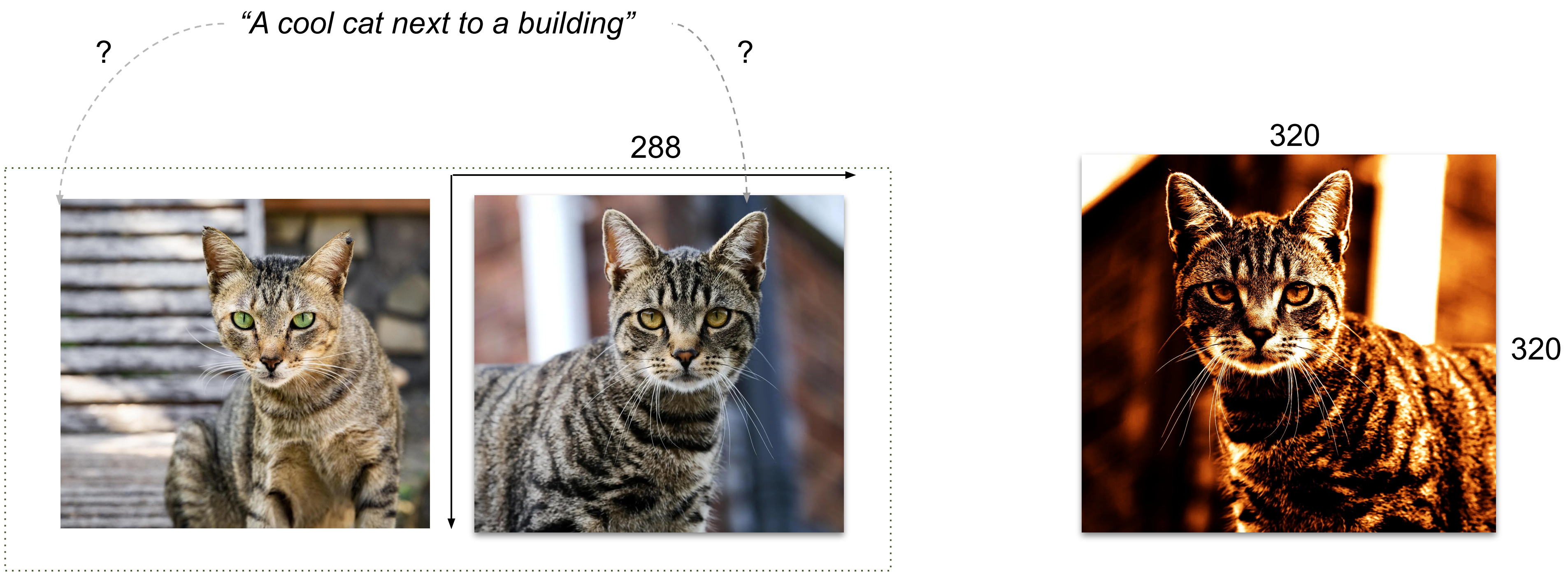
$$\phi : [1, L] \mapsto [0, N]$$

Method



Method - Training Flamingo Models

- Pre-processing & augmentation
  - Random flips, Increased resolution, Color augmentation
  - **Random text links 50% probability.**
  - 8 frames are sampled from training videos.



Method - Training Flamingo Models

- A big chunk is coming from **frozen** LLM. (Chinchilla[1])
- Vision encoder (NFNet-F6) and Perceiver Resampler are same for all.

	Requires model sharding	Frozen Language	Frozen Vision	Trainable GATED XATTN-DENSE	Trainable Resampler	Total count
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**

Method - Training Datasets

- Different data types

Dataset	What Data?	Property
Long Text & Image Pairs (in-house)	312 million Image and text pairs	better quality and longer descriptions
ALIGN[6]	1.8 billion images paired with alt-text	Relatively noisy pairs.
MultiModal MassiveWeb(M3W)	Massive Web dataset, multimodal	***MASSIVE,*** some ambiguous links
Video & Text Pairs (VTP)	22 million short videos with paired text	On average 22 seconds

Method - Data Deduplication

- LTIP and ALIGN are deduplicated (M3W and VTP Not!).

Datasets (**EVALUATION**)	Is deduplicated against?	Used for?
ImageNet	✓	(train, valid)
COCO	✓	(train, valid, test)
OK-VQA	✓	(train, valid, test)
VQAv2	✓	(train, valid, test)
Flickr30k	✓	(valid, test)
VisDial	✓	(valid, test)
VizWiz	❌	(test)
HatefulMemes	❌	(test)
TextVQA	❌	(test)

Method - Training Flamingo Models

- Possible input combinations:

Method - Training Flamingo Models

- Optimizer: AdamW[14]
- **Linear-warm up** followed by a constant learning rate.
- Dataset mixture weights:

Dataset	Weight
M3W	1.0
LTIP	0.2
VTP	0.03
ALIGN	0.2

Paramater	Value
Weight Decay*	0.1
Learning Rate	10^-4

*: No weight decay for Perceiver Resampler

[14] Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Method - Training Flamingo Models

- Loss function
  - Weighted sum of per-dataset expected negative log-likelihoods **of text**, **given the visual inputs**.

