max planck institut
informatik

SIC Saarland Informatics Campus

# High Level Computer Vision

# Vision Language Models

# @ July 10, 2024

**Bernt Schiele**

**cms.sic.saarland/hlcvss24/**

**Max Planck Institute for Informatics & Saarland University,**
**Saarland Informatics Campus Saarbrücken**

# Overview of Today's Lecture

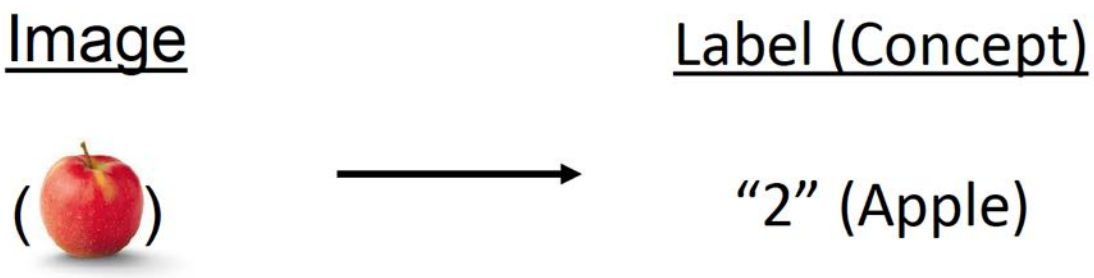
- Vision Language Learning for Computer Vision
  - Supervised learning vs. vision-language learning
  - CLIP [icml'21] - https://arxiv.org/abs/2103.00020
  - ALIGN [icml'21] - https://arxiv.org/abs/2102.05918
  - Modality gap discussion — https://arxiv.org/abs/2404.07983 (Apr'24)

- Large Vision Language Models — Leveraging Large Language Models
  - Flamingo [neurips'22] — https://arxiv.org/abs/2204.14198
  - Gemini 1.0 — https://arxiv.org/abs/2312.11805 (Dec'23 & Jun'24)
  - Gemini 1.5 — https://arxiv.org/abs/2403.05530 (Mar'24 & Jun'24)

slide credit: Jianwei Yang

# Supervised Learning

Map an image to a discrete label
which is associated a visual concept



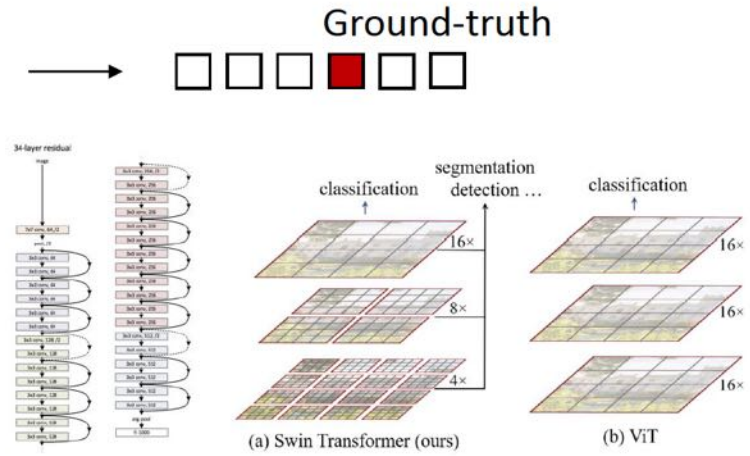
slide credit: Jianwei Yang

# Supervised Learning

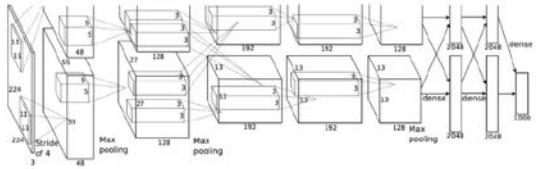
**MNIST**. LeCun *et al.*

**CIFAR-10**. Krizhevsky *et al.*


**ImageNet**. Deng *et al.*


**ResNet**. He *et al.* **Swin**. Liu *et al.*



**AlexNet.** Krizhevsky *et al.*

slide credit: Jianwei Yang

# Supervised Learning

- Pros
  - Densely labeled samples for each category
- Cons
  - Requires a lot of human effort
  - Limited number of categories

slide credit: Jianwei Yang

# Zero-Shot Learning (Canonical)

Map an image to description of a visual concept

| Image | | Descriptions (Concept) |
| --- | --- | --- |
|  | → | Fruit, Red, Sphere (Apple) |
|  | → | Fruit, Yellow (Orange) |

slide credit: Jianwei Yang

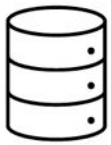
# Zero-Shot Learning (Canonical)

**CUB-200-2011**. Wah *et al.*

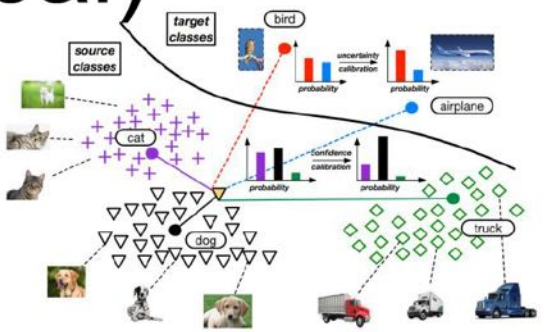
**AwA2**. Xian *et al.*

**aPY**. Farhadi *et al.*

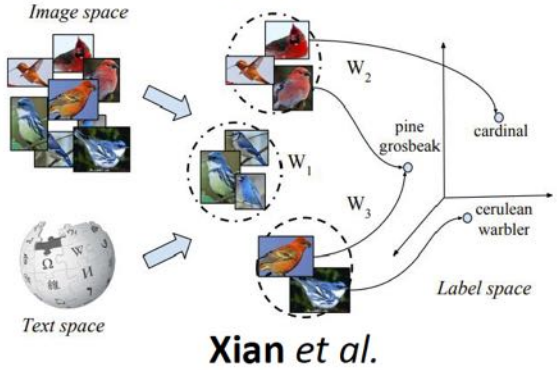
Label & descriptions



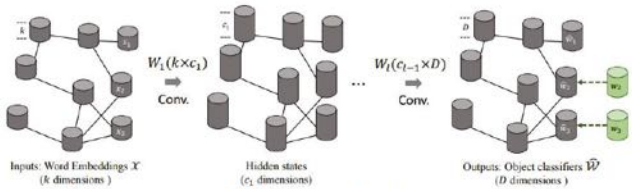
images




**Liu** *et al.*




**Xian** *et al.*




**Wang** *et al.*

slide credit: Jianwei Yang

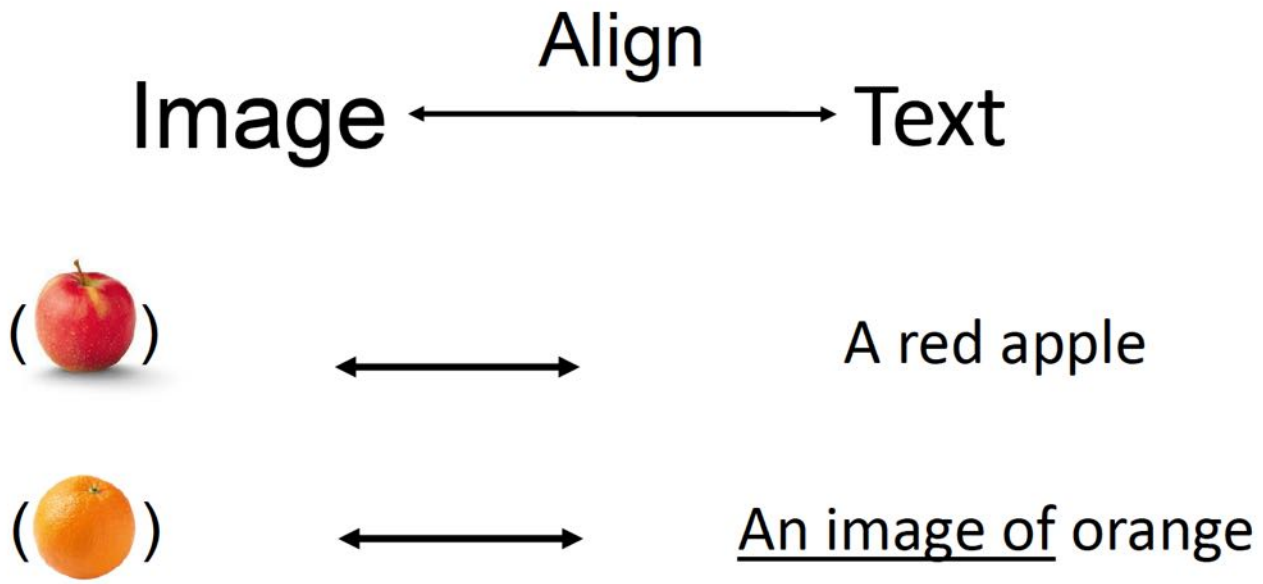
# Zero-Shot Learning (Canonical)

- Pros
  - Directly learn the visual-semantic matching

- Cons
  - Small scale with limited vocabulary
  - Fixed visual and text encoder

slide credit: Jianwei Yang

# Contrastive Vision-Language Learning



slide credit: Jianwei Yang

# Contrastive Vision-Language Learning

Align

Image ⟷ Text

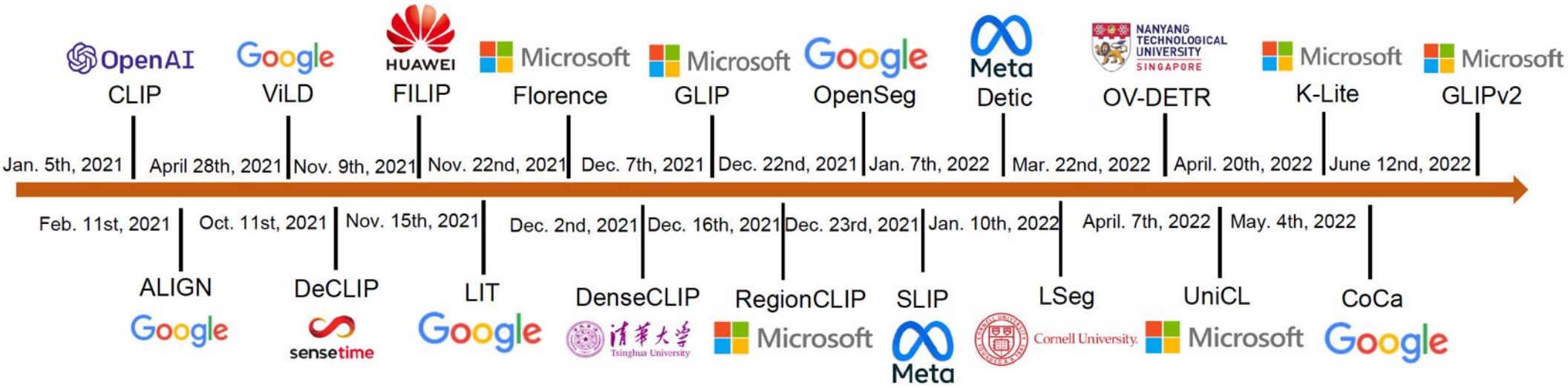
End-to-end learning on large-scale corpus

(  ) ⟷ An image of orange

slide credit: Jianwei Yang

# The most recent art

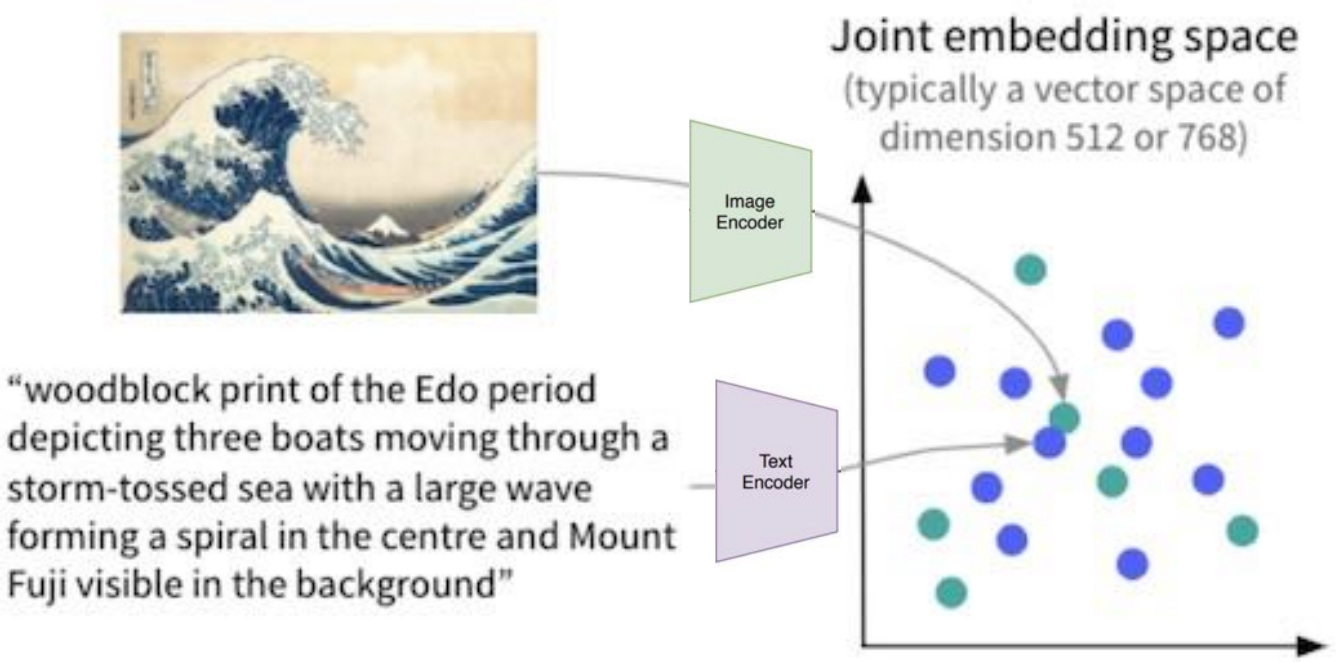
## Contrastive Vision-Language Learning




A lot of research works come along the line of vision-language learning for vision

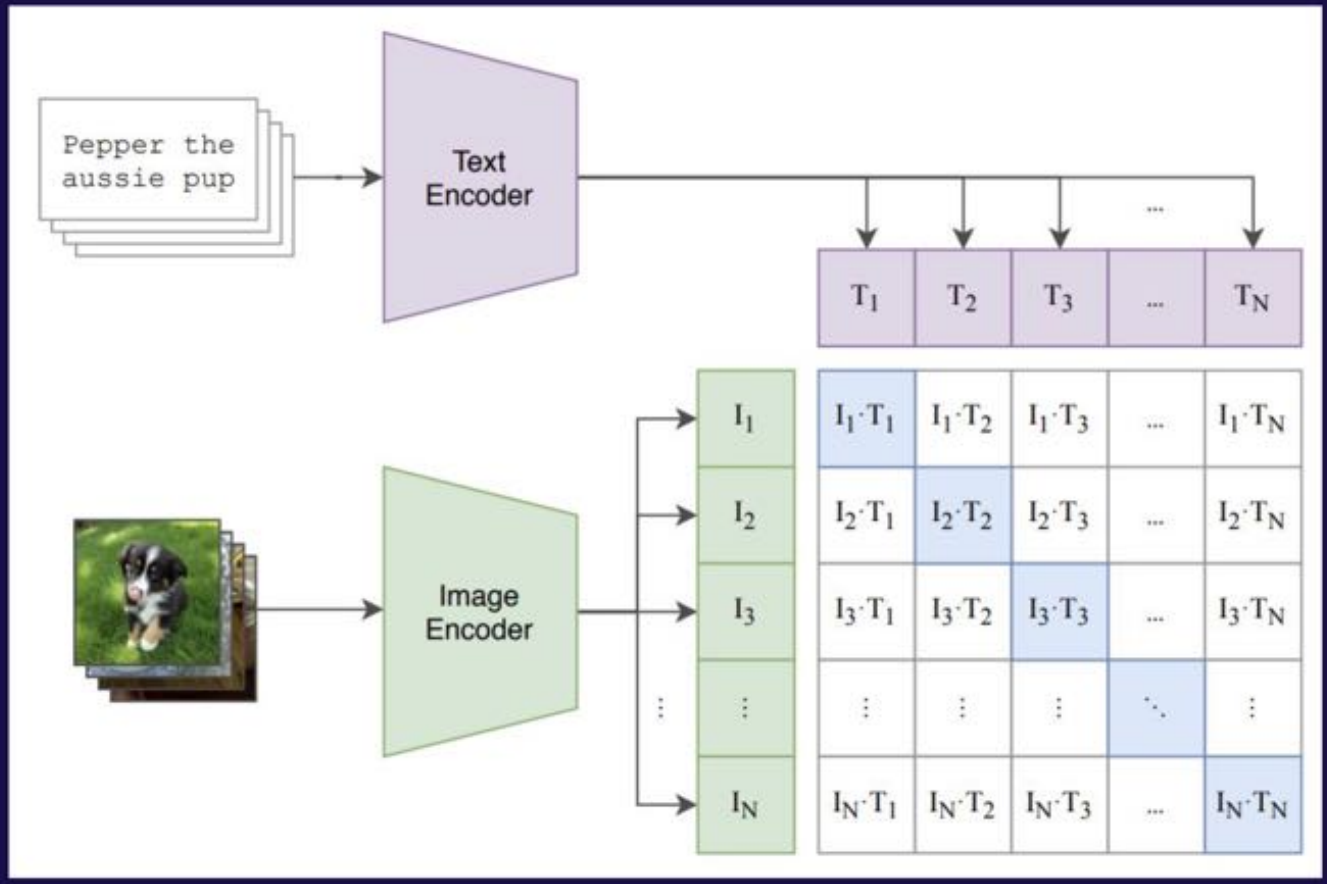
# Key Idea — Learning of a Joint Embedding Space

- Map both Images and Text into the same Embedding Space:



Learning Transferable Visual Models From Natural Language Supervision: Radford et al@ICML'21

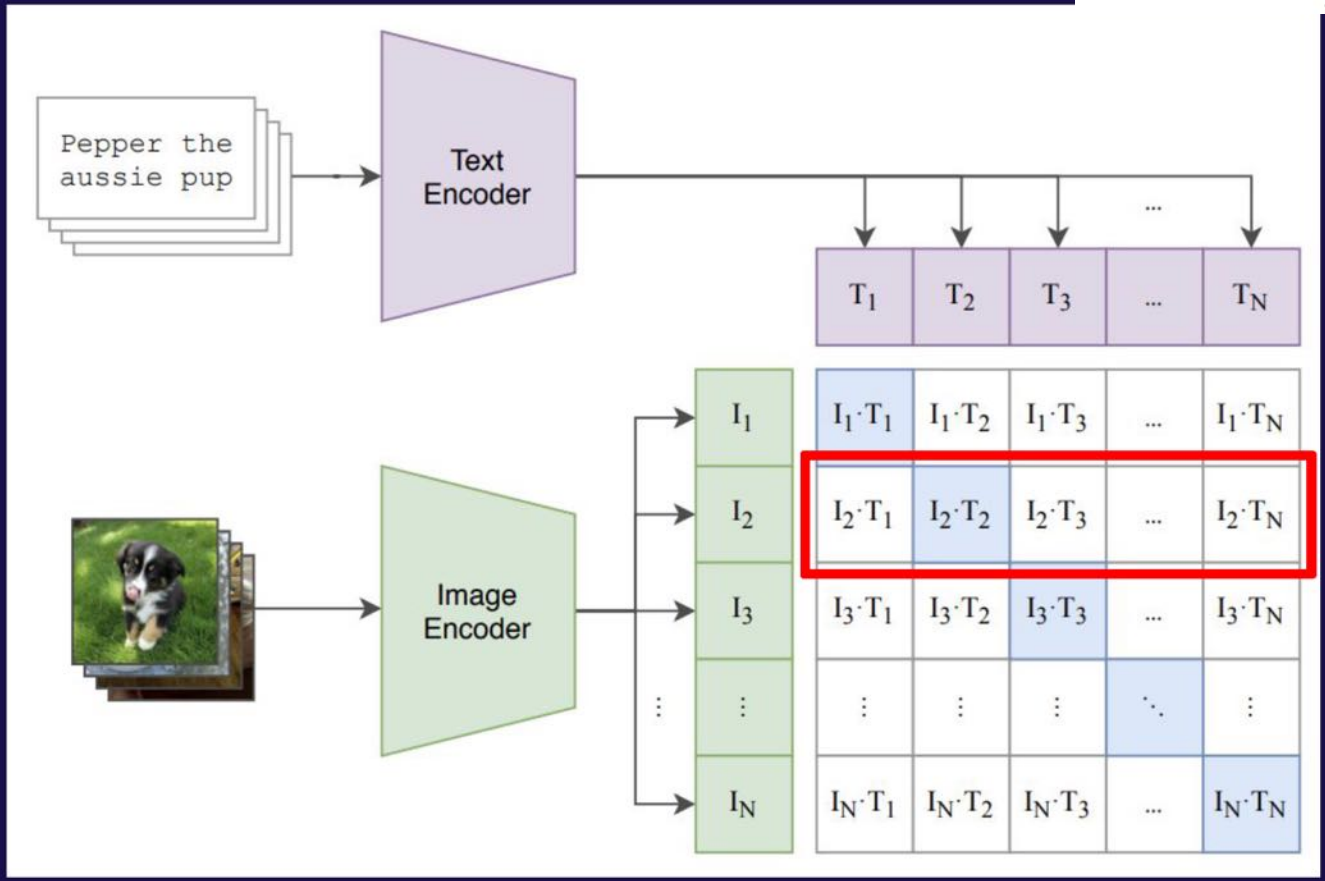
# CLIP: Contrastive Language-Image Pre-training



Learning Transferable Visual Models From Natural Language Supervision: Radford et al@ICML'21

# CLIP: Contrastive Language-Image Pre-training

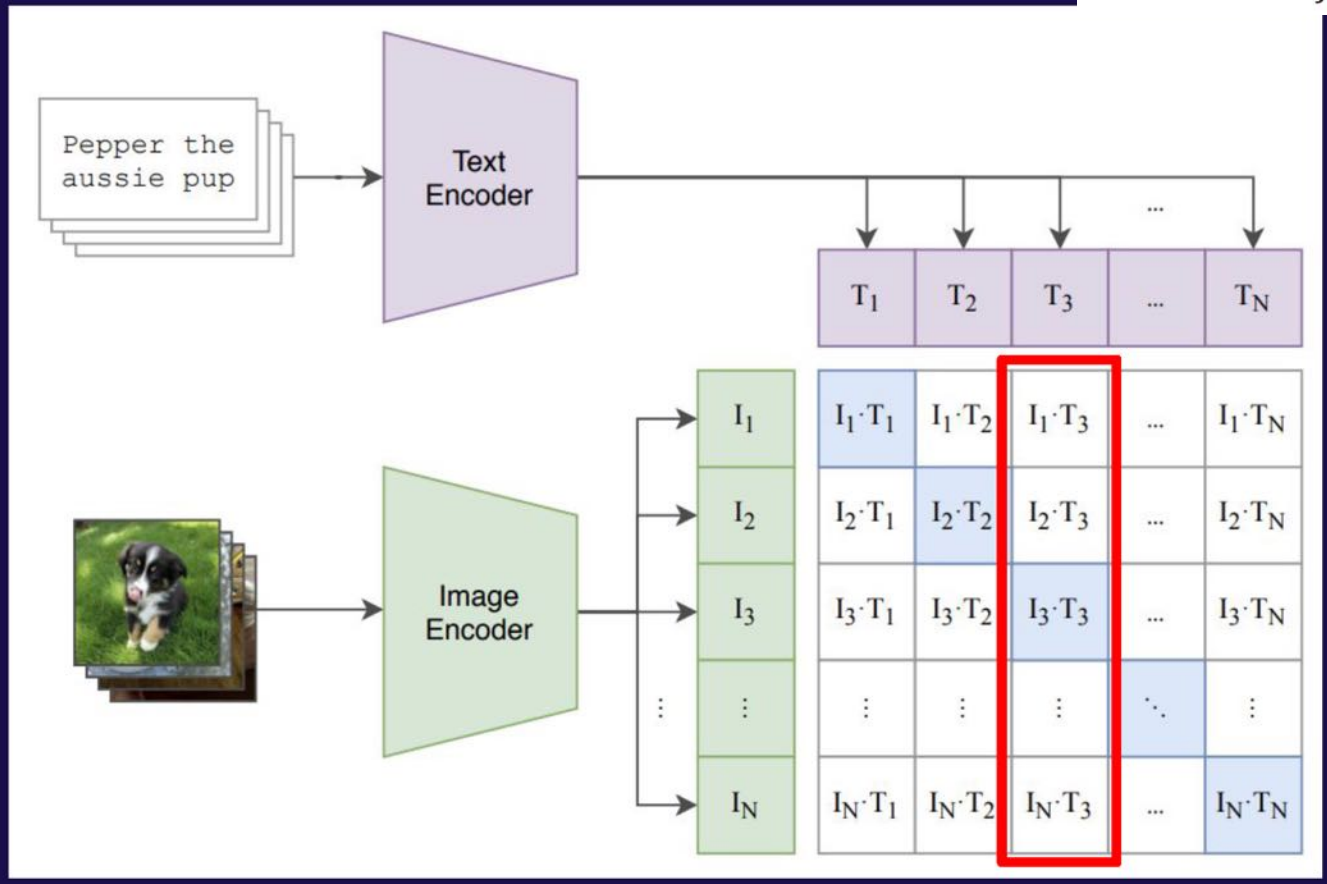
$$L_{i2t} = -\sum_j \log \frac{\exp I_j T_j^T}{\sum_k \exp I_j T_k^T}$$