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

The weights we discussed

Tex Input

Visual Inputs

Method - Ready for Evaluation!

Evaluation - Established Methods

- **Close-Ended Tasks**: Response from a **pre-defined** space.
    - Text after the query image is used.
    - Final Selection: Beam search.
    - *ex.* Classification
        - Zero-shot
        - Few-shot

- **Open-Ended Tasks**: **Without** a pre-defined response space.
    - Final Selection: log-likelihood
    - *ex.* VQA, Open-ended dialog.
    - Zero or Few-Shot Generalization, new task learning

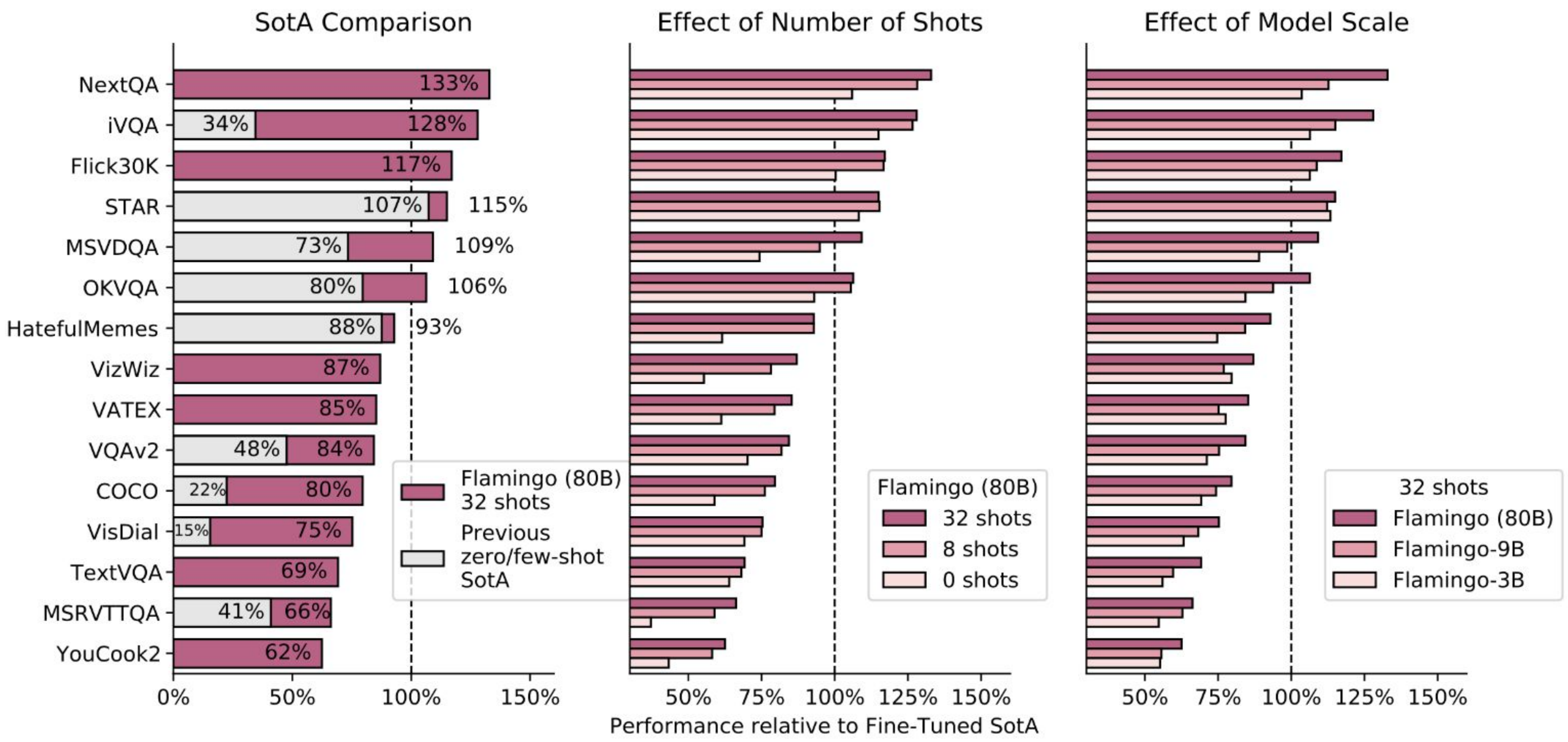
Evaluation - Few-shot Classification

- RICES (Retrieval In-Context Example Selection)
- Flamingo is **not trained with** a contrastive loss.
  - Requires **well distributed training data**.

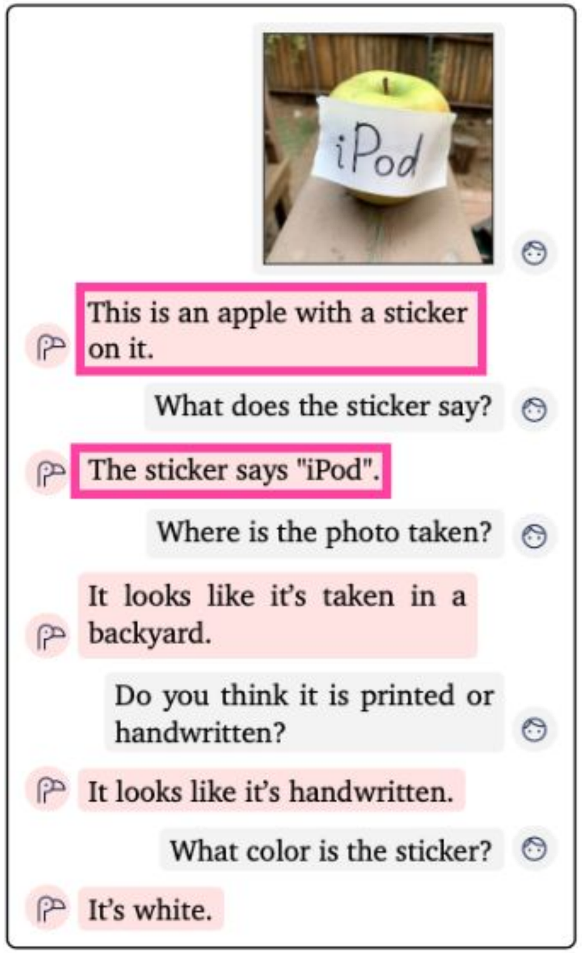
Model	Method	Prompt size	shots/class	ImageNet top 1	Kinetics700 avg top1/5
SotA	Fine-tuned	-	full	91.0 soup	89.0 MTV
SotA	Contrastive	-	0	**85.7** BASIC	**69.6** CLIP
NFNetF6	Our contrastive	-	0	77.9	62.9
Flamingo-3B	RICES	8	1	70.9	55.9
		16	1	71.0	56.9
		16	5	72.7	58.3
Flamingo-9B	RICES	8	1	71.2	58.0
		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

Evaluation

- Superior in 6 tasks
  - Zero-shot, Few-shot image understanding.