Learning Transferable Visual Models From Natural Language Supervision: Radford et al@ICML'21

# CLIP: Contrastive Language-Image Pre-training

$$L_{t2i} = -\sum_j \log \frac{\exp I_j T_j^T}{\sum_k \exp I_k T_j^T}$$



Learning Transferable Visual Models From Natural Language Supervision: Radford et al@ICML'21

## Some CLIP details

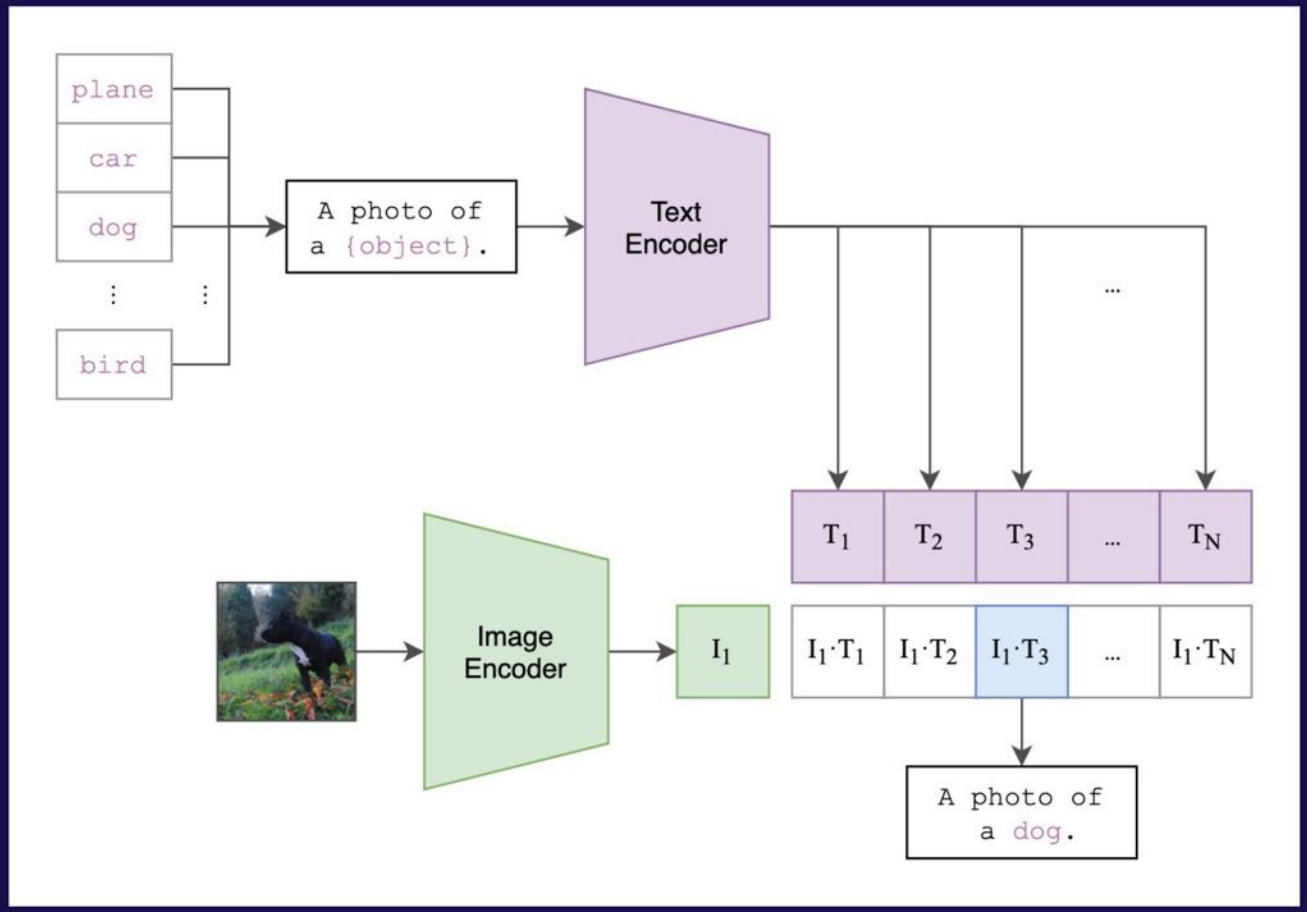
Training

- Trained on 400M image-text pairs from the internet
- Batch size of 32,768
- 32 epochs over the dataset
- Cosine learning rate decay

Architecture

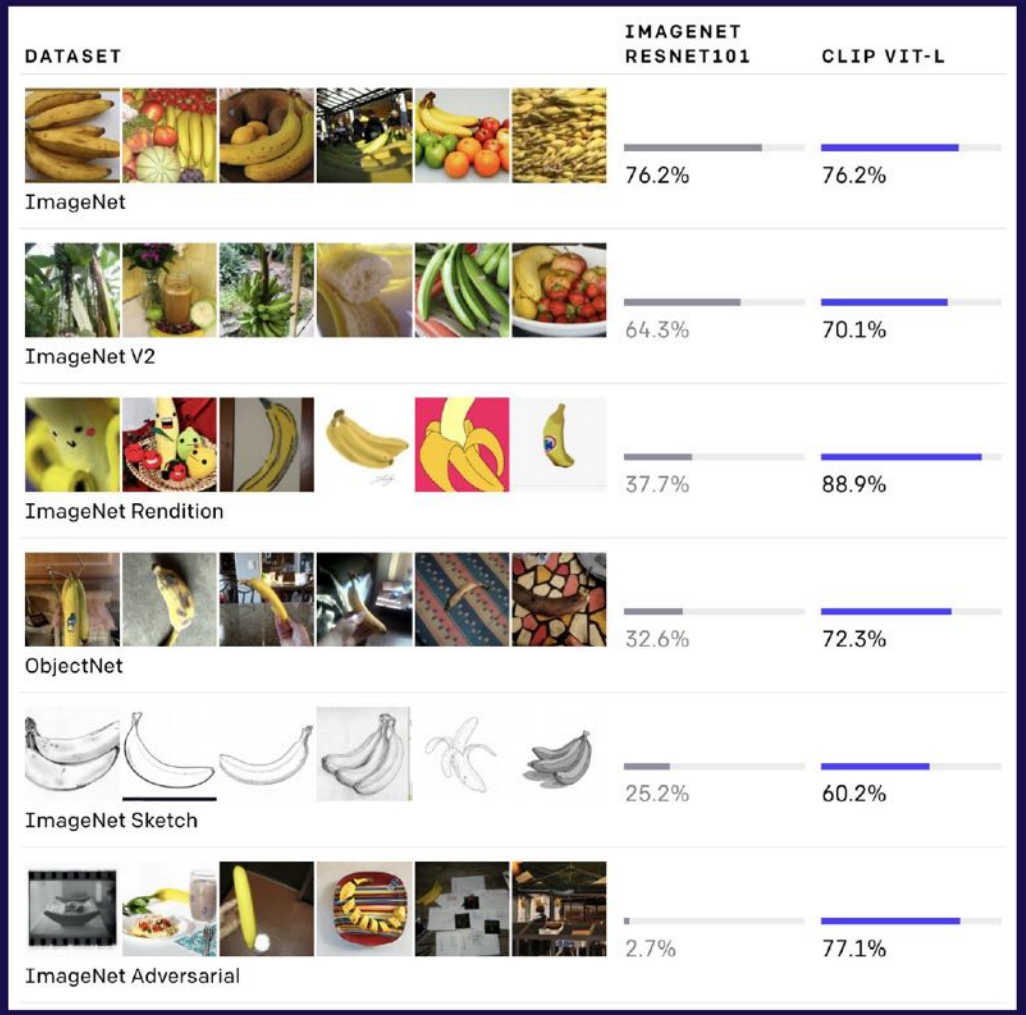
- ResNet-based or ViT-based image encoder
- Transformer-based text encoder

# Zero-shot image classification



Learning Transferable Visual Models From Natural Language Supervision: Radford et al@ICML'21

# Zero-shot CLIP is much more robust



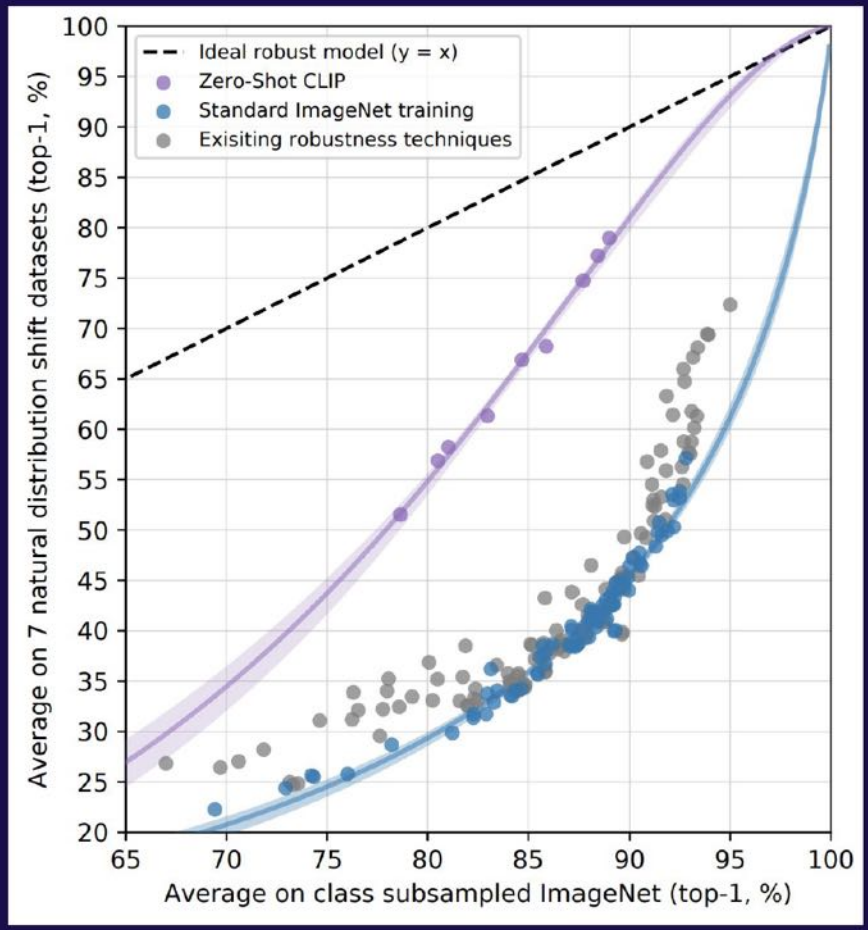
Learning Transferable Visual Models From Natural Language Supervision: Radford et al@ICML'21

# Robustness to natural distribution shift

CLIP is significantly more robust!

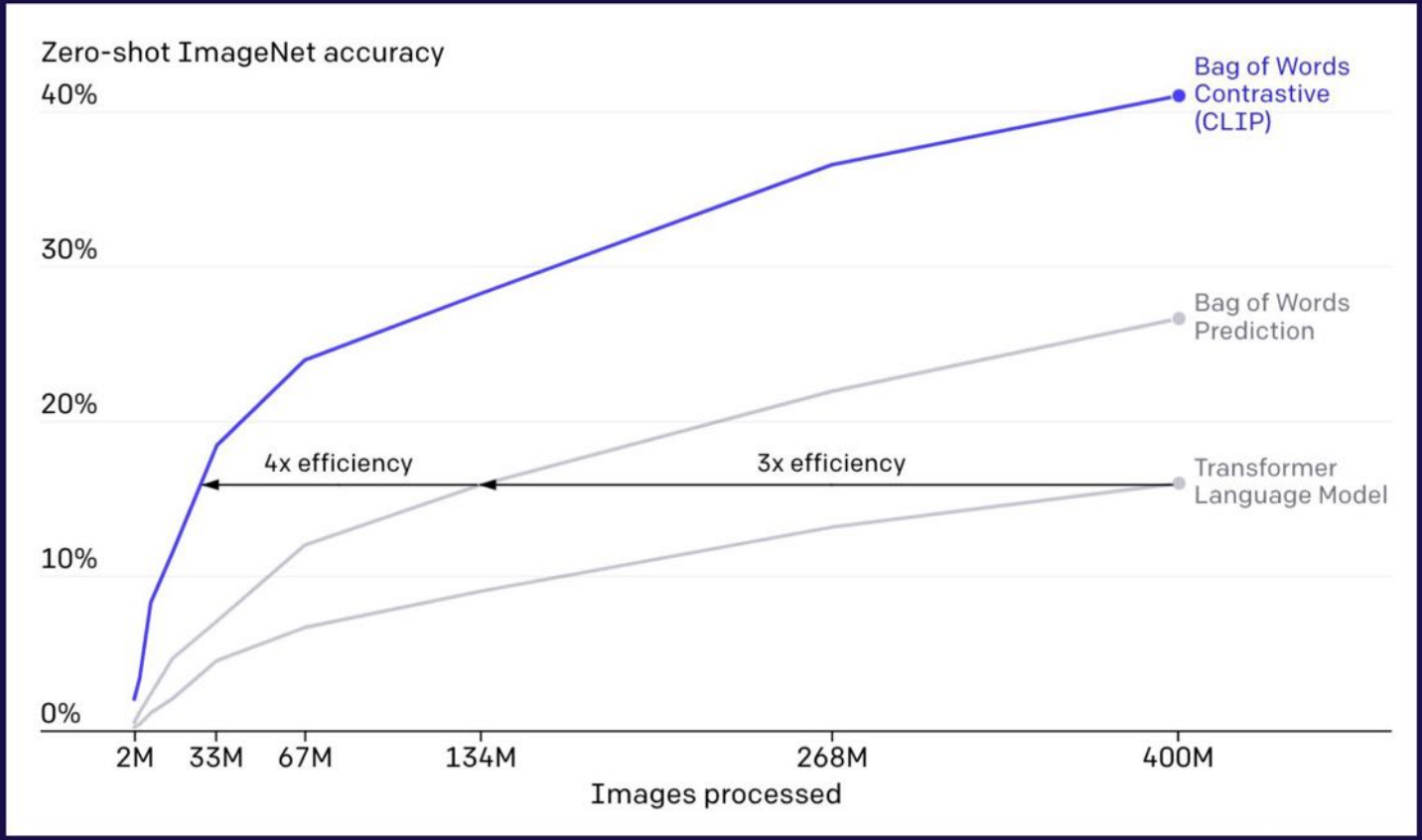
7 ImageNet-like Datasets (Taori et al.)

- ImageNetV2
- ImageNet-A
- ImageNet-R
- ImageNet Sketch
- ObjectNet
- ImageNet Vid
- Youtube-BB



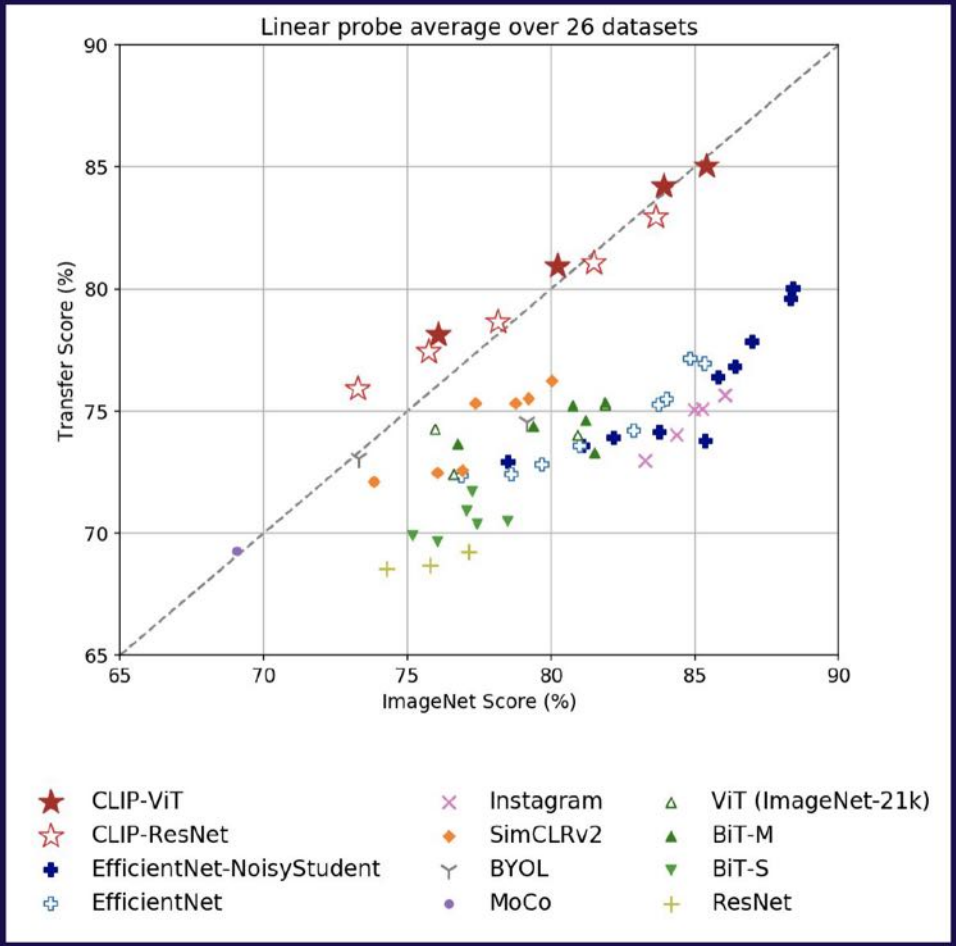
Learning Transferable Visual Models From Natural Language Supervision: Radford et al@ICML'21

# Why contrastive



Learning Transferable Visual Models From Natural Language Supervision: Radford et al@ICML'21

# vs ImageNet score

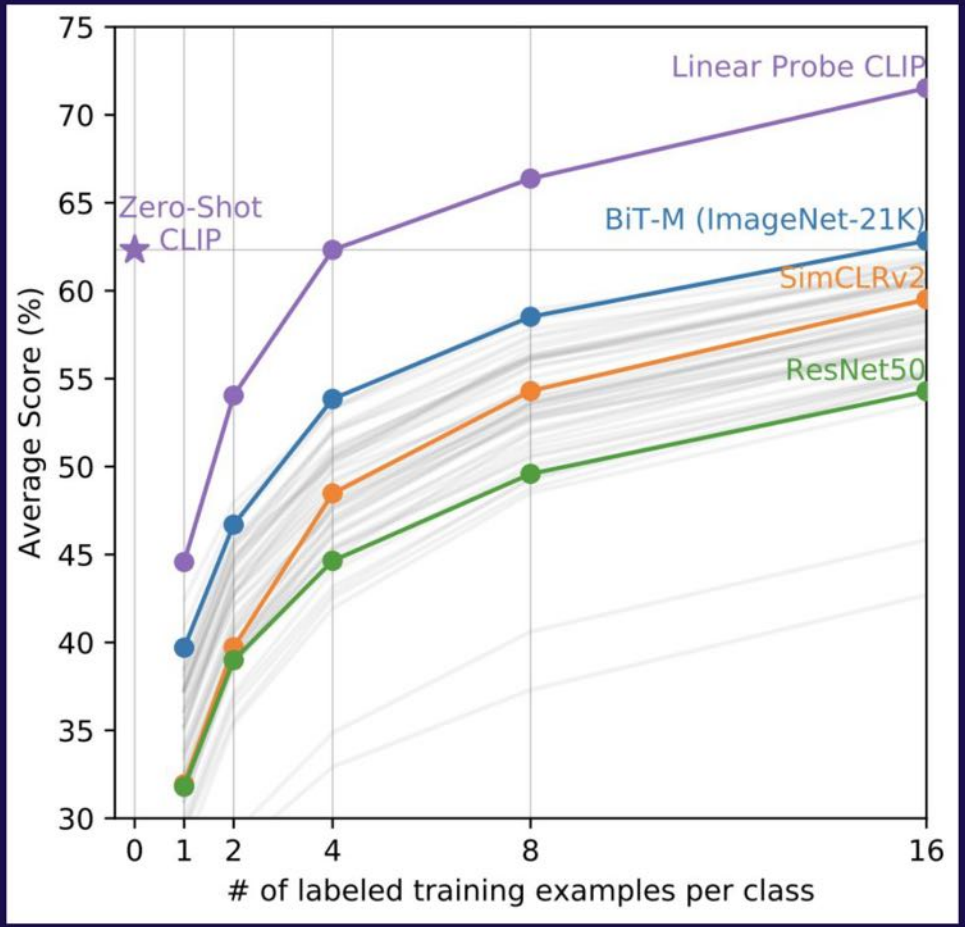


Learning Transferable Visual Models From Natural Language Supervision: Radford et al@ICML'21

# Zero-shot CLIP vs Few-shot linear probes

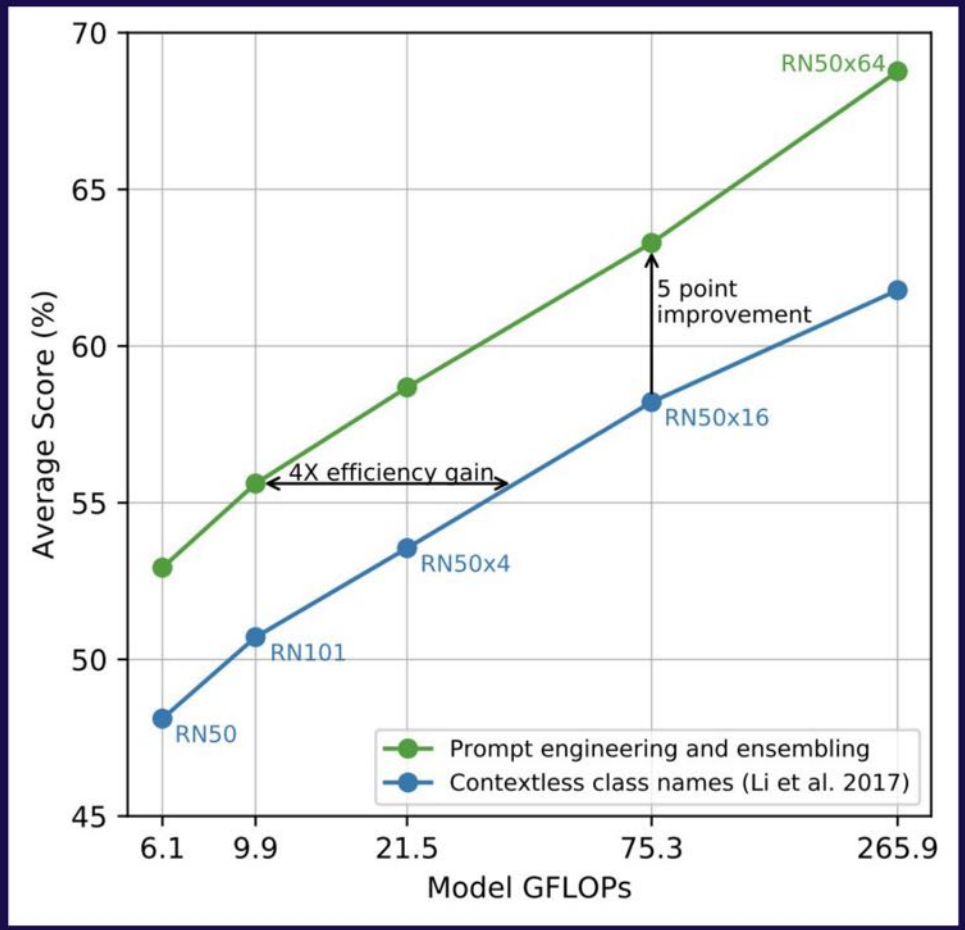
Zero-shot CLIP is as good as

- 4-shot linear-probe CLIP
- 16-shot BiT-M



Learning Transferable Visual Models From Natural Language Supervision: Radford et al@ICML'21