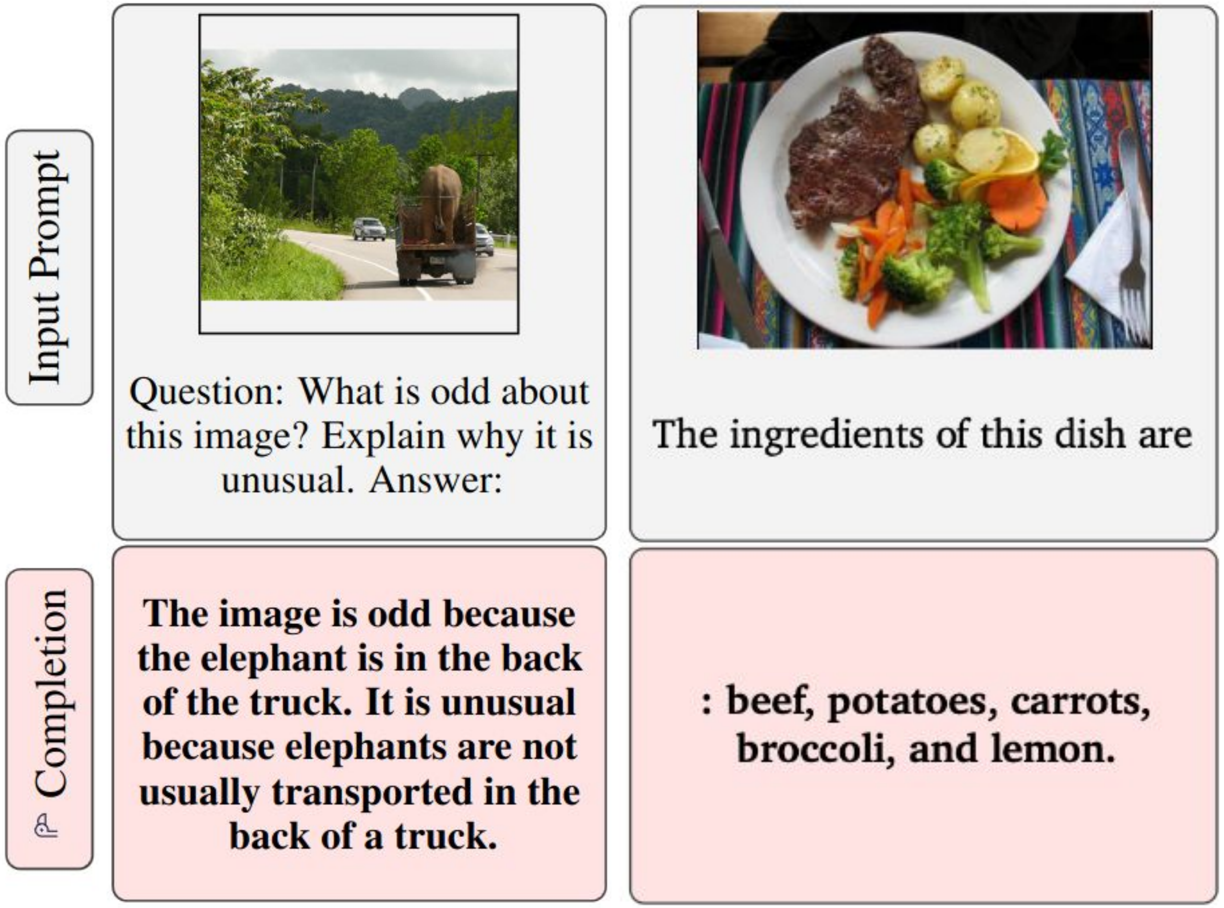


Evaluation - Open Dialog



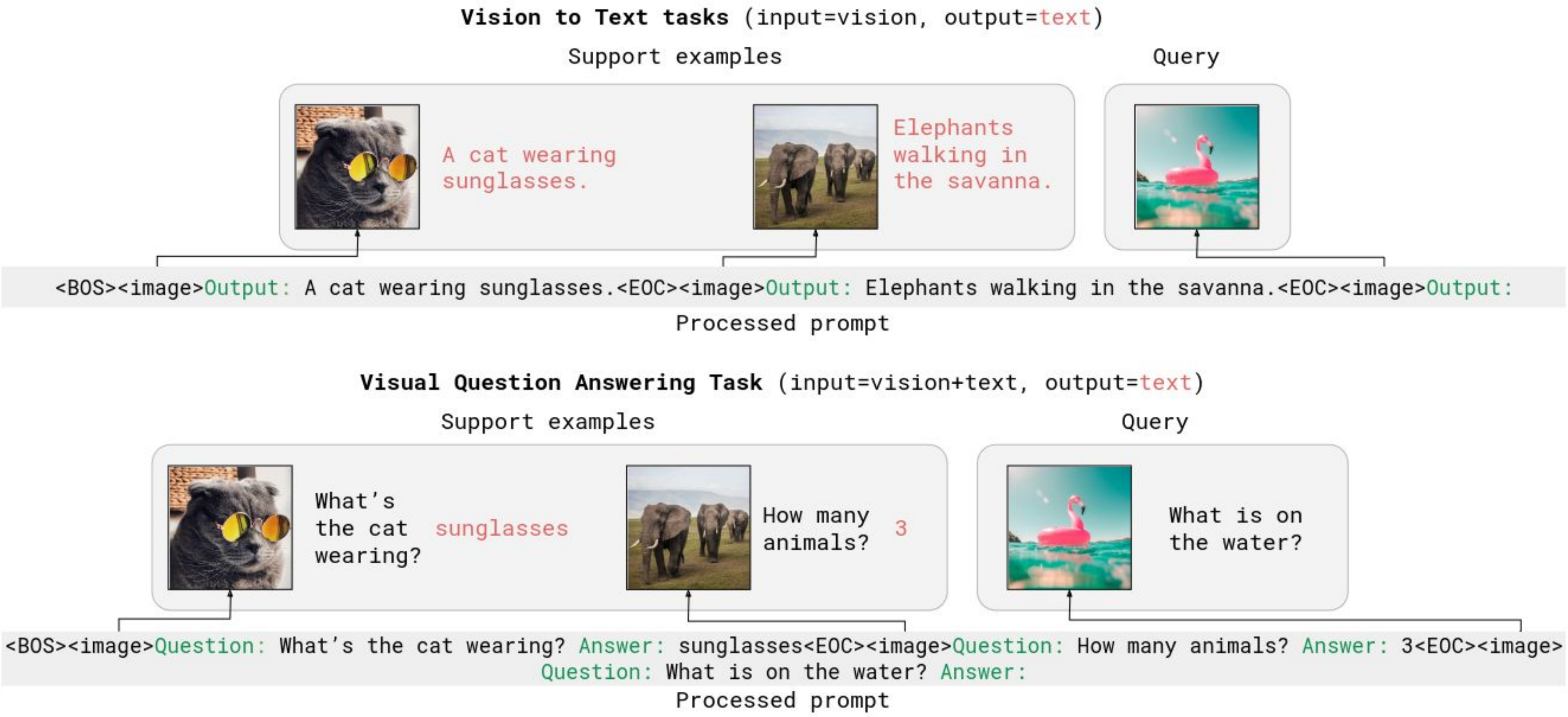
Evaluation - VQA

- Text Completion and VQA Prompts



Evaluation - Few Shot Generalization

- Examples are highly important.
  - Known to have example bias: **the last example has to be good** and relevant.



Evaluation - Few Shot Generalization

- Outperforms **pre-trained** SOTA in some cases.
- ↑ Shot, ↑ Performance

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)	-

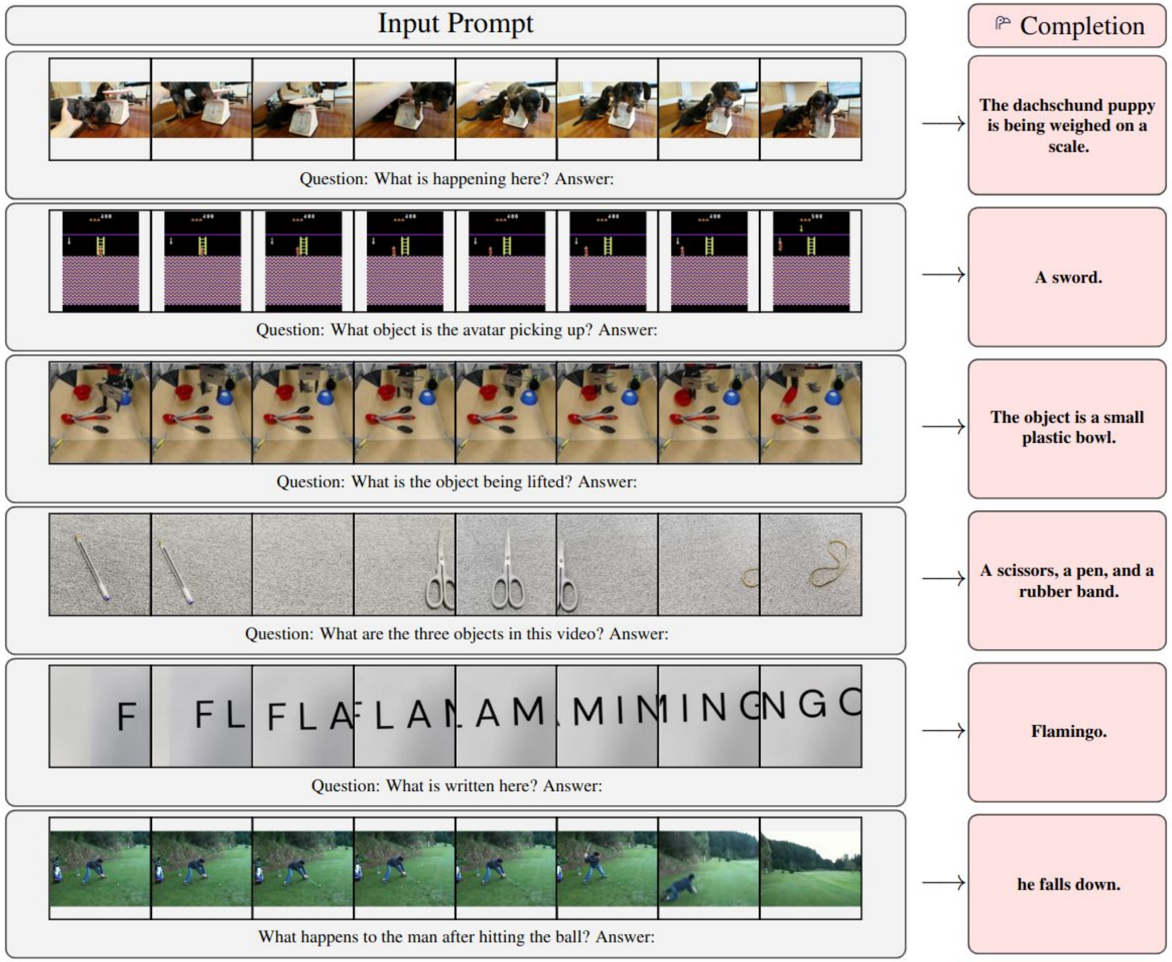
Evaluation - Zero Shot Generalization

- Prompt engineering is important.
  - The way you present the question matters.

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

Evaluation - Video / Text Input

- Videos are **sequence of single images**.



Ablation Studies

- Ablation study on Flamingo-3B Model.

	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 - Dataset Combining Strategy

- **Data merged: Merging** examples from **each dataset**.
- **Round-robin**[15]**: Alternate** examples from **each dataset**.
- **Accumulation: The gradients** from **each dataset** are **weighted** and **summed**.

Dataset	Combination strategy	ImageNet accuracy top-1	COCO image-to-text R@1	COCO image-to-text R@5	COCO image-to-text R@10	COCO text-to-image R@1	COCO text-to-image R@5	COCO text-to-image 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

Bigger but more noisy.

5x Smaller, higher in quality.

[15] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In International Conference on Machine Learning, 2021