# Prompt engineering



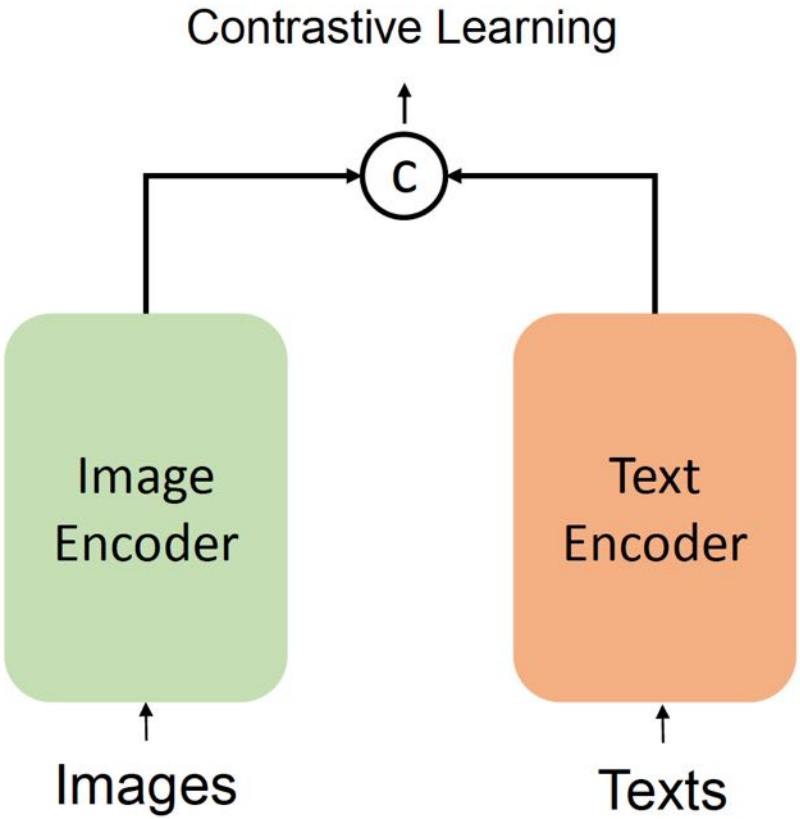
slide credit: Jianwei Yang

# The Lesson from CLIP

- Image recognition can be formulated as an **image-text matching problem** instead of image-label mapping problem
- Image recognition does not require human-annotated image-label data but **huge amount of (noisy) image-text pairs**
- **Contrastive learning is a good learning objective** for multi-modal learning strategy compared with generative learning
- **Two-tower model without fusion** is sufficient to learn good and generic visual and language representations

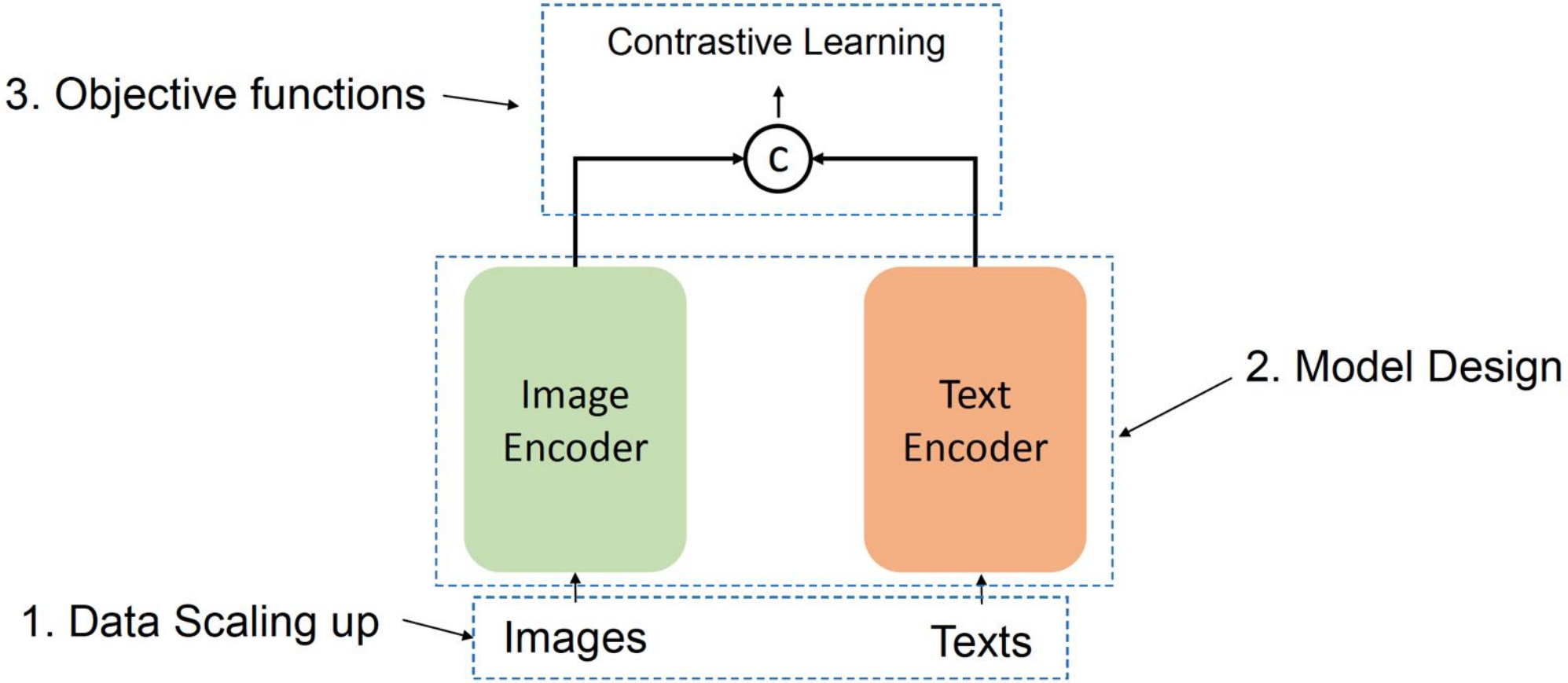
slide credit: Jianwei Yang

# The most recent art



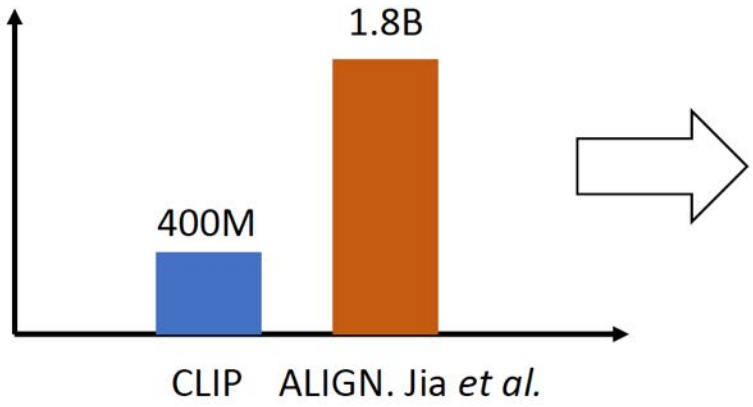
slide credit: Jianwei Yang

# The most recent art



ALIGN = Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision - Jia et al@ICML'21 slide credit: Jianwei Yang

# Data Scaling-Up




## Larger scale but more noisy data


"motorcycle front wheel"

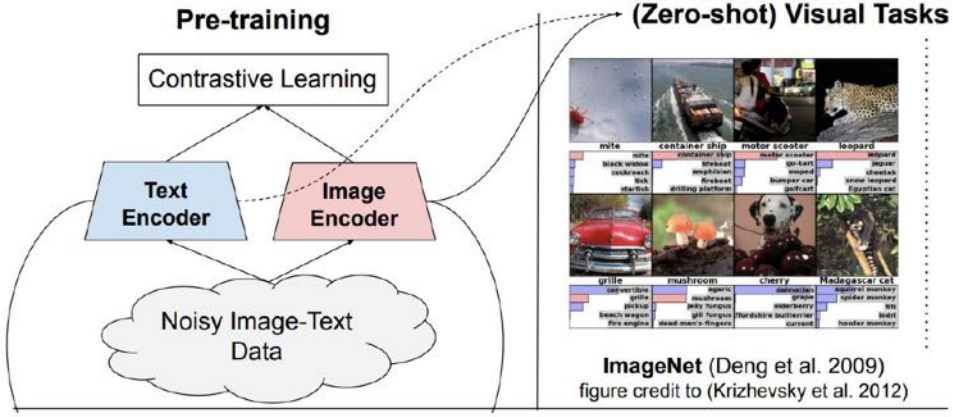

"*thumbnail for version as of 21 57 29 june 2010*"


"file frankfurt airport skyline 2017 05 jpg"


"file london barge race 2 jpg"


"moustache seamless wallpaper design"


"st oswalds way and shops"




| Model | ImageNet | ImageNet-R | ImageNet-A | ImageNet-V2 |
|---|---|---|---|---|
| CLIP | 76.2 | 88.9 | **77.2** | **70.1** |
| **ALIGN** | **76.4** | **92.2** | 75.8 | **70.1** |

## Zero-shot image classification on ImageNet

| Model (backbone) | Acc@1 w/ frozen features | Acc@1 | Acc@5 |
|---|---|---|---|
| WSL (ResNeXt-101 32x48d) | 83.6 | 85.4 | 97.6 |
| CLIP (ViT-L/14) | 85.4 | - | - |
| BiT (ResNet152 x 4) | - | 87.54 | 98.46 |
| NoisyStudent (EfficientNet-L2) | - | 88.4 | 98.7 |
| ViT (ViT-H/14) | - | 88.55 | - |
| Meta-Pseudo-Labels (EfficientNet-L2) | - | **90.2** | **98.8** |
| **ALIGN** (EfficientNet-L2) | **85.5** | 88.64 | 98.67 |

## Image classification finetuning

# Overview of Today's Lecture

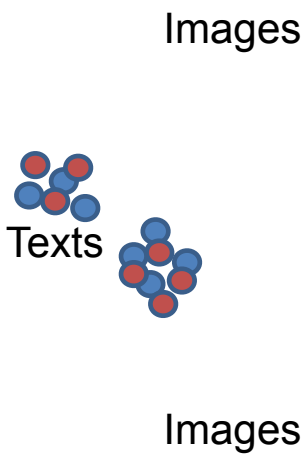
- Vision Language Learning for Computer Vision
  - Supervised learning vs. vision-language learning
  - CLIP [icml'21] - https://arxiv.org/abs/2103.00020
  - ALIGN [icml'21] - https://arxiv.org/abs/2102.05918
  - Modality gap discussion — https://arxiv.org/abs/2404.07983 (Apr'24)

- Large Vision Language Models — Leveraging Large Language Models
  - Flamingo [neurips'22] — https://arxiv.org/abs/2204.14198
  - Gemini 1.0 — https://arxiv.org/abs/2312.11805 (Dec'23 & Jun'24)
  - Gemini 1.5 — https://arxiv.org/abs/2403.05530 (Mar'24 & Jun'24)

# Modality gap

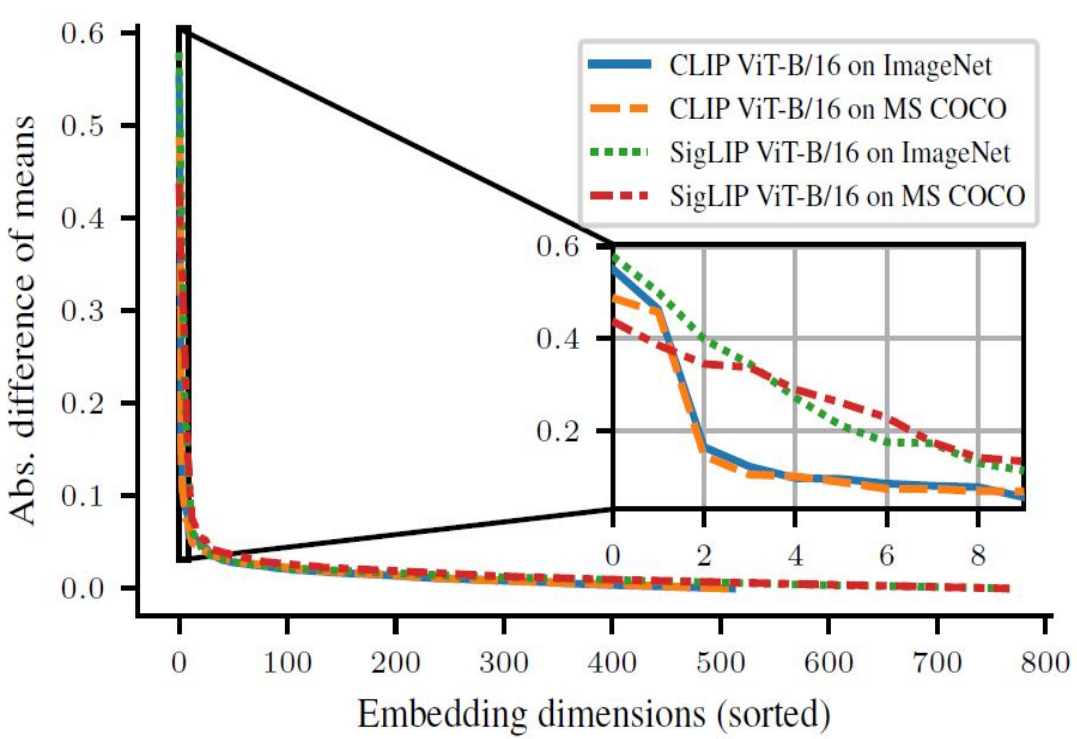
slide credit: Thomas Brox

- Vision-language models trained with contrastive loss (e.g. CLIP) separate image and text in the embedding space (Liang et al. 2022)
- How does this happen?
- How critical is it?
- What causes the gap?

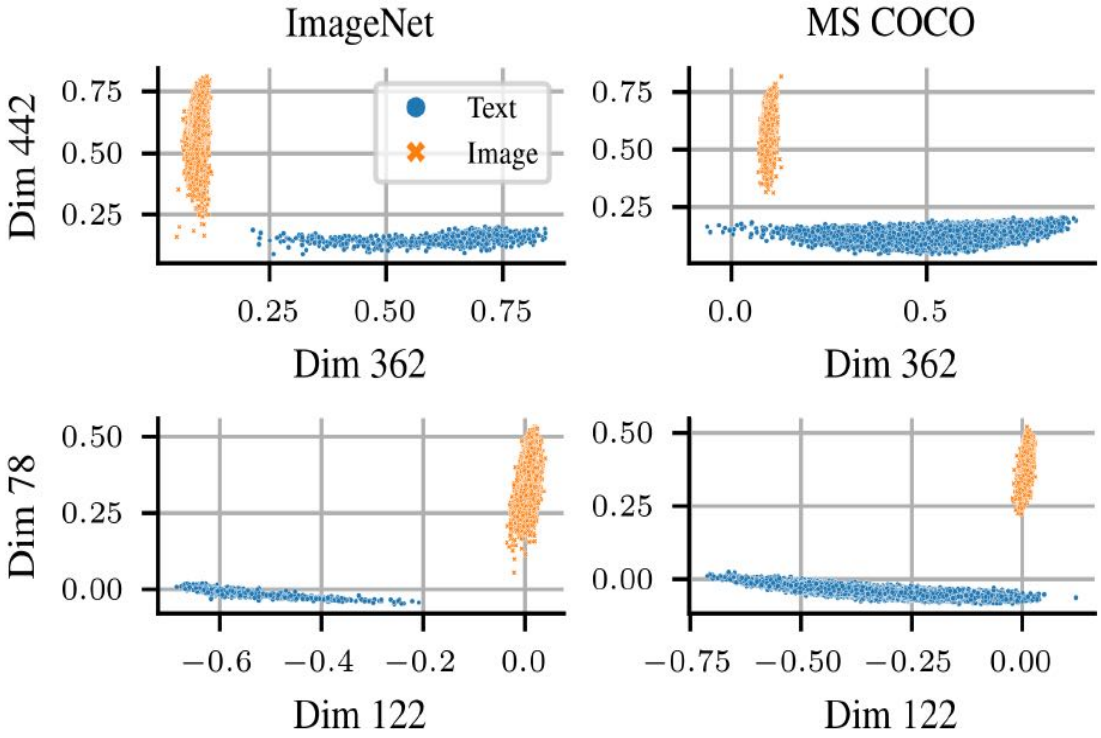


# Modality gap due to few (mostly two) dimensions

slide credit: Thomas Brox



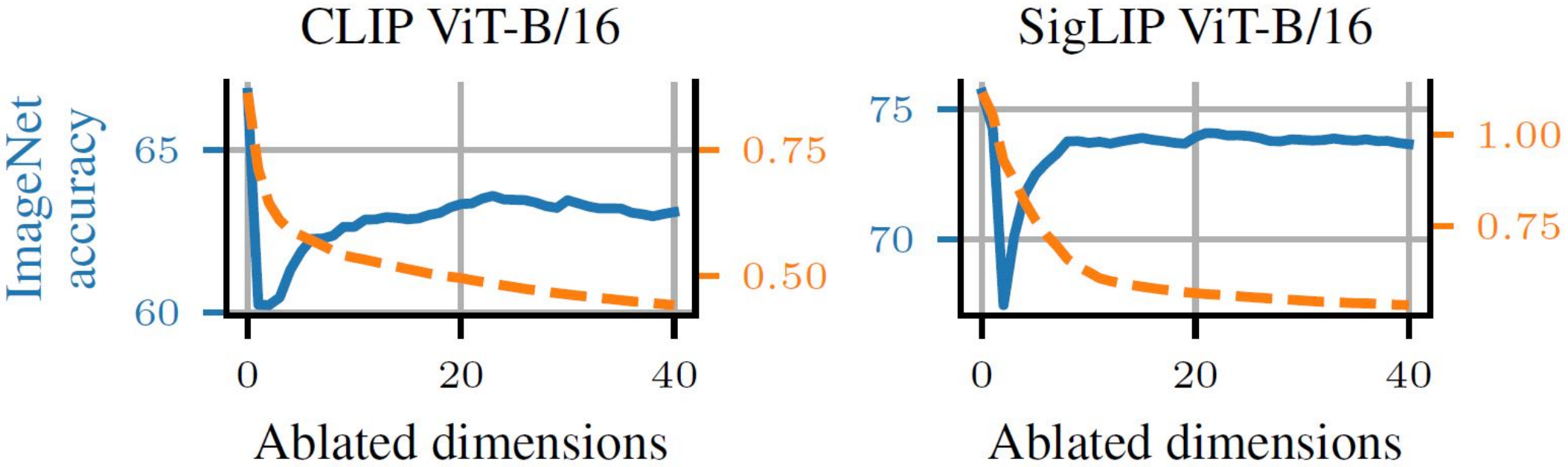

(a) Abs. difference of means of embedding dims.




(b) Some dims. perfectly separate the modalities.

# Can we just remove these dimensions?

slide credit: Thomas Brox

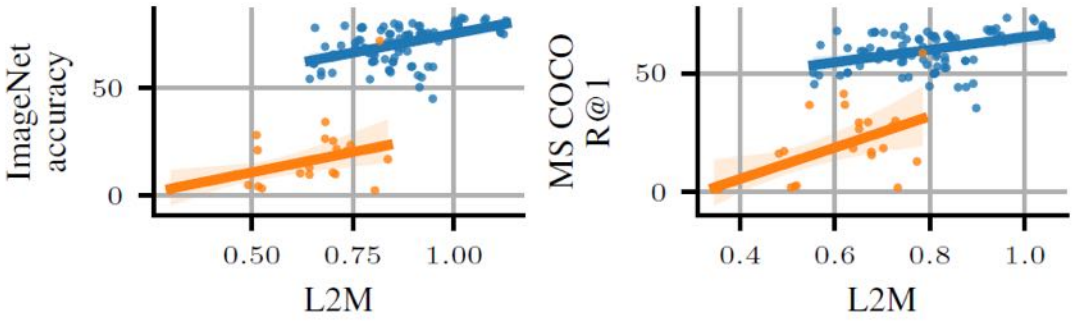



→ These dimensions are most important

# Does the modality gap decrease performance?

slide credit: Thomas Brox

- Meta-study using more than 100 publiclv available models



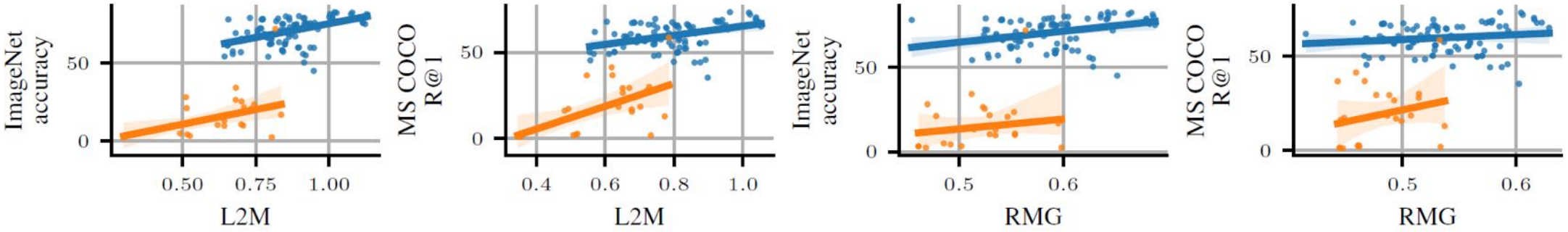

→ The modality gap correlates positively with performance

- Is it the metric?