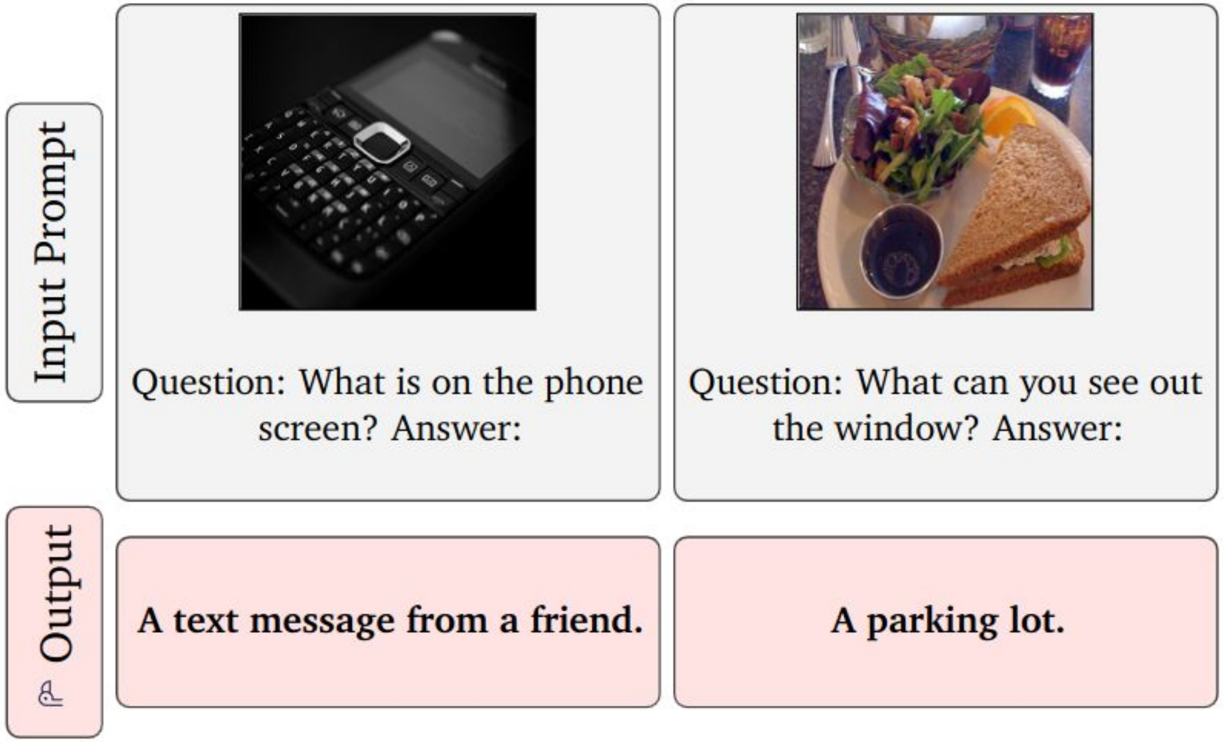
Ablation - Additional

- Additional studies

	Ablated setting	Flamingo 3B value	Changed value	Param. count ↓	Step time ↓	COCO CIDEr↑	OKVQA top1↑	VQAv2 top1↑	MSVDQA top1↑	VATEX CIDEr↑	Overall score↑
	Flamingo 3B model (short training)			3.2B	1.74s	86.5	42.1	55.8	36.3	53.4	**70.7**
(i)	Resampler size	Medium	Small	3.1B	1.58s	81.1	40.4	54.1	36.0	50.2	67.9
			Large	3.4B	1.87s	84.4	42.2	54.4	35.1	51.4	69.0
(ii)	Multi-Img att.	Only last	All previous	3.2B	1.74s	70.0	40.9	52.0	32.1	46.8	63.5
(iii)	p_{next}	0.5	0.0	3.2B	1.74s	85.0	41.6	55.2	36.7	50.6	69.6
			1.0	3.2B	1.74s	81.3	43.3	55.6	36.8	52.7	70.4
(iv)	LM pretraining	MassiveText	C4	3.2B	1.74s	81.3	34.4	47.1	60.6	53.9	62.8
(v)	Freezing Vision	✓	✗ (random init)	3.2B	4.70s*	74.5	41.6	52.7	31.4	35.8	61.4
			✗ (pretrained)	3.2B	4.70s*	83.5	40.6	55.1	34.6	50.7	68.1
(vi)	Co-train LM on MassiveText	✗	✓ (random init)	3.2B	5.34s*	69.3	29.9	46.1	28.1	45.5	55.9
			✓ (pretrained)	3.2B	5.34s*	83.0	42.5	53.3	35.1	51.1	68.6
(vii)	Dataset and Vision encoder	M3W+ITP+VTP and NFNetF6	LAION400M and CLIP	3.1B	0.86s	61.4	37.9	50.9	27.9	29.7	54.7
			M3W+LAION400M+VTP and CLIP	3.1B	1.58s	76.3	41.5	53.4	32.5	46.1	64.9