# Model size and data

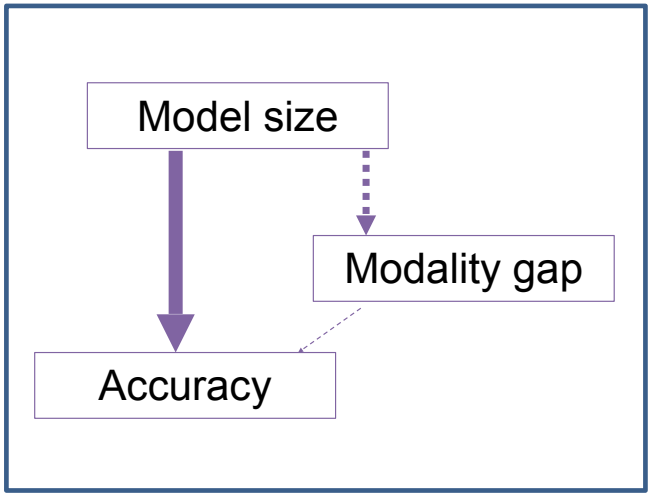
slide credit: Thomas Brox

- Model size correlates strongly with accuracy




- Model size also correlates with the modality gap and is a strong confounder (as is the dataset)

- Confounders are hard to control
  → Causal relationship stays unclear




Causal graph

# What causes the modality gap?

slide credit: Thomas Brox

- InfoNCE loss:

$$\mathcal{L}(f_x, f_y) := \mathop{\mathbb{E}}_{(\mathbf{x}_i, \mathbf{y}_i) \sim p_{\text{data}}} \left[ -\log \frac{\exp(f_x(\mathbf{x}_i)^T f_y(\mathbf{y}_i)/\tau)}{\exp(f_x(\mathbf{x}_i)^T f_y(\mathbf{y}_i)/\tau) + \sum_{j=1}^{N-1} \exp(f_x(\mathbf{x}_i)^T f_y(\mathbf{y}_j)/\tau)} \right]$$

- **Information imbalance**: many concepts visible in the image are not mentioned in the text

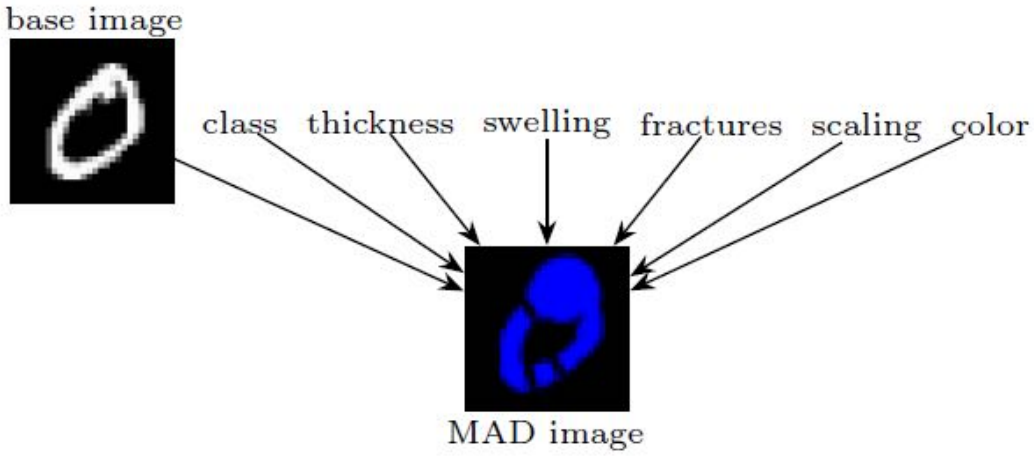
  → image and text often cannot be well aligned

- Consequence: model minimizes the denominator by maximizing the distance between images and text

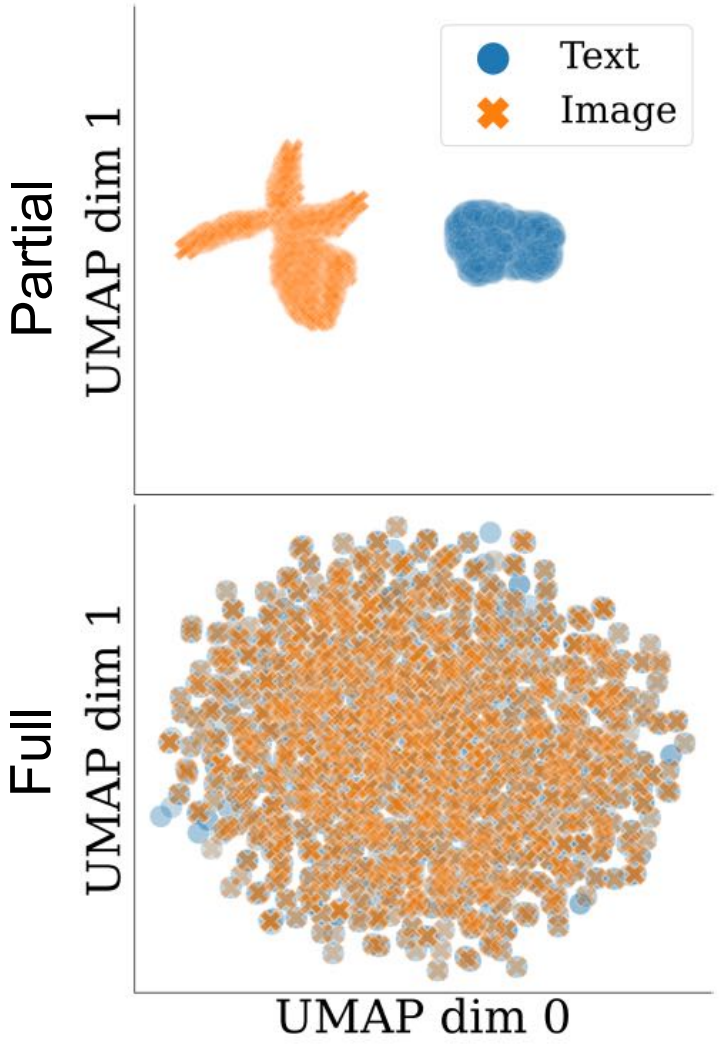
# Partially vs. fully-informed image-text pairs

slide credit: Thomas Brox

- Synthetic dataset to control information completeness




- With fully informed data pairs, the modality gap goes away



# Overview of Today's Lecture

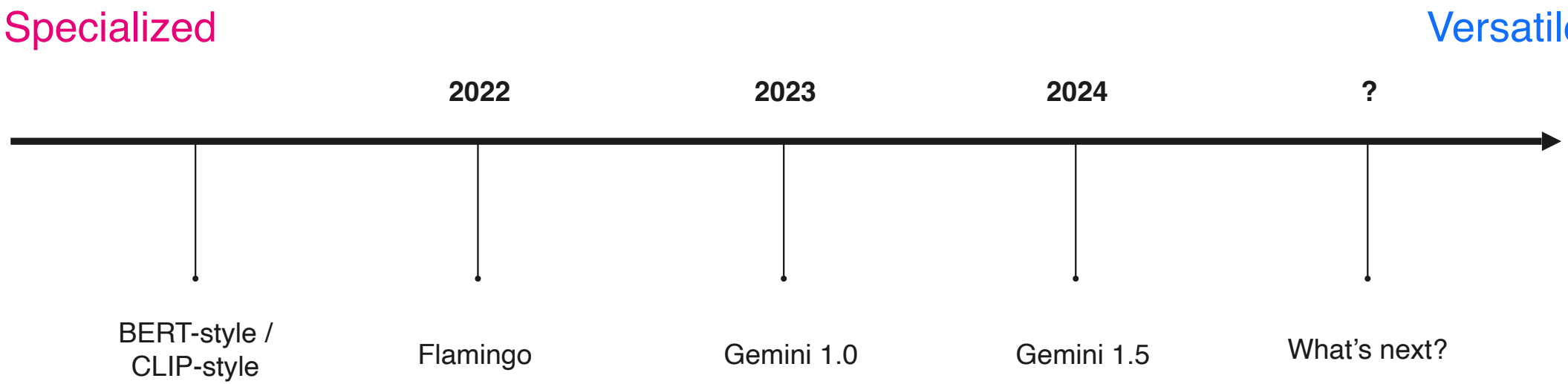
- Vision Language Learning for Computer Vision
  - Supervised learning vs. vision-language learning
  - CLIP [icml'21] - https://arxiv.org/abs/2103.00020
  - ALIGN [icml'21] - https://arxiv.org/abs/2102.05918
  - Modality gap discussion — https://arxiv.org/abs/2404.07983 (Apr'24)

- Large Vision Language Models — Leveraging Large Language Models
  - Flamingo [neurips'22] — https://arxiv.org/abs/2204.14198
  - Gemini 1.0 — https://arxiv.org/abs/2312.11805 (Dec'23 & Jun'24)
  - Gemini 1.5 — https://arxiv.org/abs/2403.05530 (Mar'24 & Jun'24)

# Overview of story…

slide credit: Antoine Miech

Google DeepMind



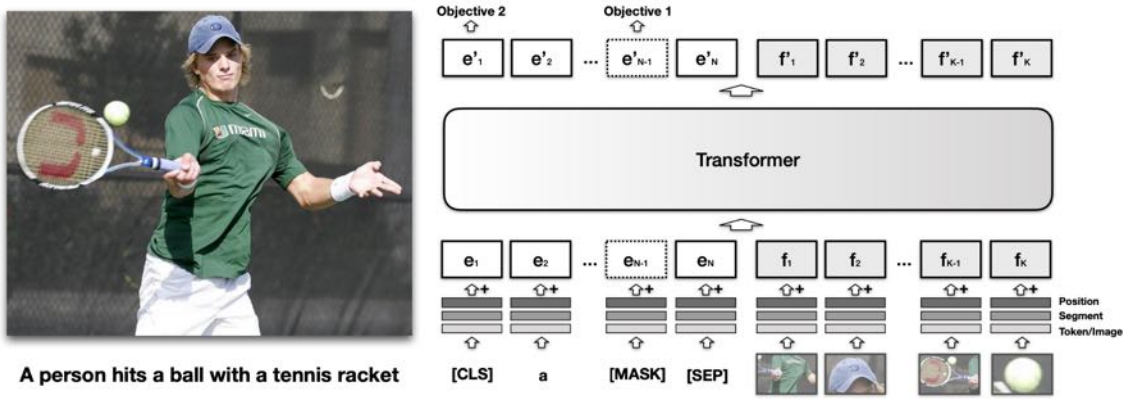
slide credit: Antoine Miech

Google DeepMind

# BERT-Like text-vision models

- **Encoder only model**
- Trained with a masked loss, similarly as BERT, typically fine-tuned for a downstream task

Single Stream (Visual BERT)




- VisualBert (ss)
- ViLBERT (ds)
- VL-BERT
- UNITER
- OSCAR
- VideoBERT
- ActBERT
- Unicoder-VL
- MERLOT
- HERO
- ALBEF
- **Many more…**

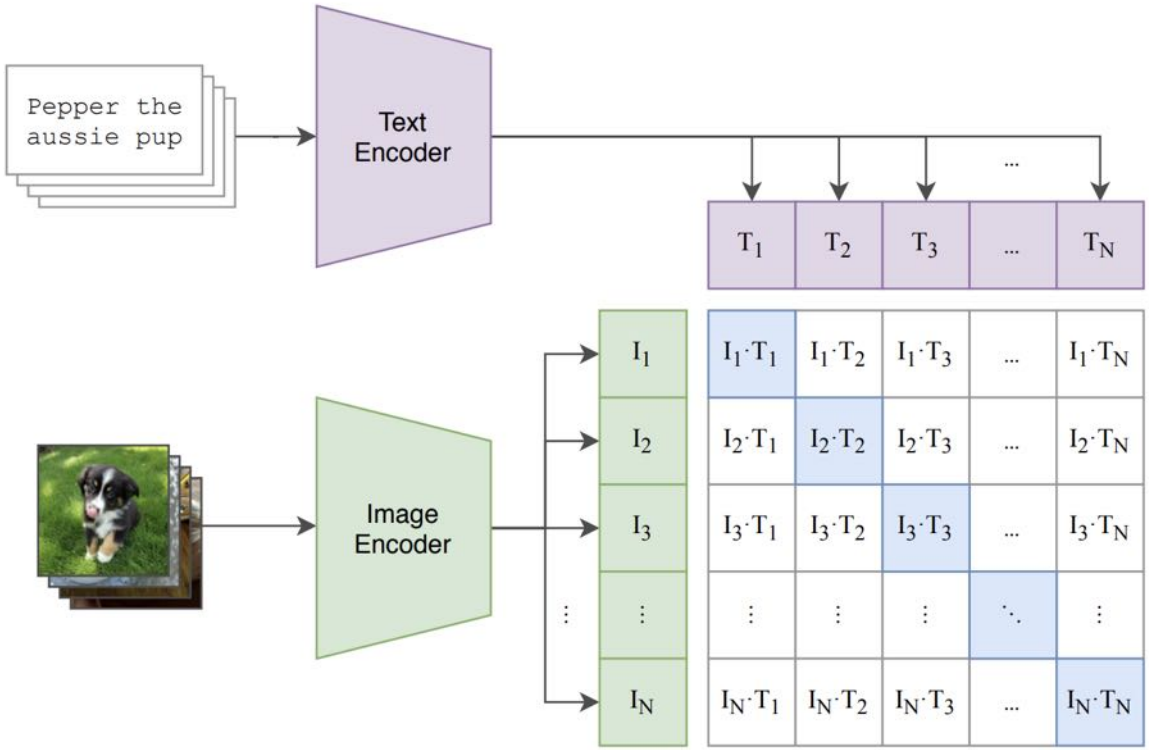
Li et al. "VisualBERT: A Simple and Performant Baseline for Vision and Language." arXiv preprint:1908.03557 (2019).

slide credit: Antoine Miech

Google DeepMind

# Text-Vision contrastive models

Efficient for training image classification and image retrieval models




Contrastive loss matrix

- CLIP
- ALIGN
- CoCa
- Florence
- BASIC
- LiT
- FILIP

Alec Radford, et al. "Learning Transferable Visual Models From Natural Language Supervision" arXiv preprint arXiv:2103.00020 (2021)

# Autoregressive Visual Language Models

slide credit: Antoine Miech

Google DeepMind

*What do these creatures usually do?*

Autoregressive VLM

**text**

*They usually chill on your sofa.*

A generic interface to solve **many** visual + language problems.

- SimVLM
- Virtex
- MAGMA
- Frozen
- VisualGPT
- GIT
- ClipClap
- VC-GPT
- CM3
- BLIP
- Uni-Perceiver
- VL-BART
- VL-T5
- VLM
- PaLI, PaLI-X
- Flamingo
- GPT4-V
- Gemini
- Claude 3
- And many more...

slide credit: Antoine Miech

Google DeepMind

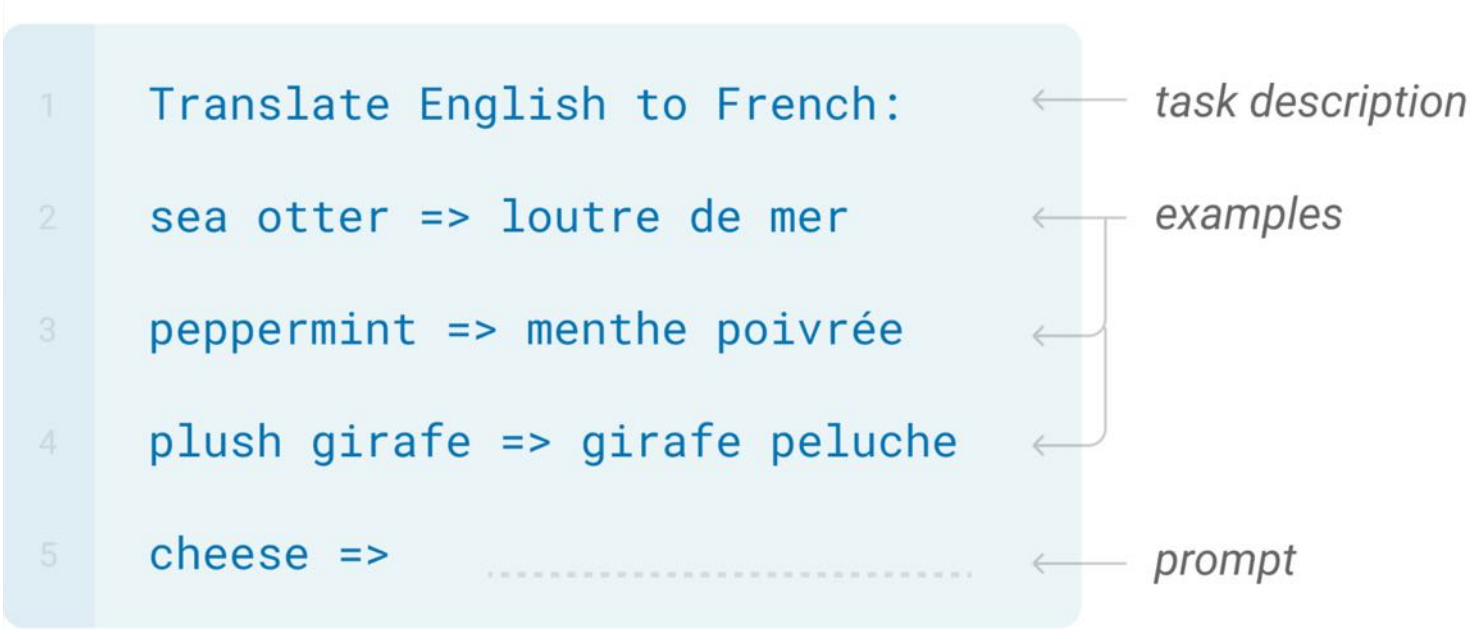
# 🦩 Flamingo: A Visual Language Model for Few-Shot Learning

**J**ean-Baptiste Alayrac*, Jeff Donahue*, Pauline Luc*, Antoine Miech*, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, Karen Simonyan*, NeurIPS 2022

*: Equal contributions

# The motivation: In-context learning



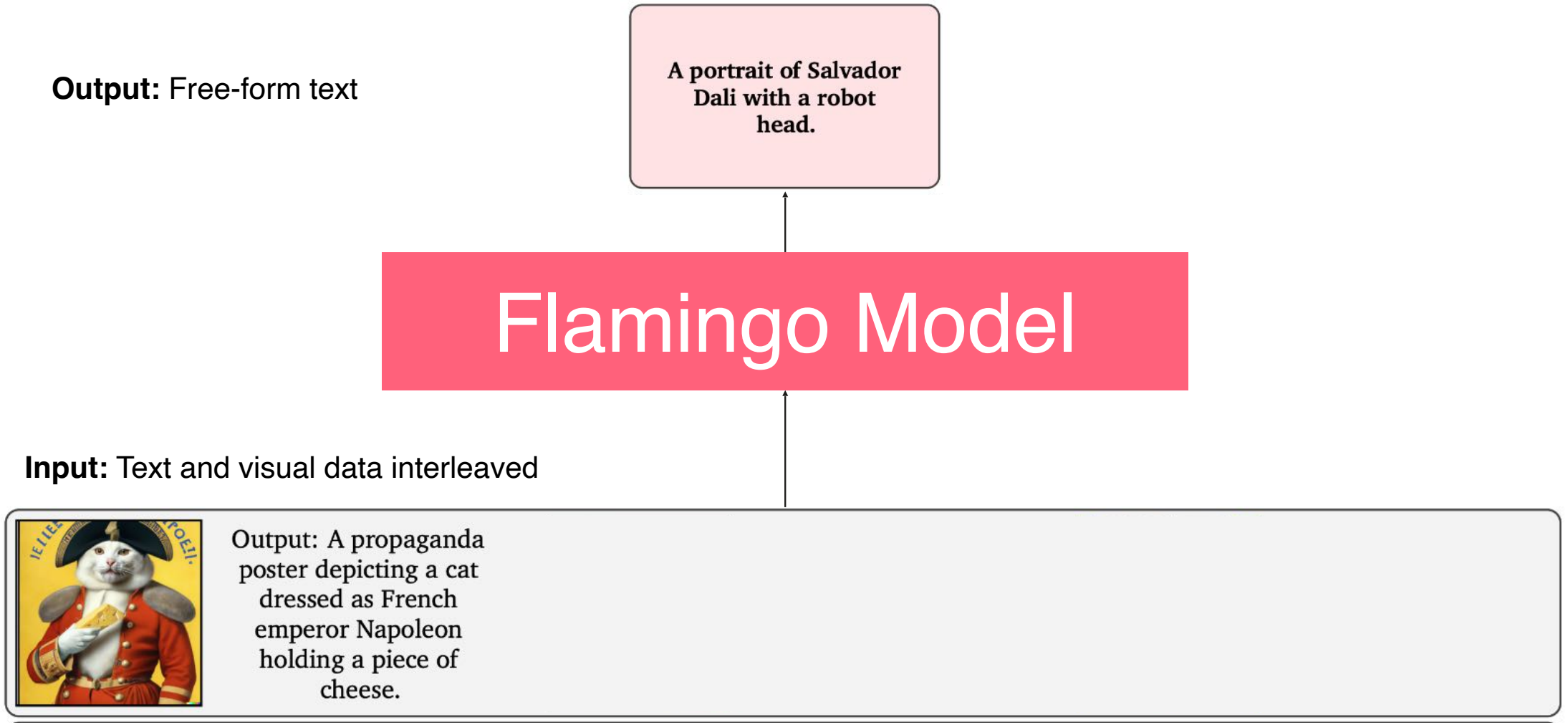



*Language Models are Few-Shot Learners*, Brown et al., NeurIPS 2020

# What is Flamingo?

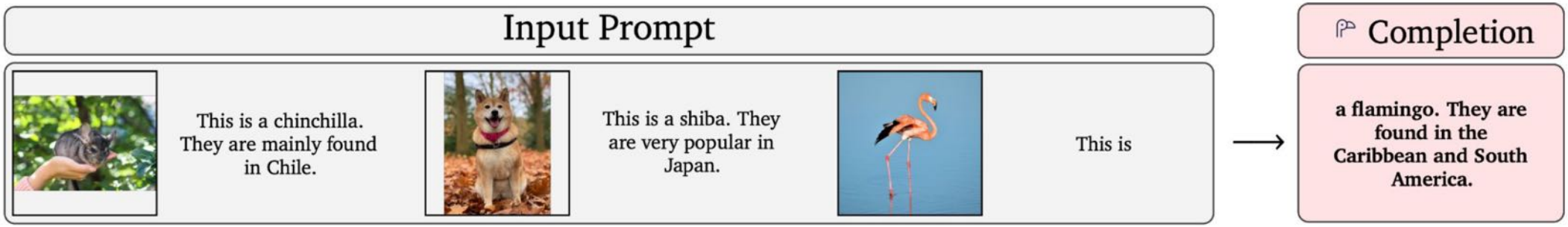
Google DeepMind

**Output:** Free-form text



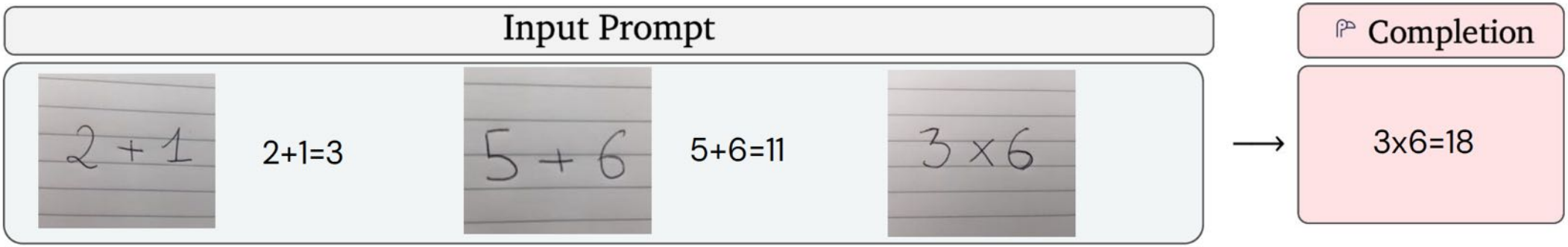
slide credit: Jean-Babtiste Alayrac

# Flamingo comes in many colours depending on what it eats



slide credit: Jean-Babtiste Alayrac

# Flamingo comes in many colours depending on what it eats



slide credit: Jean-Babtiste Alayrac

# Flamingo comes in many colours depending on what it eats

Input Prompt

What happens to the man after hitting the ball? Answer:

→

Completion

he falls down.

slide credit: Jean-Babtiste Alayrac

# Why did we build it?

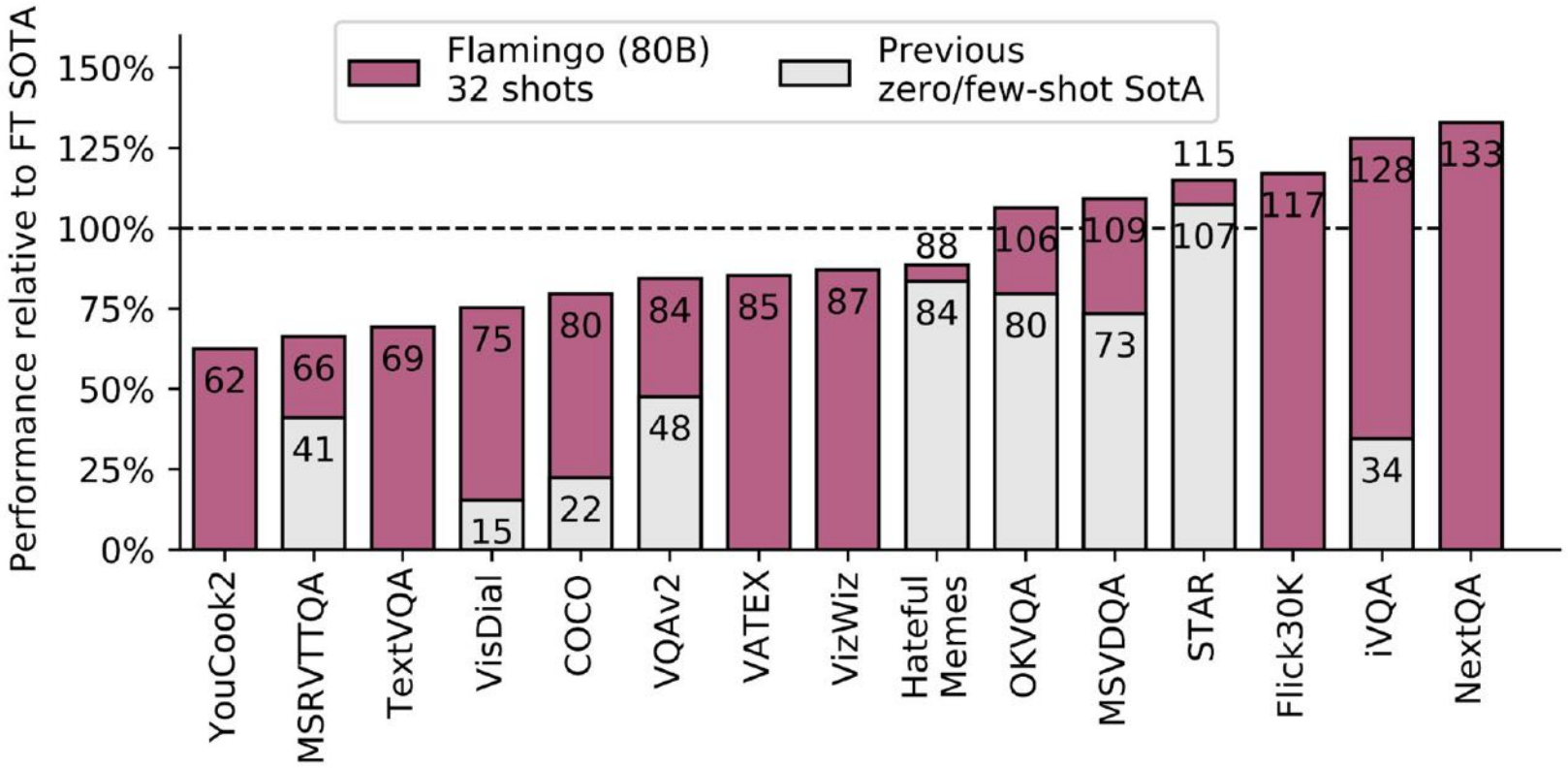
Build a state-of-the-art, generalist Visual Language Model that can **be rapidly adapted to different multimodal tasks via few-shot learning**

- **Visual Language Model**: ingest visual data (images or videos) along with a language input, and produce language output.