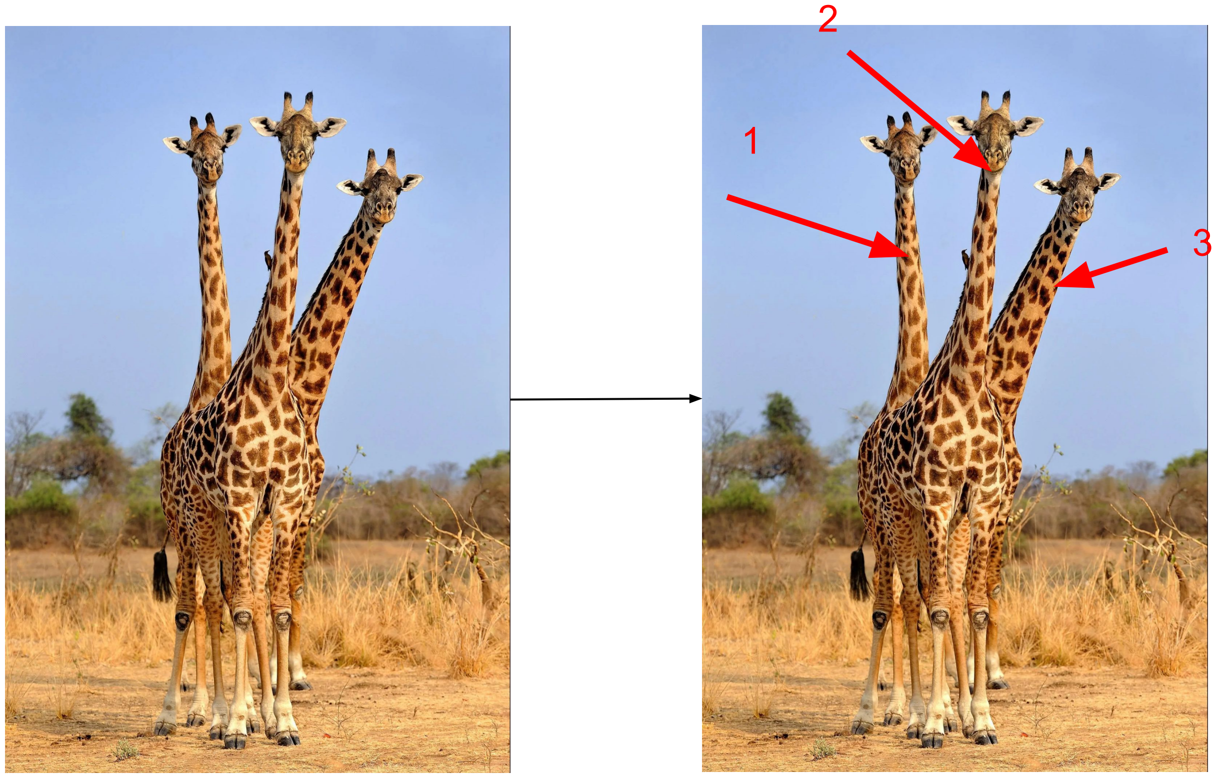
Discussion - Limitations

- Inherits **weaknesses** of LLMs.
    - **Hallucinations** and ungrounded guesses.
    - Fixed number of tokens.
    - Bad sample efficiency.



Discussion - Limitations

- Limited Visual and language interface
    - No visual context about the output prompt.

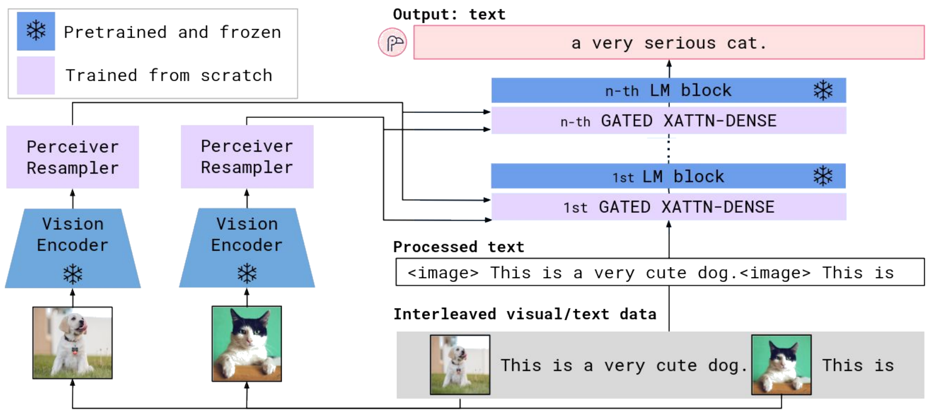


“How many giraffe?”

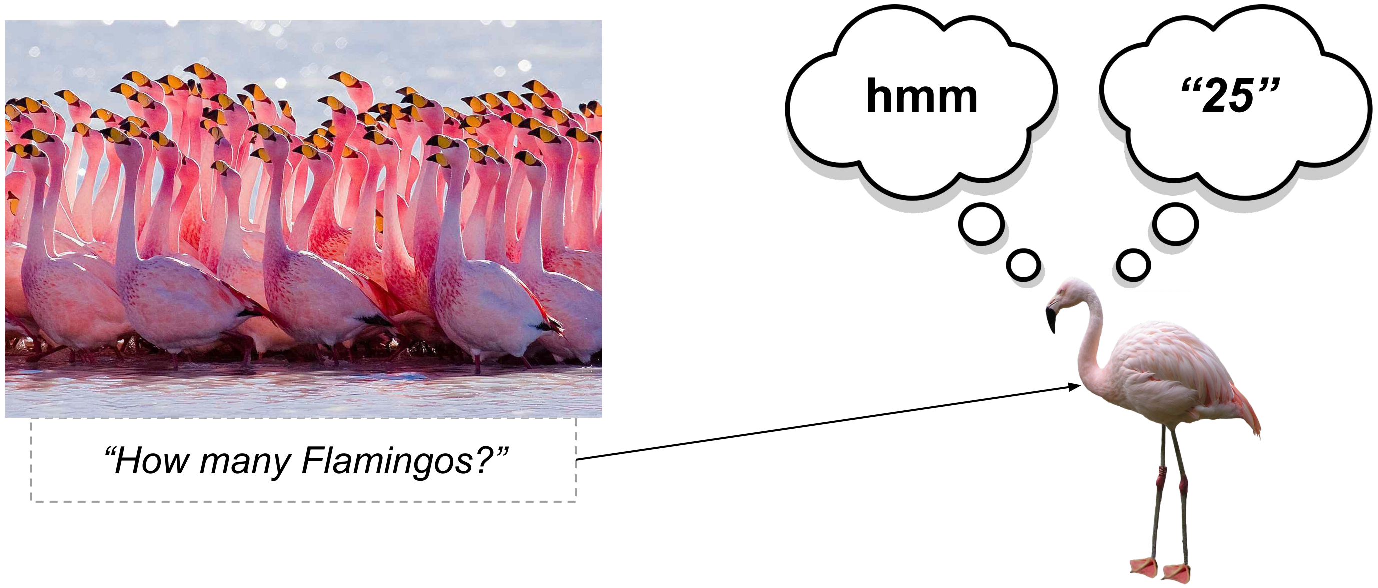
“3”

Conclusion

- A framework on how to extend LLMs -> VLMs.
  - Cross-attention allows for VLM extension!
- Perceiver Sampler based fixed tokens successful!.
  - Video input enabled!
- Data size matters.
  - Data **quality** matters more.
- Could perform better than **Fine Tuned** SOTA!
- Direct **inheritance** of **bad habits** of LLMs
  - Racism
  - Hallucinations



Thank you for listening!



References

- [1] Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., ... & Sifre, L. (2022). Training compute-optimal large language models. arXiv preprint arXiv:2203.15556.
- [2] Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... & Simonyan, K. (2022). Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems, 35, 23716-23736.
- [3] Li, J., Li, D., Savarese, S., & Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597.
- [4] Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. arXiv preprint arXiv:1906.02243.
- [5] Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., ... & Irving, G. (2021). Scaling language models: Methods, analysis & insights from training gopher. arXiv preprint arXiv:2112.11446.
- [6] Jia, C., Yang, Y., Xia, Y., Chen, Y. T., Parekh, Z., Pham, H., ... & Duerig, T. (2021, July). Scaling up visual and vision-language representation learning with noisy text supervision. In International Conference on Machine Learning (pp. 4904-4916). PMLR.
- [7] (CLIP) A. Radford et al., "Learning transferable visual models from natural language supervision", ICML (2021)
- [8] (CM3) A. Aghajanyan et al., "CM3: A Causal Masked Multimodal Model of the Internet", arxiv (2022)
- [9] (VLKD) Dai, W., Hou, L., Shang, L., Jiang, X., Liu, Q., & Fung, P. (2022). Enabling multimodal generation on CLIP via vision-language knowledge distillation. arXiv preprint arXiv:2203.06386.
- [10] (Frozen) M. Tsimpoukelli et al., "Multimodal few-shot learning with frozen language models", NeurIPS (2021)
- [11] Brock, A., De, S., Smith, S. L., & Simonyan, K. (2021, July). High-performance large-scale image recognition without normalization. In International Conference on Machine Learning (pp. 1059-1071). PMLR.
- [12] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- [13] Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., & Carreira, J. (2021, July). Perceiver: General perception with iterative attention. In International conference on machine learning (pp. 4651-4664). PMLR.
- [14] Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Appendix - Compute

- A big chunk is coming from LLM.
- Vision encoder is same for all. (NFNet-F6)

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

L: Layers, **D**: Transformer Hidden Size, **H**: Number of heads

Appendix - Training the Image Encoder

- Details:
  - ALIGN and LTIP datasets
  - Resolution: 288 x 288
  - Embedding Size: 1376
  - Adam Opt.
  - Gradient clipping

- Evaluation
  - Zero-shot image classification -> Image-text retrieval

- Why use BERT?
  - To be able to extract contextual features rather than pure geometric features.
  - **If trained with a LLM** it generalizes **better as a visual conditioner**.

Appendix - Training the Image Encoder

- Trained from scratch with BERT language encoder
  - Text-to-image contrastive loss

$$L_{\text{contrastive:txt2im}} = -\frac{1}{N}\sum_{i}^{N}\log\left(\frac{\exp\left(L_i^\top V_i \beta\right)}{\sum_j^N \exp\left(L_i^\top V_j \beta\right)}\right)$$

  - Image-to-text contrastive loss

$$L_{\text{contrastive:im2txt}} = -\frac{1}{N}\sum_{i}^{N}\log\left(\frac{\exp\left(V_i^\top L_i \beta\right)}{\sum_j^N \exp\left(V_i^\top L_j \beta\right)}\right)$$

Trainable inverse temperature parameter

Appendix - Visually Conditioned Large Language Models

- Examples are from Flamingo [2]