- **Generalist … rapidly adapted to different multimodal tasks:** one model can address multiple tasks (captioning, visual dialogue, classification) with the same weights and without any post-hoc training.

- **Few-shot learning:** condition the model to solve various tasks with only a few input-output examples (32 examples are used)

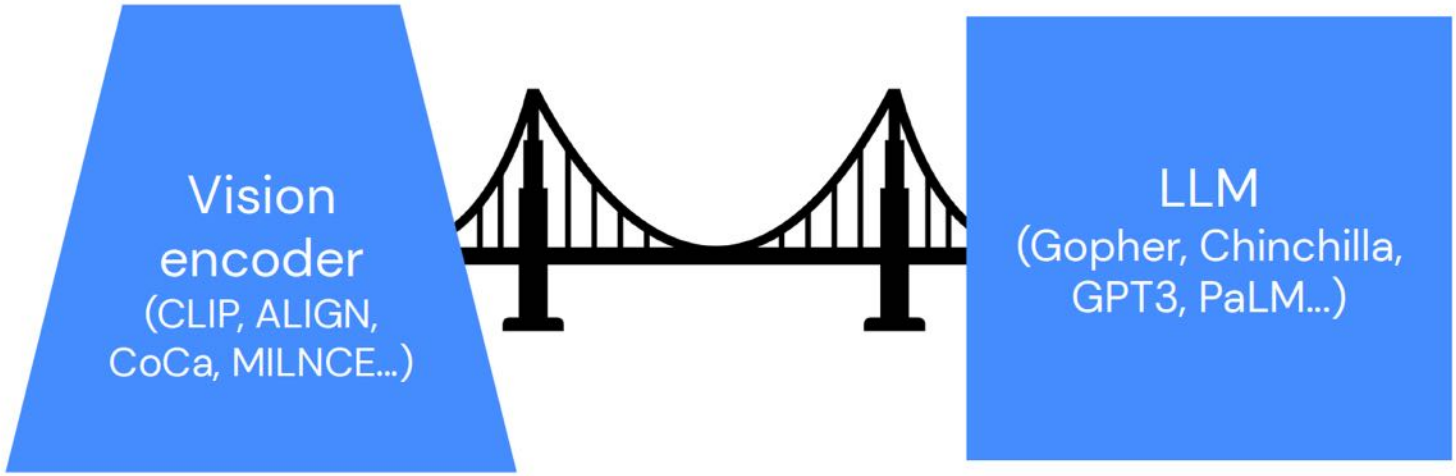
# Main results: Few-shot results



slide credit: Jean-Babtiste Alayrac

# How does it work?

**Model overview**
Pretrained parts of the model are frozen:
the Vision Encoder and the LLM.




👀

**The perception**

🧠

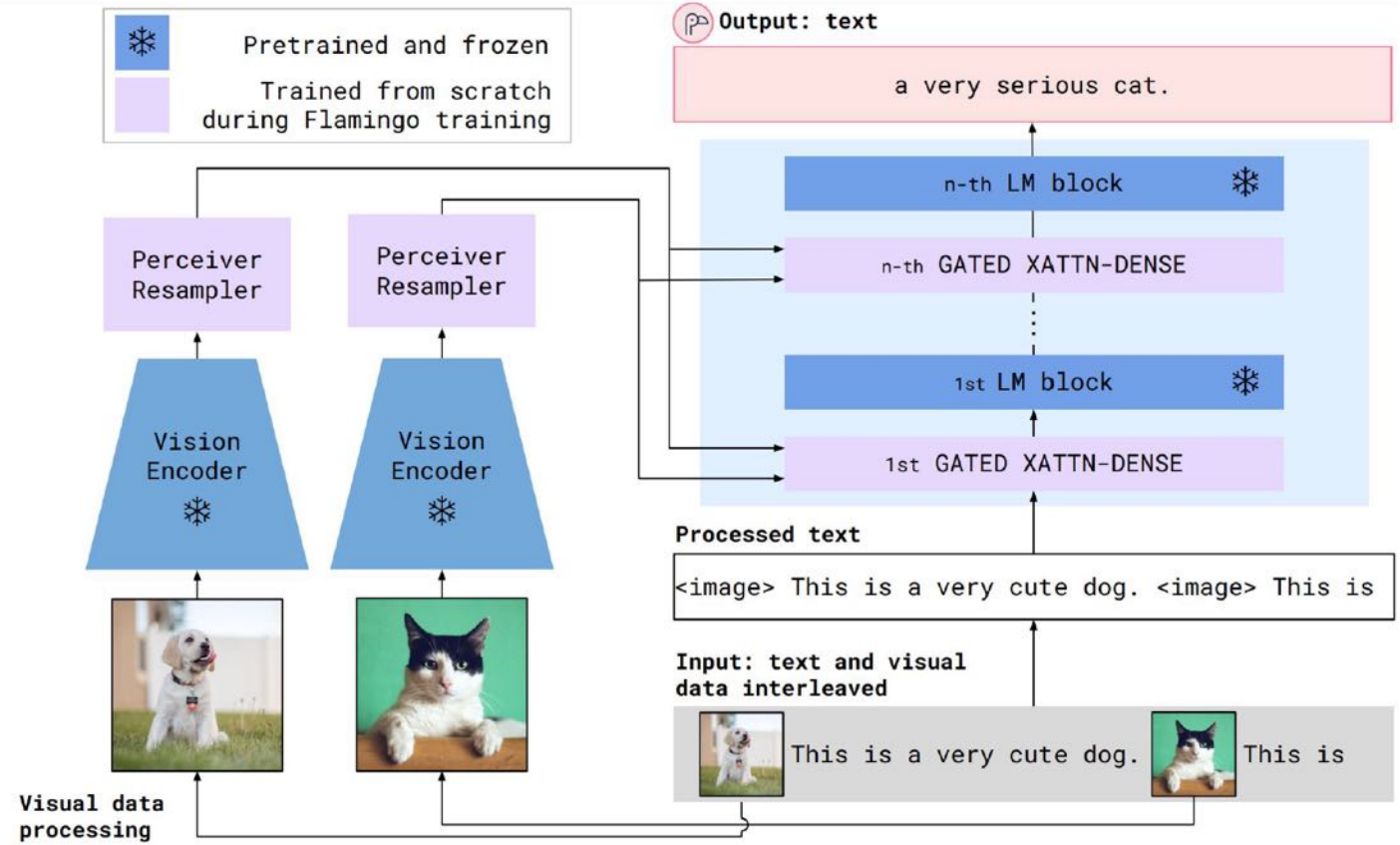
**The "reasoning part" and "knowledge source"**

slide credit: Jean-Babtiste Alayrac

# How does it work?

## Model overview

Pretrained parts of the model are frozen:
the Vision Encoder and the LLM.




## Training datasets


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

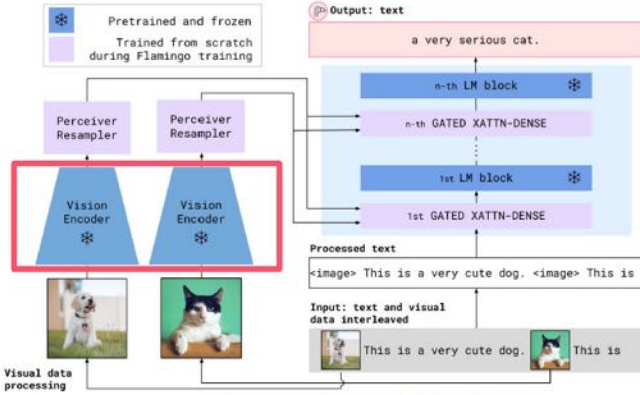

Video-Text Pairs dataset
[N=1, T>1, H, W, C]


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

# Visual processing

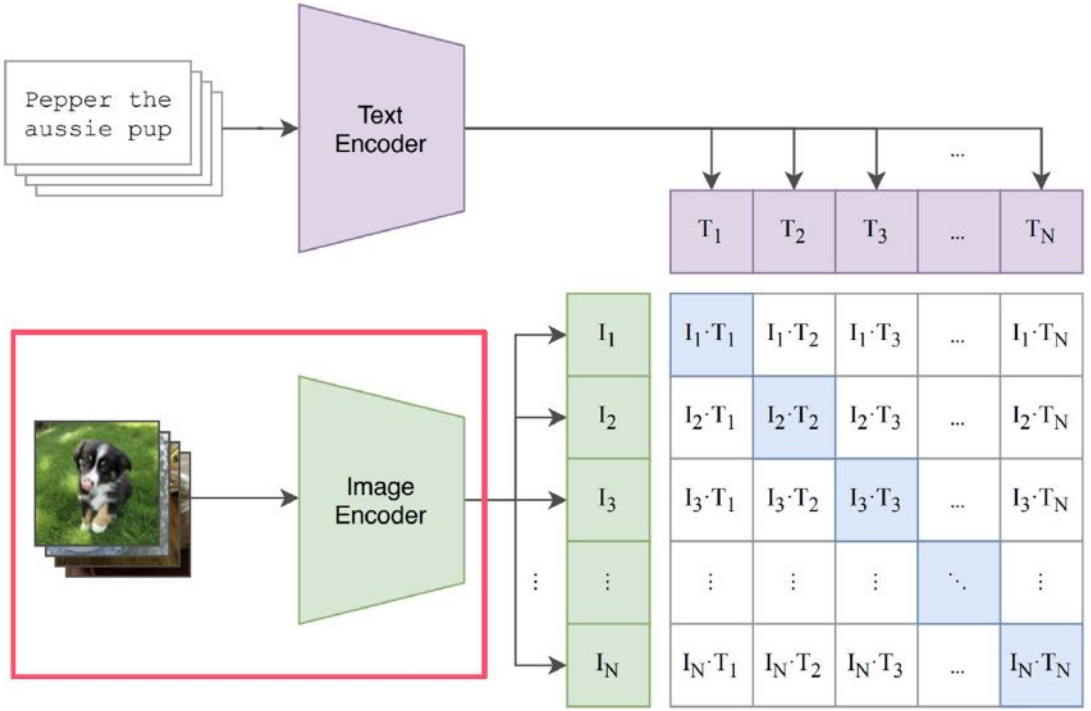
## Vision Encoder:

Pretrained with image-text contrastive training (CLIP-like) and kept frozen during Flamingo training.




Overall architecture

We only keep the vision encoder and discard the text encoder.

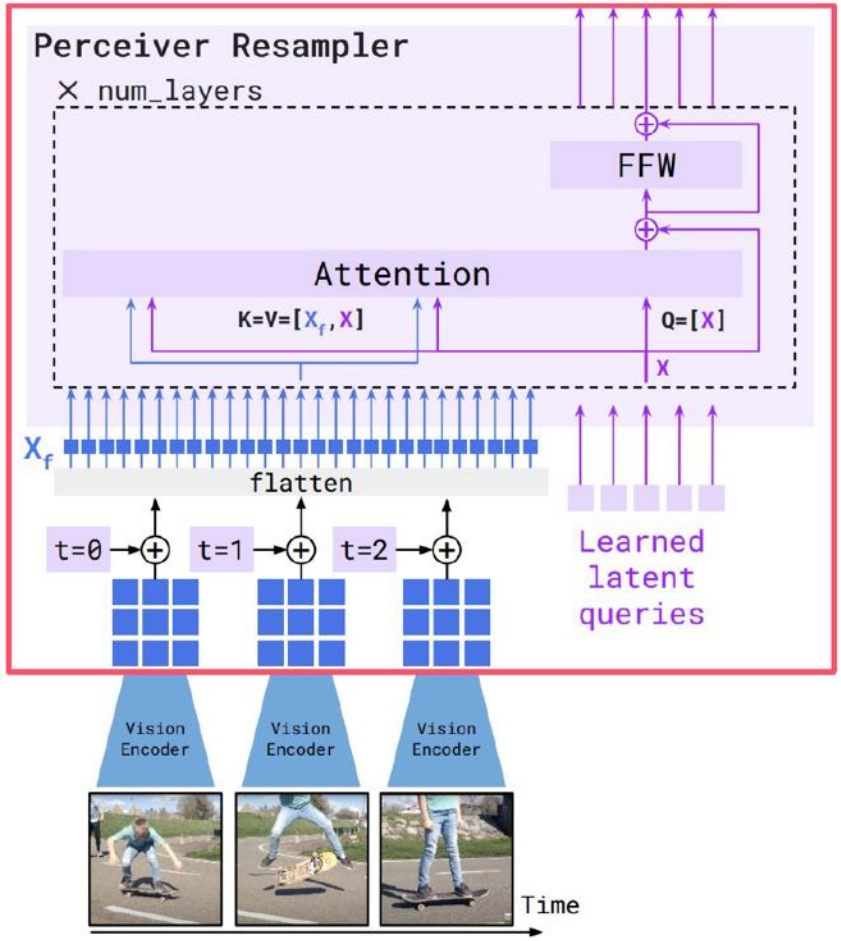


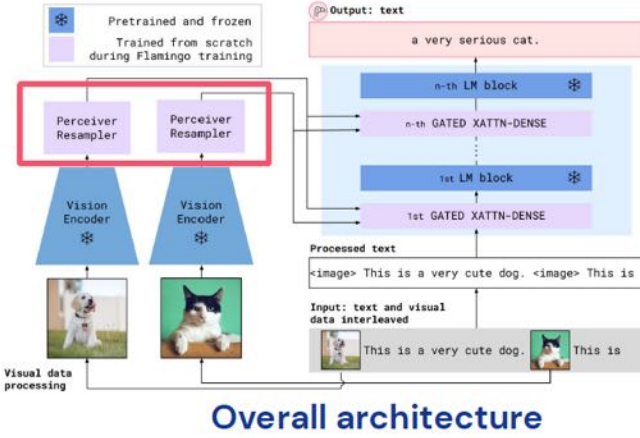

slide credit: Jean-Babtiste Alayrac

# Visual processing

**Perceiver Resampler:**

Takes as *input a variable number of features* (image or videos) and *outputs a fixed number* of "visual tokens".

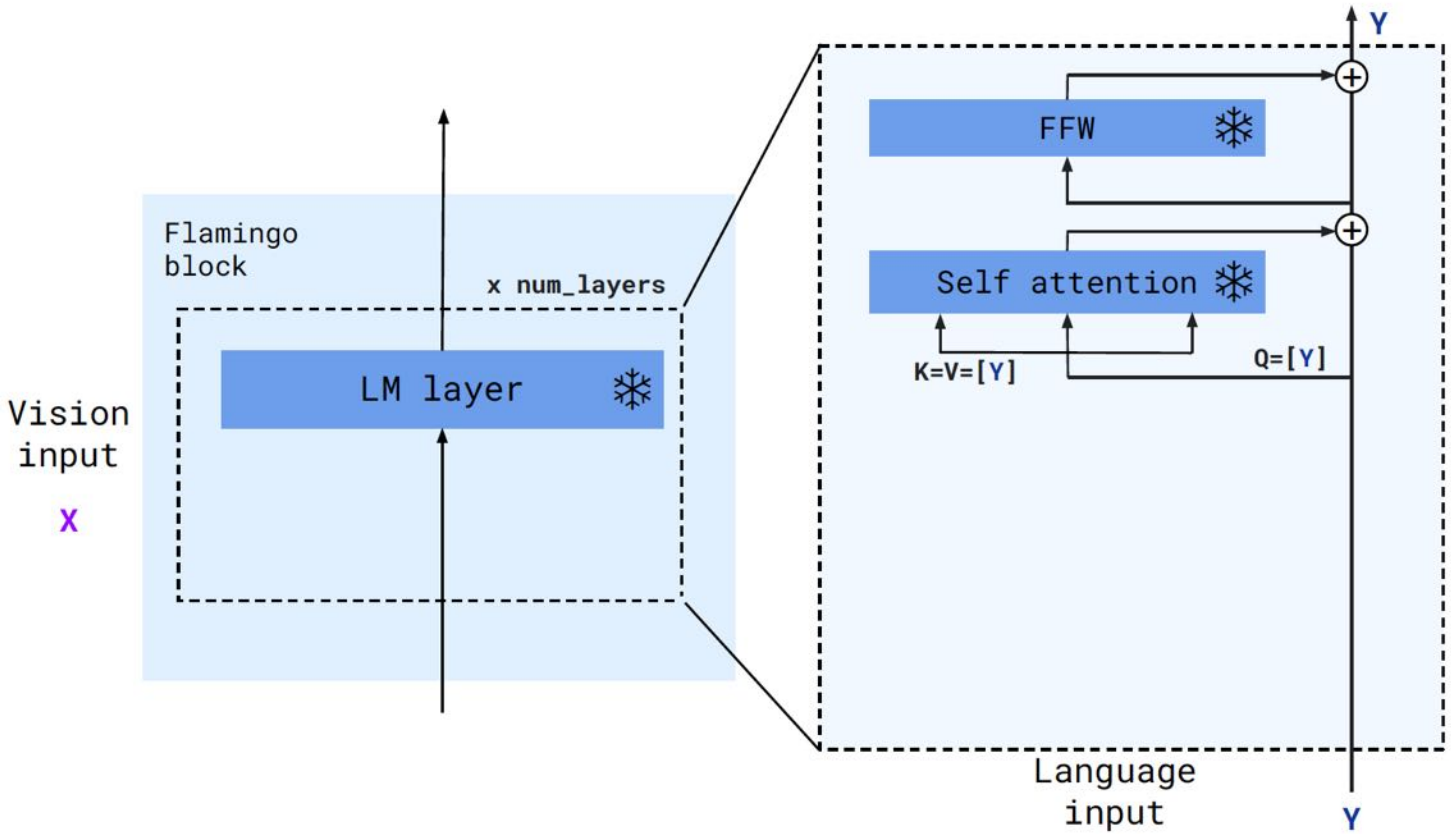


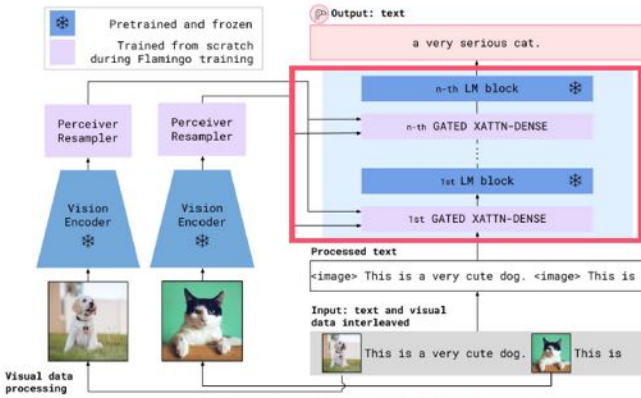




Overall architecture

slide credit: Jean-Babtiste Alayrac

# Leveraging an existing language model

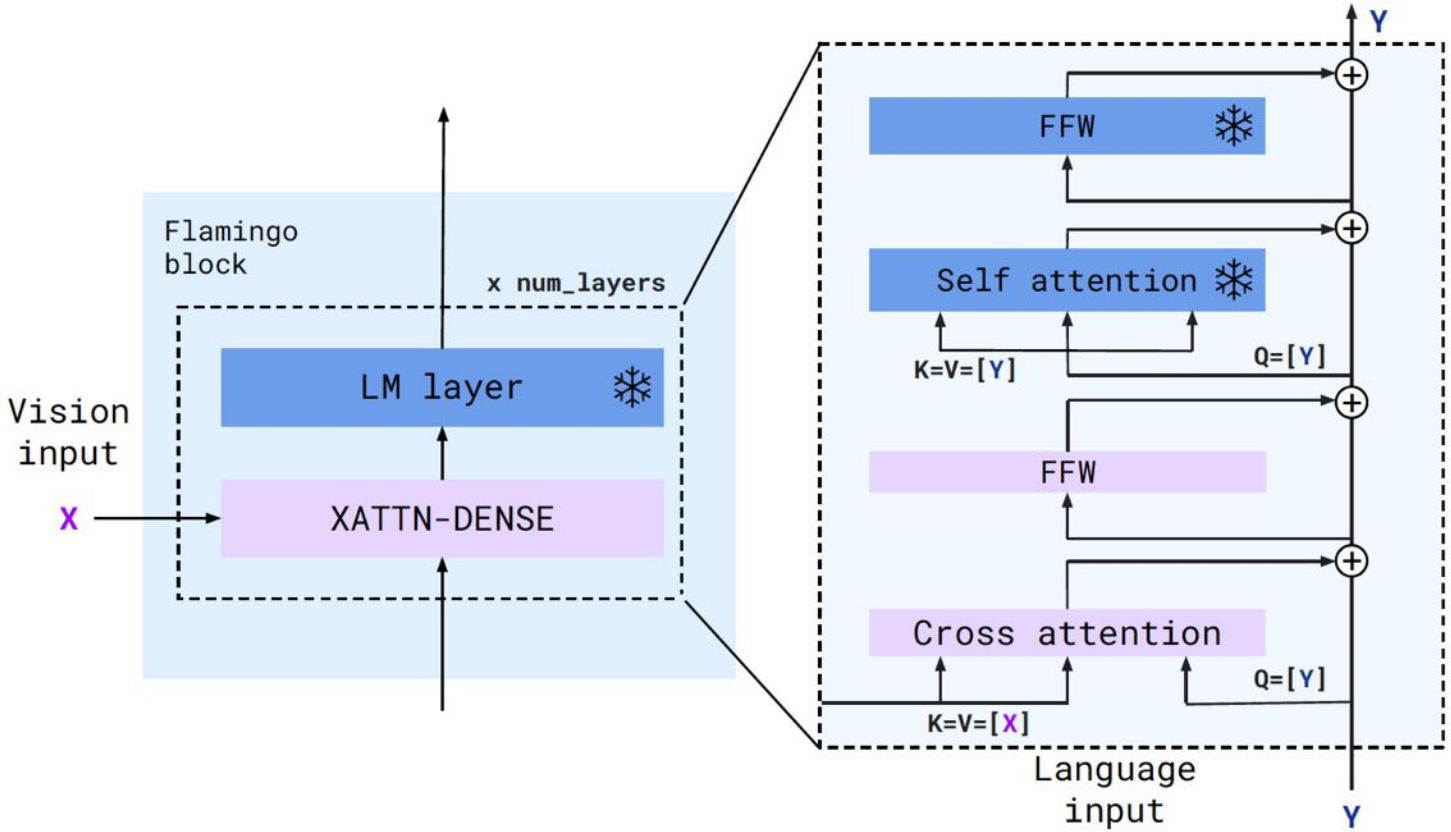


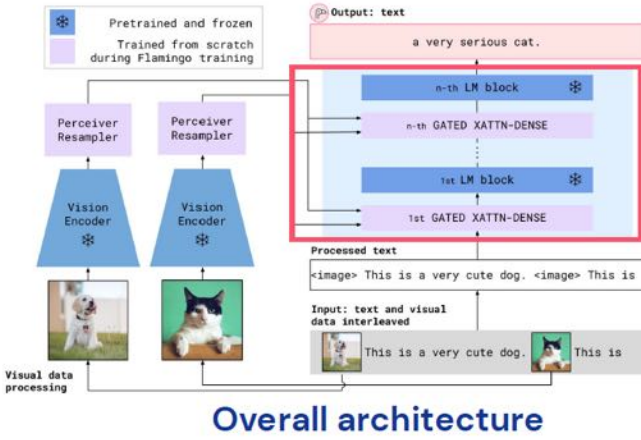




Overall architecture

slide credit: Jean-Babtiste Alayrac

# Leveraging an existing language model

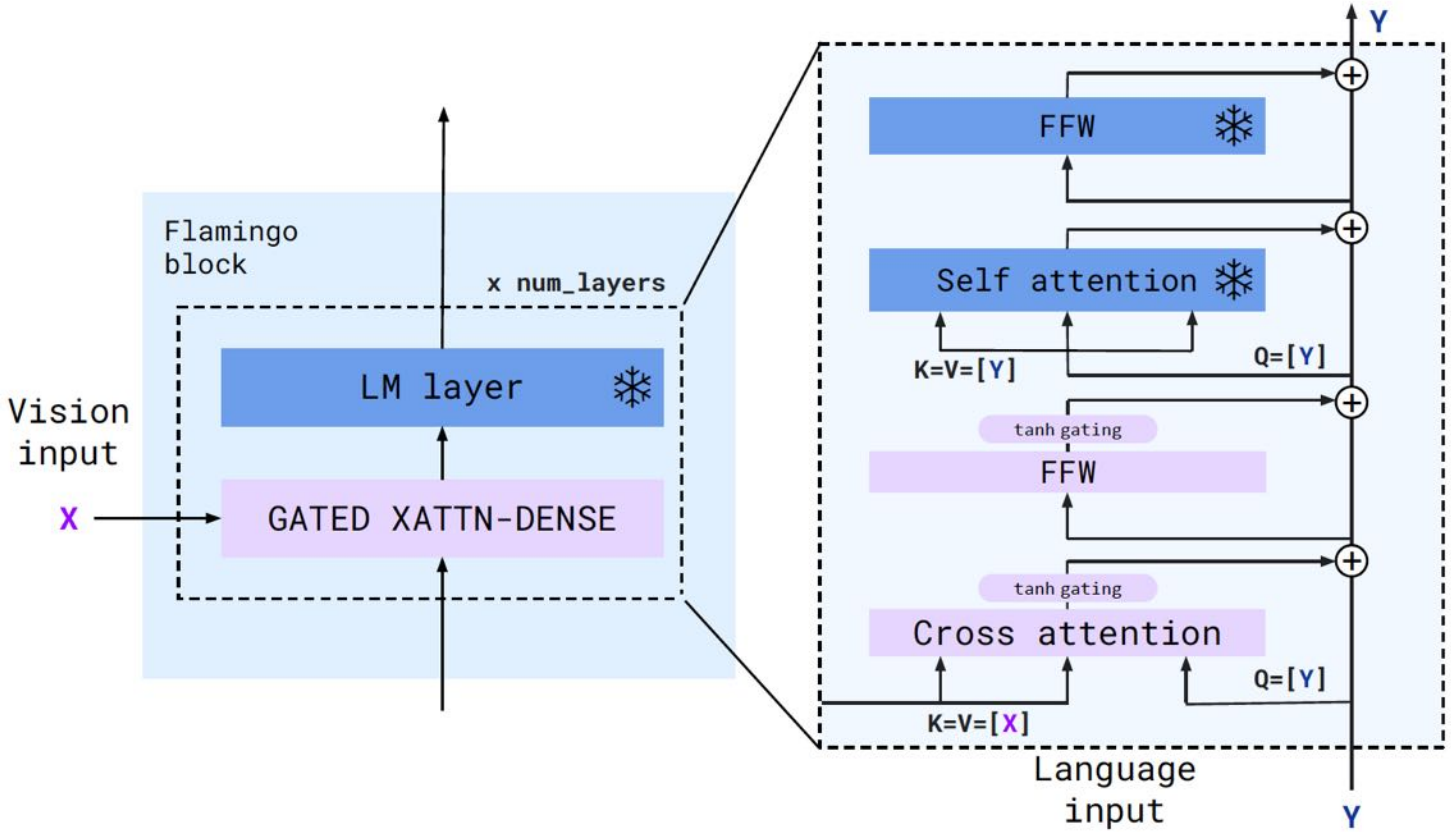


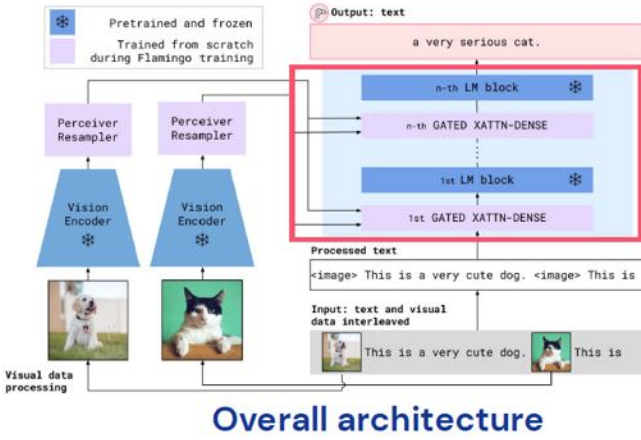




Overall architecture

slide credit: Jean-Babtiste Alayrac

# Leveraging an existing language model

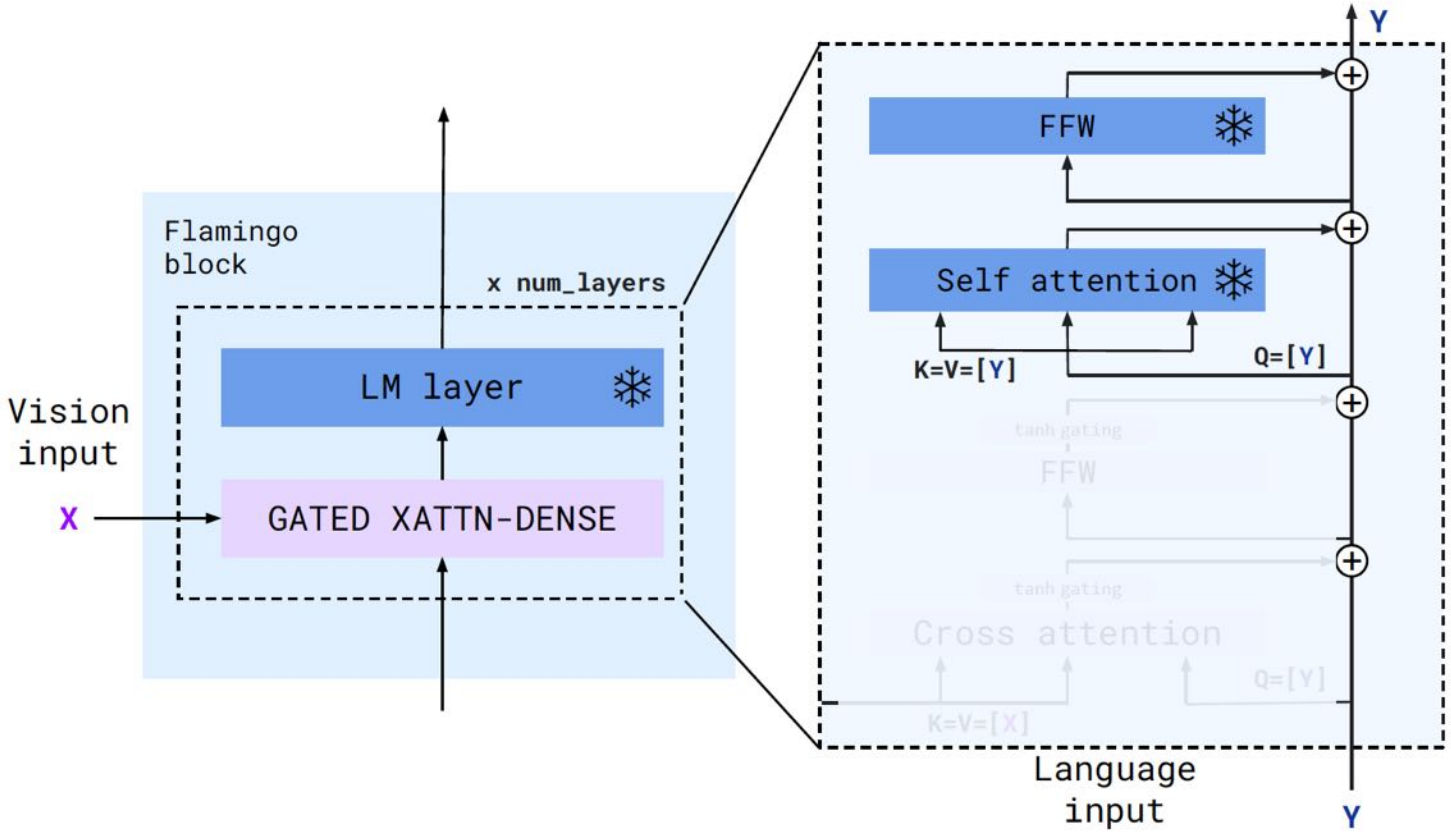


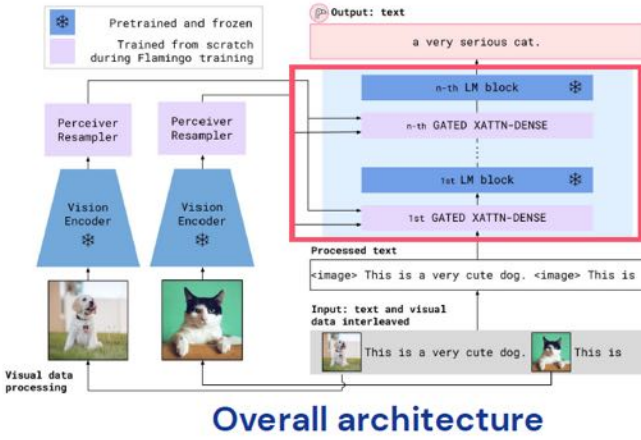




Overall architecture

slide credit: Jean-Babtiste Alayrac

# Leveraging an existing language model



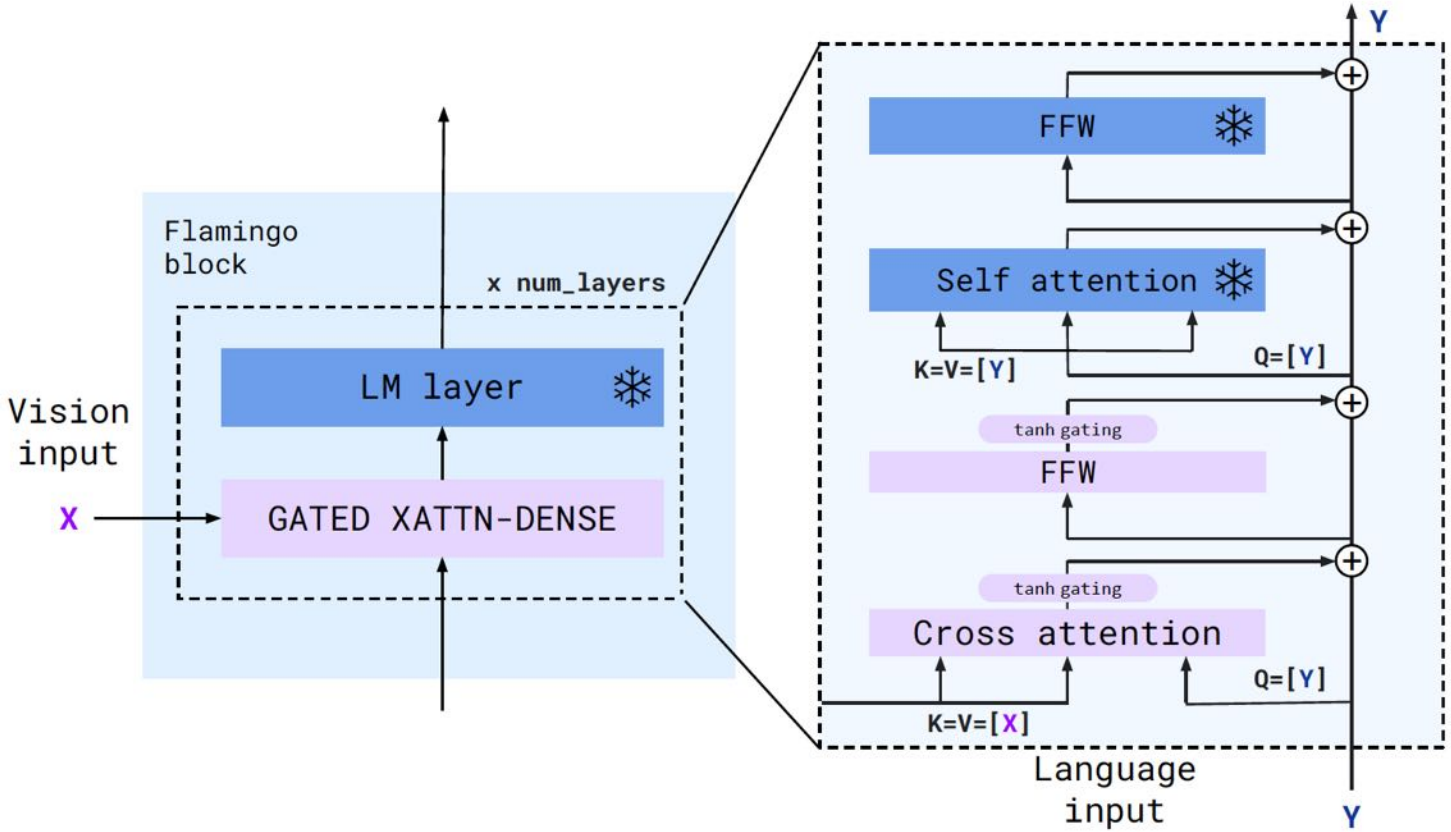



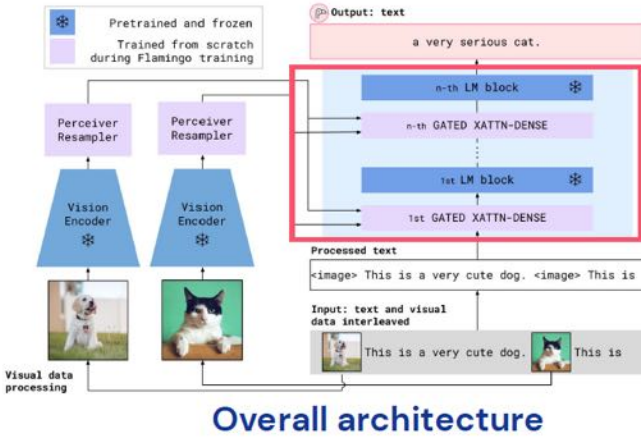

Overall architecture

At initialisation, tanh gates are all 0.

slide credit: Jean-Babtiste Alayrac

# Leveraging an existing language model




Overall architecture

At initialisation, tanh gates are all 0.

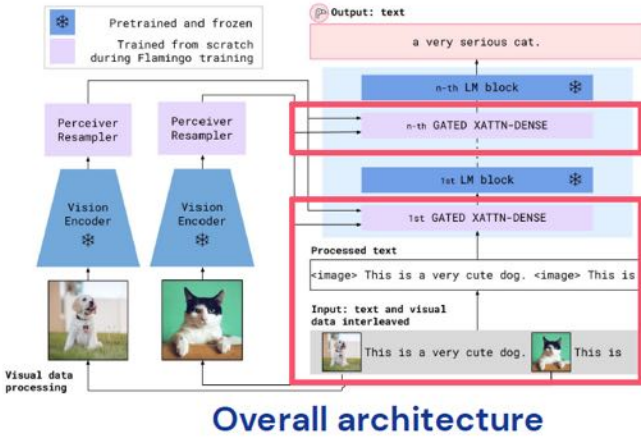
They slowly open as training progresses.

slide credit: Jean-Babtiste Alayrac

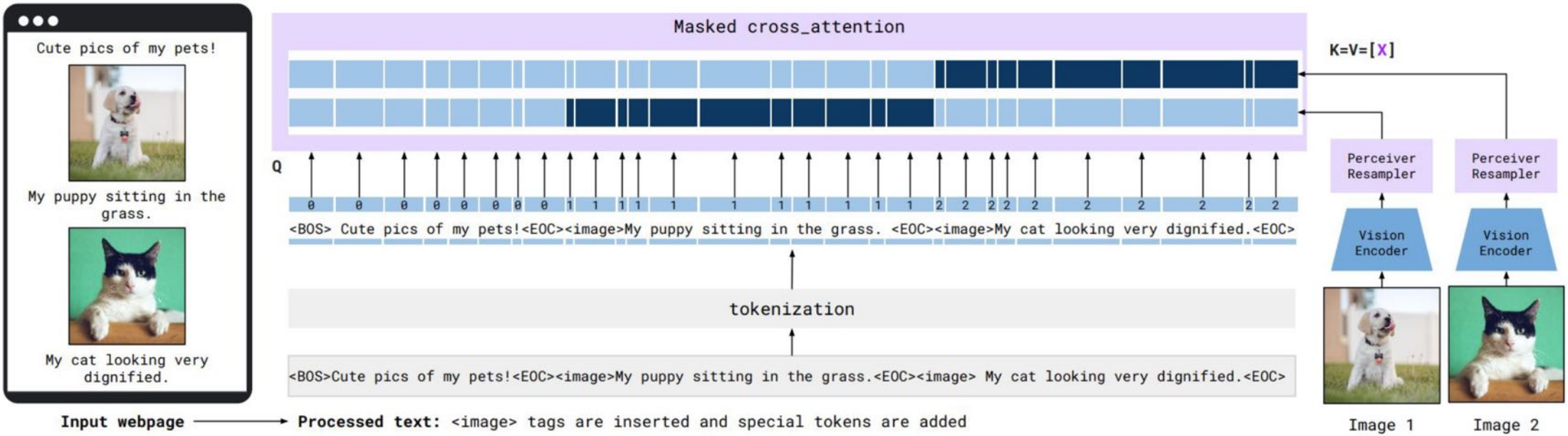
# Deal with interleaved visual and text sequence

Each text token cross-attend to the image that precedes it in the interleaved sequence.

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




Overall architecture

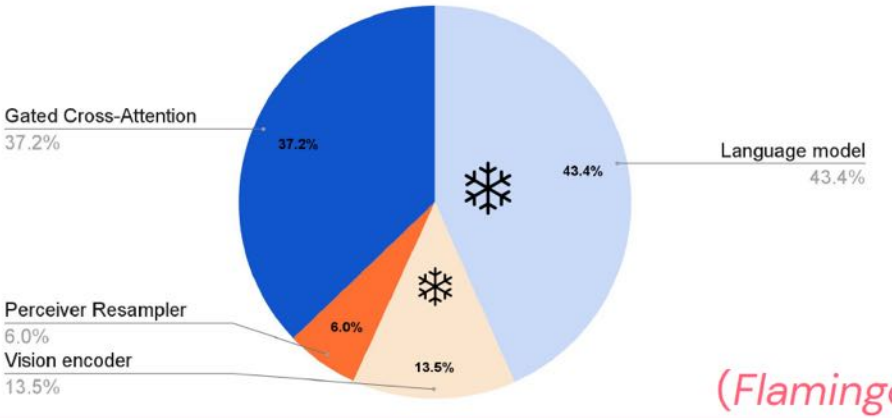



slide credit: Jean-Babtiste Alayrac

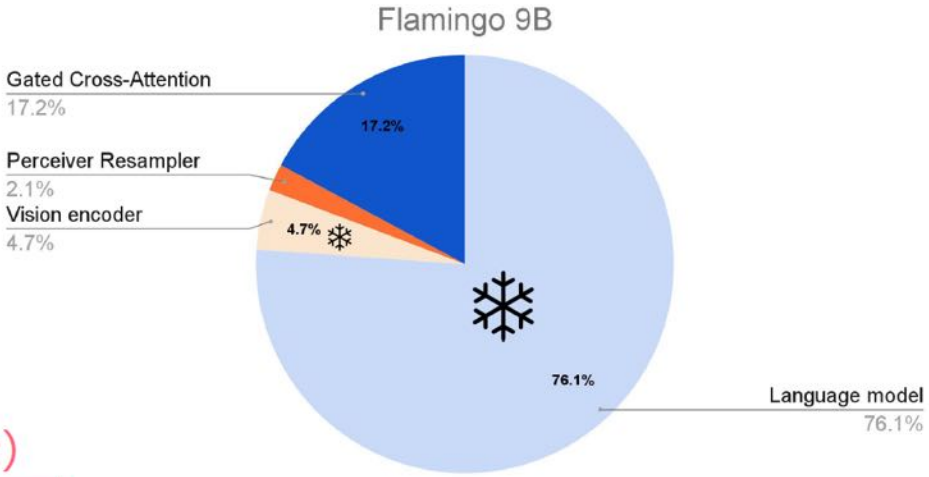
slide credit: Jean-Babtiste Alayrac

# The Flamingo family

## Flamingo 3B

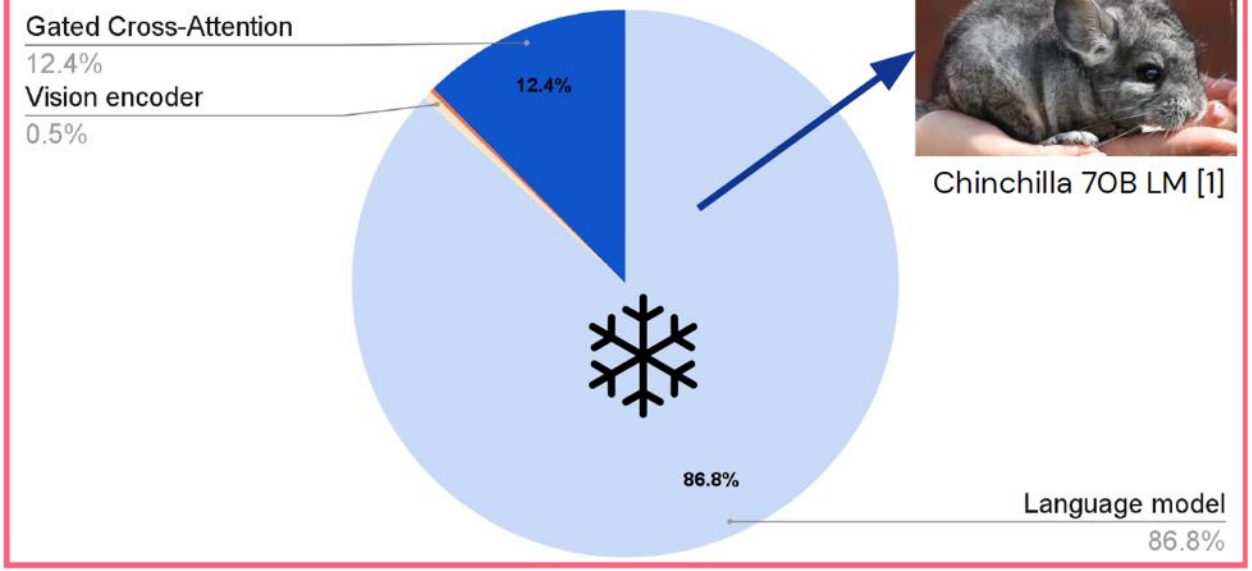



## Flamingo 9B




(*Flamingo* in short)

## Flamingo 80B






Chinchilla 70B LM [1]

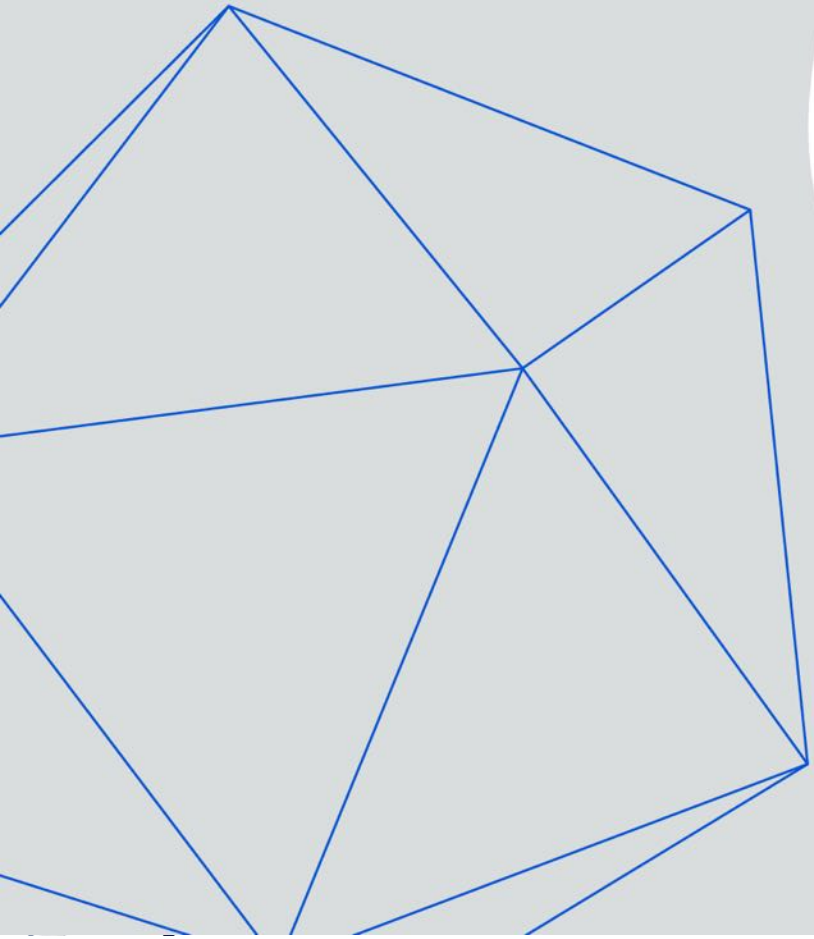
- Vision encoder (NFNet-F6) size fixed.
- Resampler size fixed.
- Focus on scaling the frozen language model.

[1] Hoffmann et al., Training Compute-Optimal Large Language Models, 2022

slide credit: Jean-Babtiste Alayrac

DeepMind

# Training data



Flamingo training data

slide credit: Jean-Babtiste Alayrac

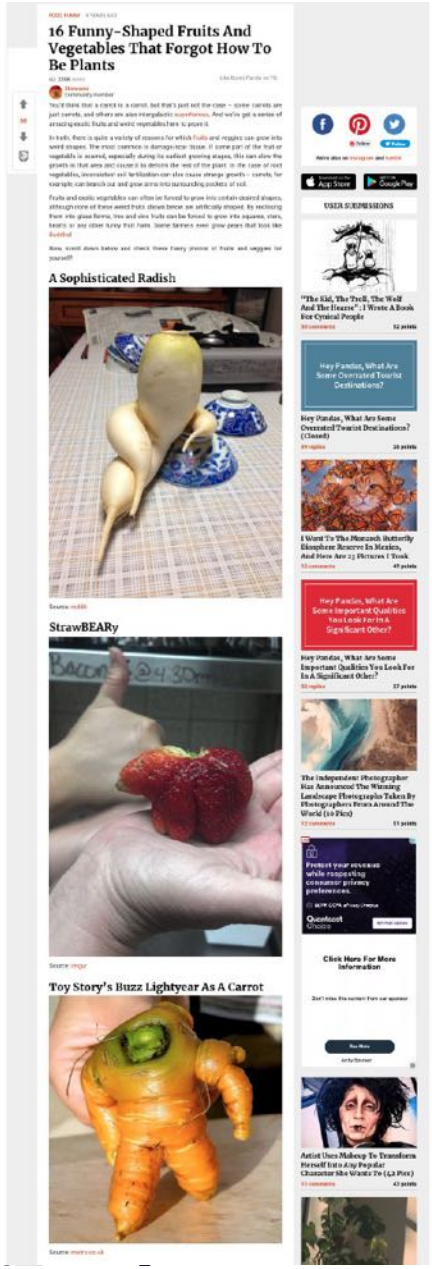
# Image / Video paired with Captions

An English bulldog standing on a skateboard.

Image-Text pairs (2B examples)

A little girl playing with a flashlight.

Video-Text pairs (27M examples)




# M3W: Massive MultiModal Web Dataset

slide credit: Jean-Babtiste Alayrac

**44M** scraped webpages with interleaved text and images.

**180M** images in total. (4 on average per webpage)

Processing

```
16 Funny-Shaped Fruits And Vegetables That
Forgot How To Be Plants

You'd think that a carrot is a carrot, but
that's just not the case - some carrots are
just carrots, and others are also
intergalactic superheroes. [...]. Some
farmers even grow pears that look like
Buddha!

Now, scroll down below and check these
funny photos of fruits and veggies for
yourself!

A Sophisticated Radish

<IMAGE PLACEHOLDER 1>

StrawBEARy

<IMAGE PLACEHOLDER 2>

Toy Story's Buzz Lightyear As A Carrot

<IMAGE PLACEHOLDER 3>
```

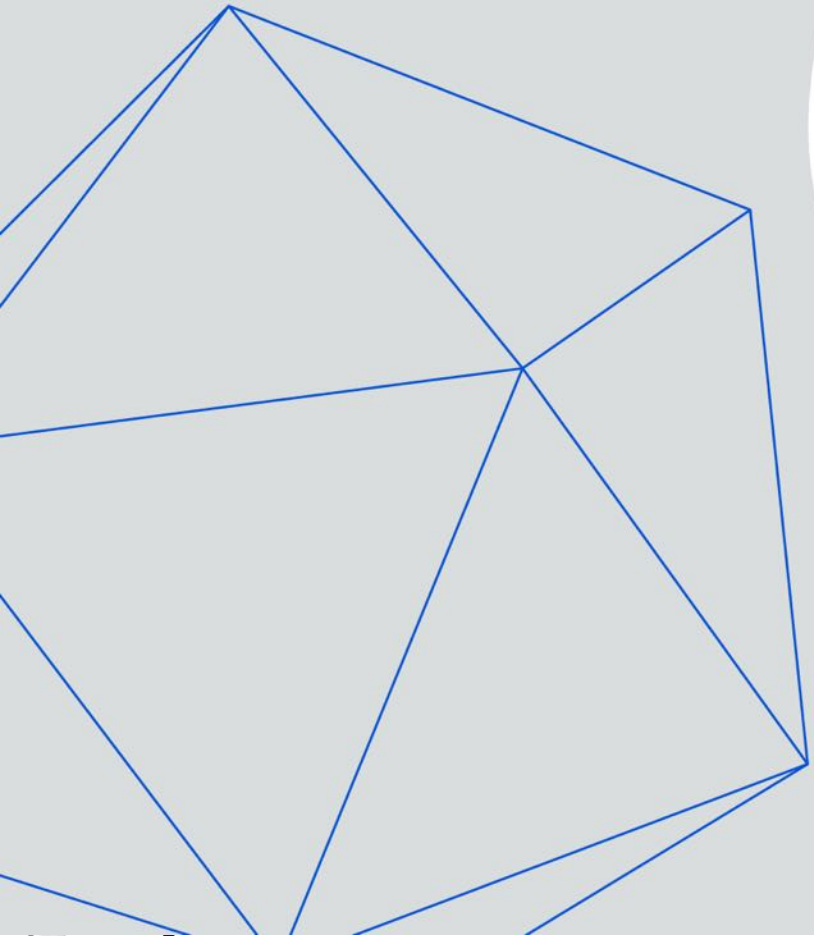
<IMAGE PLACEHOLDER 1> →

<IMAGE PLACEHOLDER 2> →

<IMAGE PLACEHOLDER 3> →

slide credit: Jean-Babtiste Alayrac

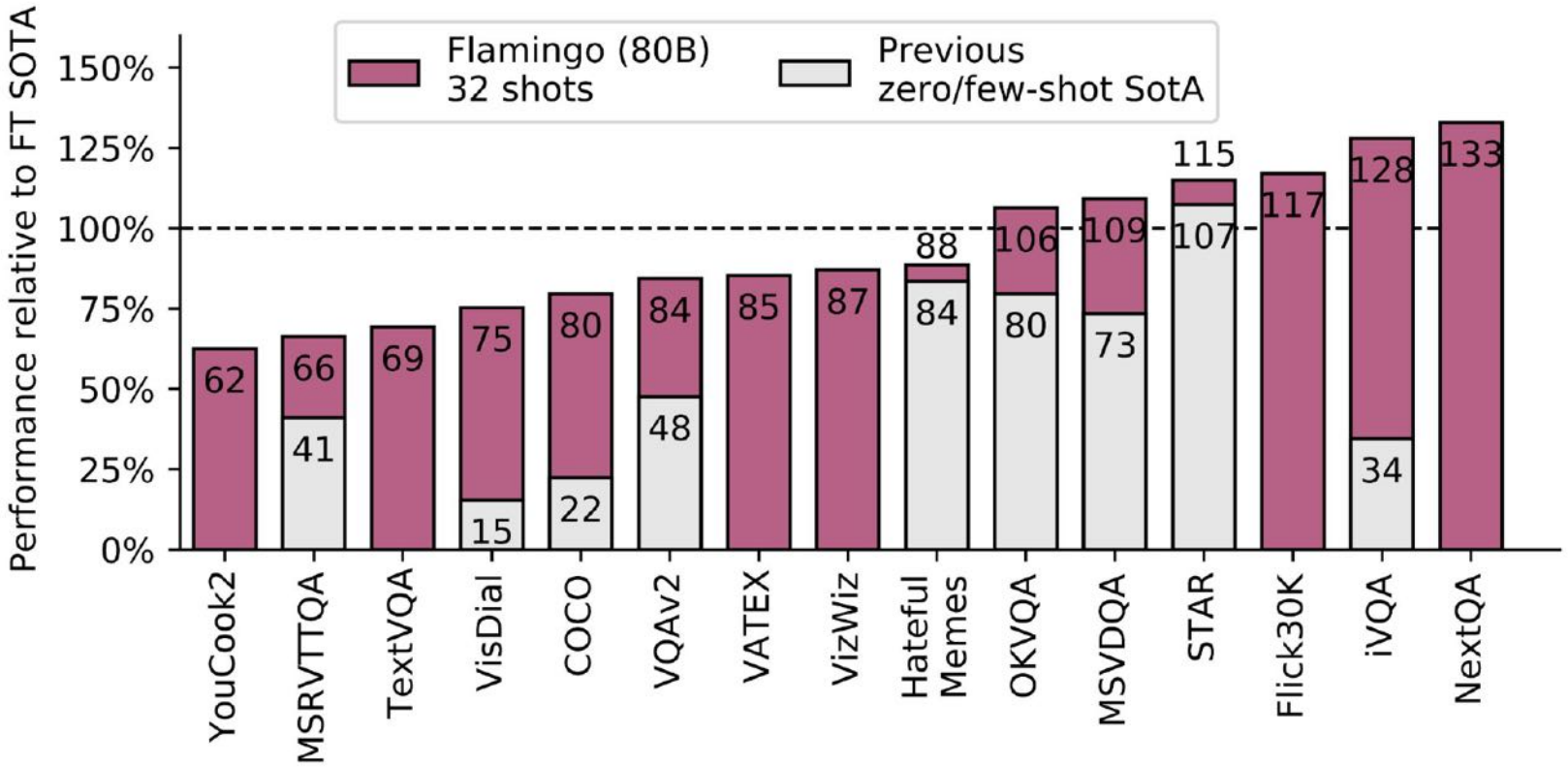
DeepMind



# Experiments Results

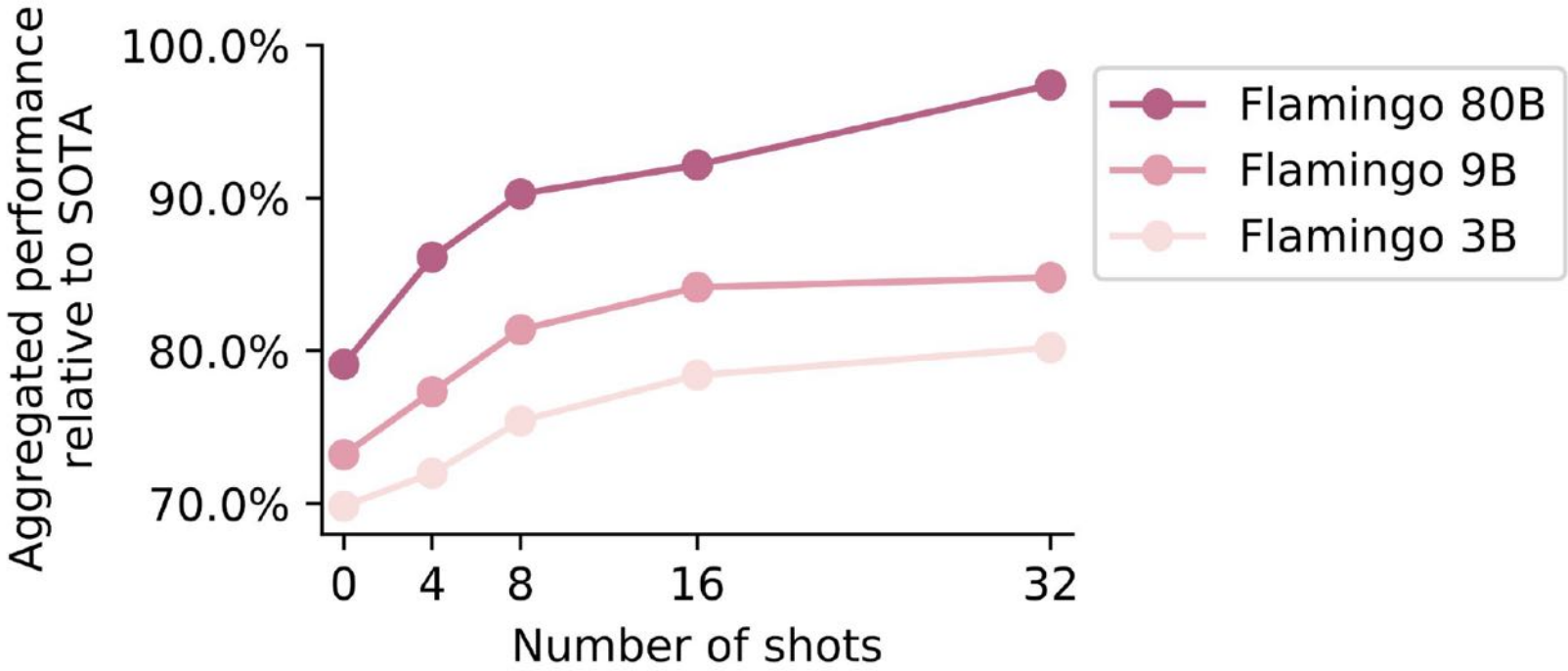
slide credit: Jean-Babtiste Alayrac

# Main results: Few-shot results



slide credit: Jean-Babtiste Alayrac

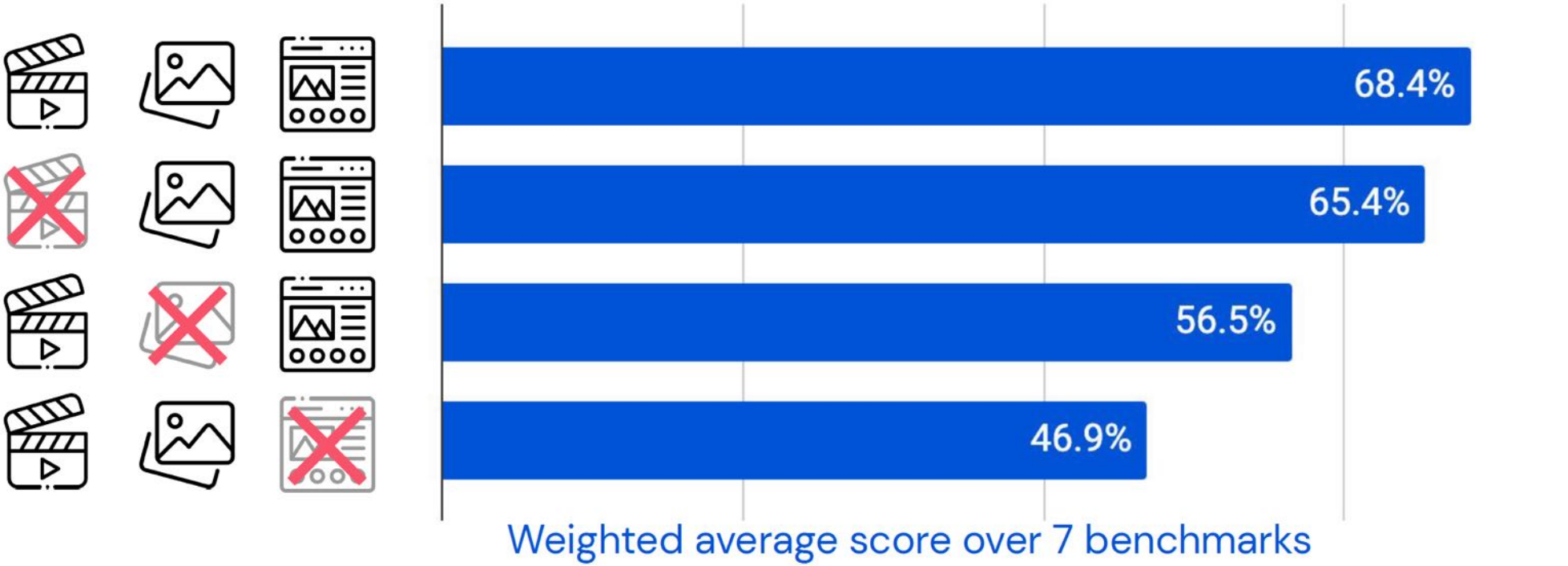
# Scaling model size / number of shot



slide credit: Jean-Babtiste Alayrac

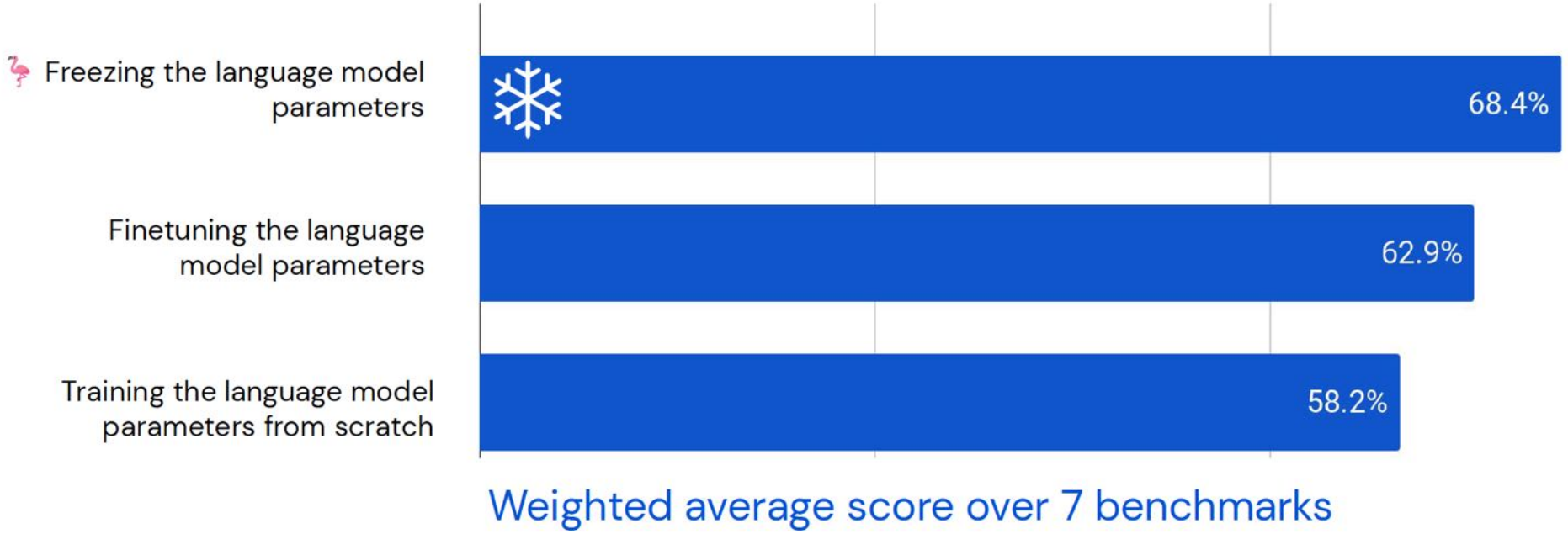
# On the importance of the training data



slide credit: Jean-Babtiste Alayrac

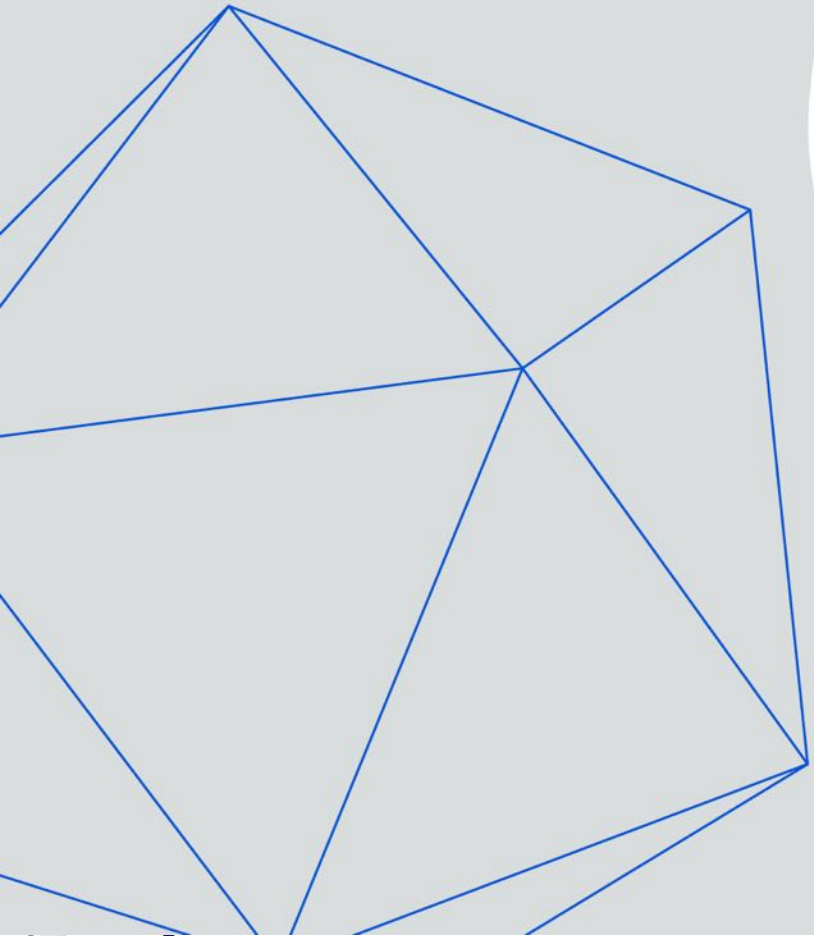
# On the importance of freezing the language model




Freezing the language model lead to the best results while also being cheaper to train.

slide credit: Jean-Babtiste Alayrac

DeepMind



# Qualitative examples

slide credit: Jean-Babtiste Alayrac

| | | | |
|---|---|---|---|
| Input Prompt |  Question: What do you think the capacities of these are? Answer: |  Question: What is odd about this image? Explain why it is unusual. Answer: |  Question: What country is this? Why do you think so? Answer: |
| Completion | **The floppy disk is 1.44MB and the CD is 700MB.** | **The image is odd because the elephant is in the back of the truck. It is unusual because elephants are not usually transported in the back of a truck.** | **It is Canada. I think so because the flag is the Canadian flag.** |
| Input Prompt |  Question: What nutrients is the fruit in the image rich in? Answer: |  Question: What is the purpose of this? Answer: |  Question: What season is this? Explain why. Answer: |
| Completion | **Vitamin C, potassium, and fiber.** | **To keep the mosquitoes out of the bed.** | **This is winter season. The Christmas tree and the presents are the main reasons why it is winter.** |

slide credit: Jean-Babtiste Alayrac


Which city is this?

This is a street sign in New York City.

Which street corner is it?

It is the intersection of Madison Avenue and East 42nd Street.

Which subway station is nearby?

The closest subway station is Grand Central Station.

How about the nearest park?

The nearest park is Bryant Park.

There is another sign. What does it say?

It says "One way".

What is the common thing about these three images?

They are all flamingos.

What is the difference between these three images?

The first one is a cartoon, the second one is a real flamingo, and the third one is a 3D model of a flamingo.

when you are fully grown but you can't resist using the playground equipment

What is funny about this meme?

This is a hamster in a slide. It is funny because the hamster is too big to fit in the slide.

Is the hamster enjoying that?

I don't think so.

How is the hamster feeling?

The hamster is feeling frustrated.

Google DeepMind

slide credit: Antoine Miech

# Limitations of the Flamingo approach

# Limitations of the Flamingo approach

slide credit: Antoine Miech

Google DeepMind

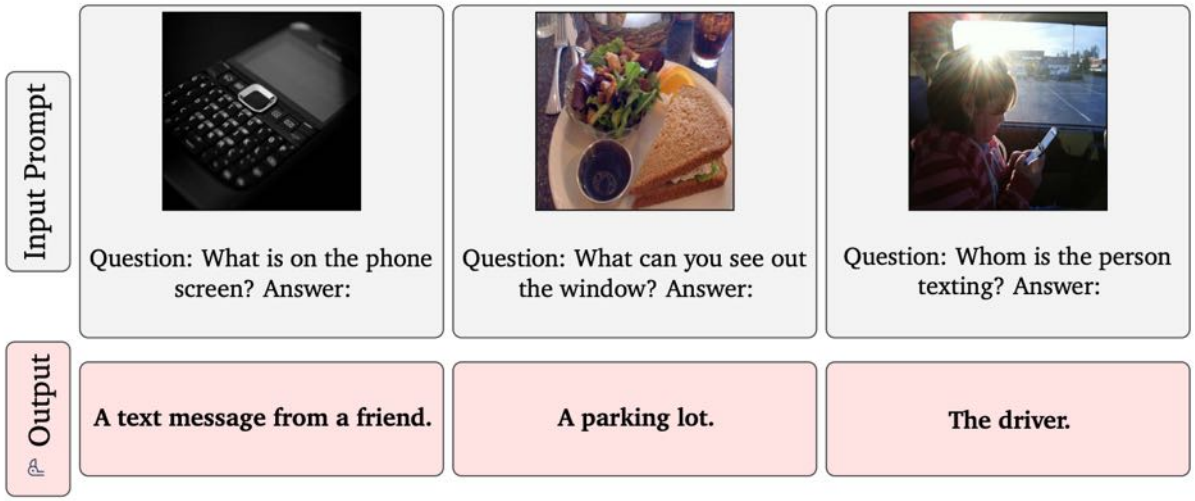
- Can only handle a few seconds of video
  (short context window)

# Limitations of the Flamingo approach

slide credit: Antoine Miech

Google DeepMind

- Can only handle a few seconds of video (short context window)
- Hallucinations

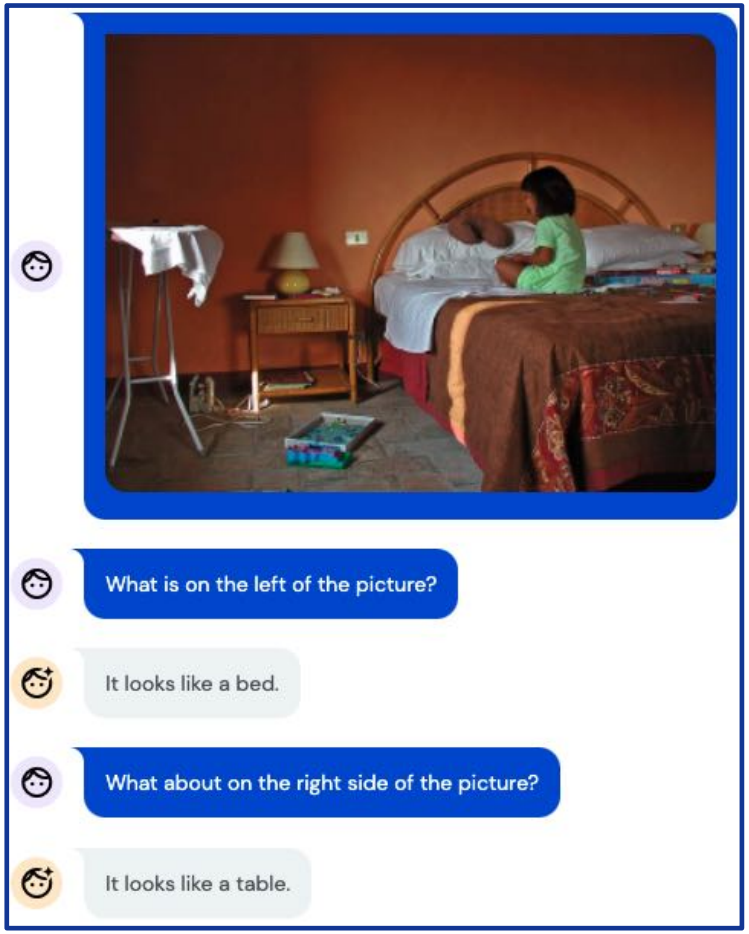


slide credit: Antoine Miech

Google DeepMind

# Limitations of the Flamingo approach

- Can only handle a few seconds of video (short context window)
- Hallucinations
- Coarse spatial understanding

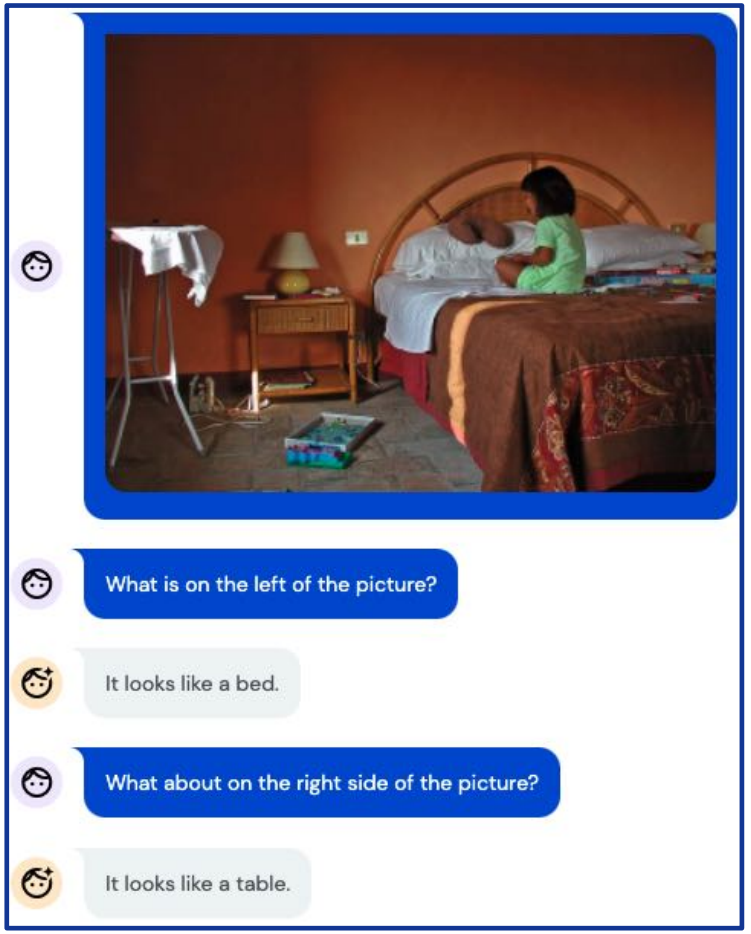


slide credit: Antoine Miech

Google DeepMind

# Limitations of the Flamingo approach

- Can only handle a few seconds of video (short context window)
- Hallucinations
- Coarse spatial understanding
- No image out / localization out
- Always require few-shot prompting to work
- English only, poor code/math



slide credit: Antoine Miech

Google DeepMind

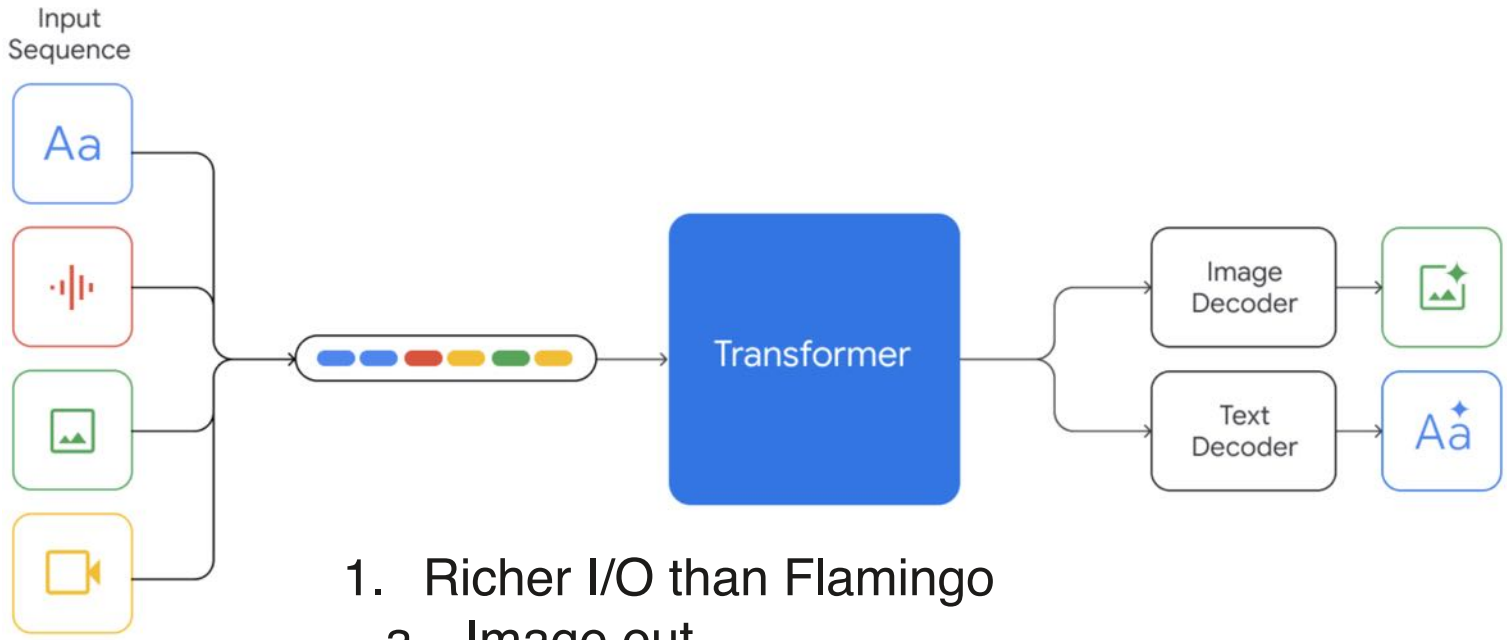
# One step further with Gemini 1.0

Gemini: a family of highly capable multimodal models, Gemini team, arXiv 2023

# Gemini model

slide credit: Antoine Miech

Google DeepMind




1. Richer I/O than Flamingo
   a. Image out
   b. Audio In
2. Natively multimodal
3. Natively multilingual
4. Better at code / math
5. Instruction tuned

# Gemini 1.0 can output arbitrarily interleaved sequences of image and text

slide credit: Antoine Miech

Google DeepMind

**Prompt**

Give me two ideas that I could do with these 2 colors

Colors: I see blue and yellow yarn

How about a cute blue cat?

Or a blue dog that would also have a yellow ear?

Give me two ideas that I could do with these 2 colors

**Response**

Colors: I see green and pink yarn

Idea 1: How about a green avocado with pink seed?

Idea 2: Or a green bunny with pink ears?

# Gemini is instruction tuned no need for few-shot examples

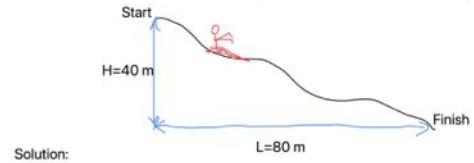
slide credit: Antoine Miech

Google DeepMind

## Prompt

Here's a solution to a Physics problem by a student.

23. A skier slides down the frictionless slope as shown. What is the skier's speed at the bottom?




Solution:

The conservation of energy means that the energy of the skier is same at the beginning and at the end.

Start: $E = mgL$

End: $E = \frac{1}{2}mv^2$

$\Rightarrow mgL = \frac{1}{2}mv^2$

$\Rightarrow v = \sqrt{2gL} = \sqrt{2 \times 9.81 \times 80} = 39.6\ m.s^{-1}$

Try to reason about the question step by step.

Did the student get the correct answer? If the solution is wrong, please explain what is wrong and solve the problem.

Make sure to use LaTeX for math and round off the final answer to two decimal places.

## Response

The student did not get the correct answer. The student made a mistake in the calculation of the potential energy at the start of the slope. The potential energy at the start of the slope is given by $mgh$, where $m$ is the mass of the skier, $g$ is the acceleration due to gravity, and $h$ is the height of the slope. The student used the length of the slope, $L$, instead of the height, $H$, in the calculation of the potential energy.

The correct solution is:

The conservation of energy means that the total energy of the skier is the same at the beginning and at the end of the slope.

At the start of the slope, the skier has only potential energy, which is given by $mgh$.

At the end of the slope, the skier has only kinetic energy, which is given by $\frac{1}{2}mv^2$

Therefore, we can write:

$$mgh = \frac{1}{2}mv^2$$

Canceling out the mass $m$ on both sides, we get:

$$gh = \frac{1}{2}v^2$$

Solving for $v$, we get:

$$v = \sqrt{2gh}$$

Substituting the given values, we get:

$$v = \sqrt{2 \times 9.81 \times 40} = 28.01\ m/s$$

Therefore, the skier's speed at the bottom of the slope is 28.01 m/s.

# State of the art VLM
## *Image Understanding*

slide credit: Antoine Miech

Google DeepMind

| | Gemini Ultra (pixel only) | Gemini Pro (pixel only) | Gemini Nano 2 (pixel only) | Gemini Nano 1 (pixel only) | GPT-4V | Prior SOTA | Flamingo 32 shots |
|---|---|---|---|---|---|---|---|
| **MMMU (val)** Multi-discipline college-level problems (Yue et al., 2023) | **59.4%** pass@1 **62.4%** Maj1@32 | 47.9% | 32.6% | 26.3% | 56.8% | 56.8% GPT-4V, 0-shot | N.A. |
| **TextVQA (val)** Text reading on natural images (Singh et al., 2019) | **82.3%** | 74.6% | 65.9% | 62.5% | 78.0% | 79.5% Google PaLI-3, fine-tuned | 36% |
| **DocVQA (test)** Document understanding (Mathew et al., 2021) | **90.9%** | 88.1% | 74.3% | 72.2% | 88.4% (pixel only) | 88.4% GPT-4V, 0-shot | N.A. |
| **ChartQA (test)** Chart understanding (Masry et al., 2022) | **80.8%** | 74.1% | 51.9% | 53.6% | 78.5% (4-shot CoT) | 79.3% Google DePlot, 1-shot PoT (Liu et al., 2023) | N.A. |
| **InfographicVQA (test)** Infographic understanding (Mathew et al., 2022) | **80.3%** | 75.2% | 54.5% | 51.1% | 75.1% (pixel only) | 75.1% GPT-4V, 0-shot | N.A. |
| **MathVista (testmini)** Mathematical reasoning (Lu et al., 2023) | **53.0%** | 45.2% | 30.6% | 27.3% | 49.9% | 49.9% GPT-4V, 0-shot | N.A. |
| **AI2D (test)** Science diagrams (Kembhavi et al., 2016) | **79.5%** | 73.9% | 51.0% | 37.9% | 78.2% | 81.4% Google PaLI-X, fine-tuned | N.A. |
| **VQAv2 (test-dev)** Natural image understanding (Goyal et al., 2017) | **77.8%** | 71.2% | 67.5% | 62.7% | 77.2% | 86.1% Google PaLI-X, fine-tuned | 67.6% |

# From Flamingo to Gemini 1.0

slide credit: Antoine Miech

Google DeepMind

- **Positive transfer from Language -> Vision:** Better LLMs lead to better VLMs: for example better reasoning, multilingual transfer.

- **Richer I/O:** Now can model images in its output

- **No need for few-shot prompting:** just tell it what to do in text!

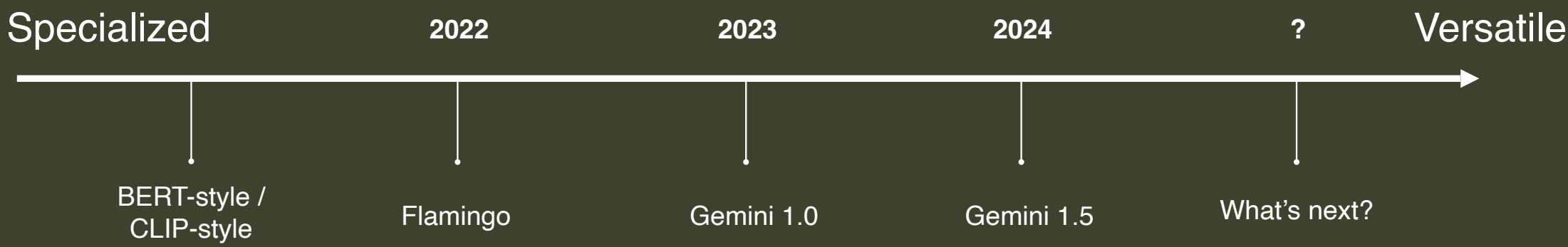
- Some of the **limitations** included hallucination, spatial understanding and issues with complex scenes have been improved but are still present

slide credit: Antoine Miech

Google DeepMind

# Gemini 1.5

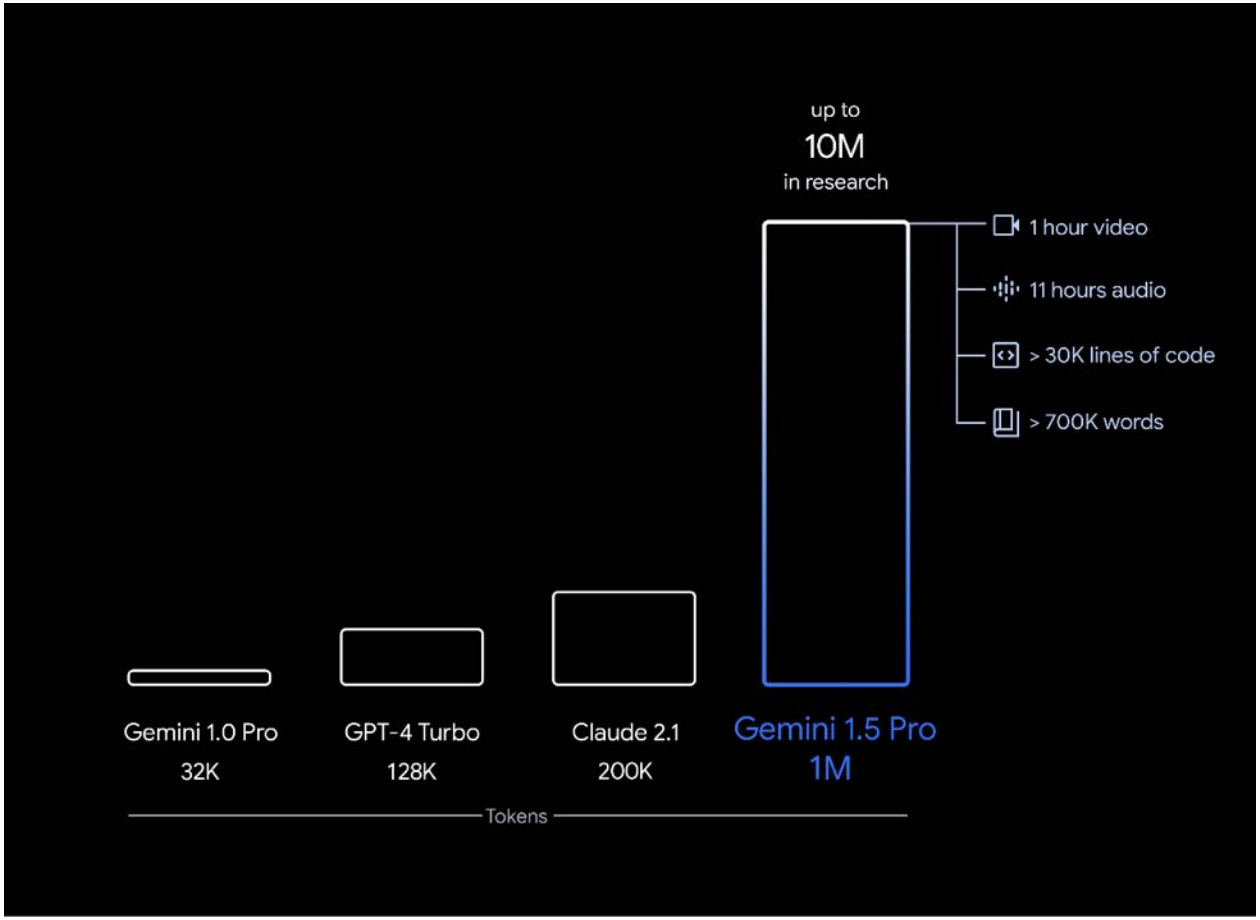
Unlocking multimodal understanding across millions of tokens of context



# Gemini 1.5, a Large Scale Multimodal Model with million-length context support

slide credit: Antoine Miech

Google DeepMind



Google DeepMind

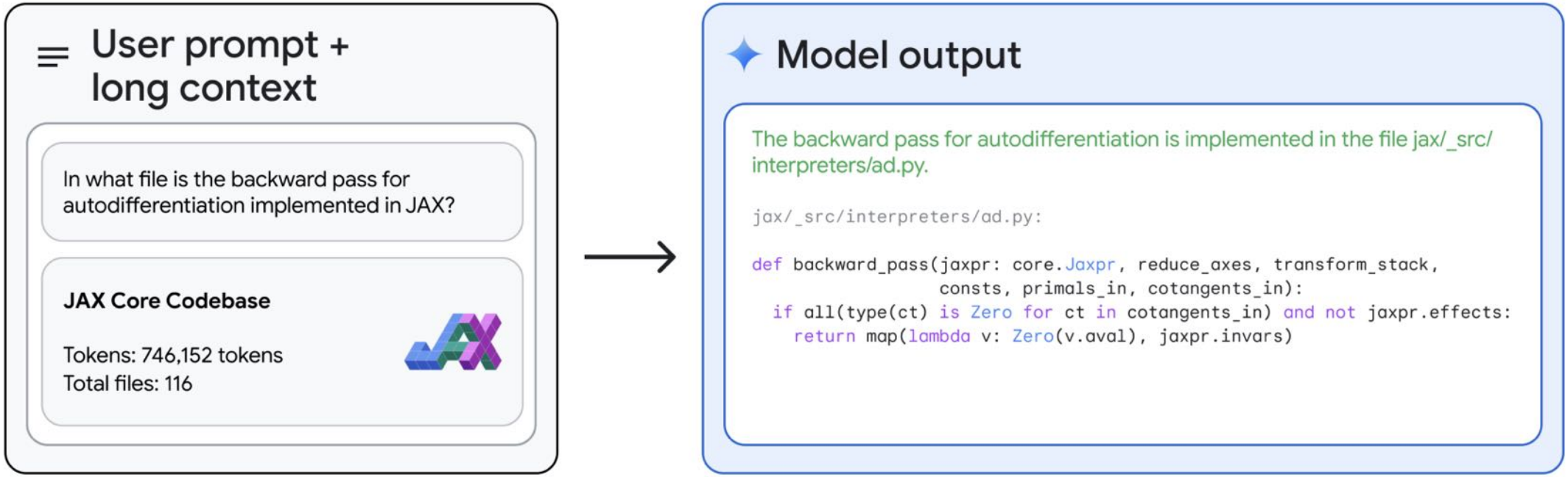
slide credit: Antoine Miech

# 10M?

# When O(1M) context length matters

slide credit: Antoine Miech

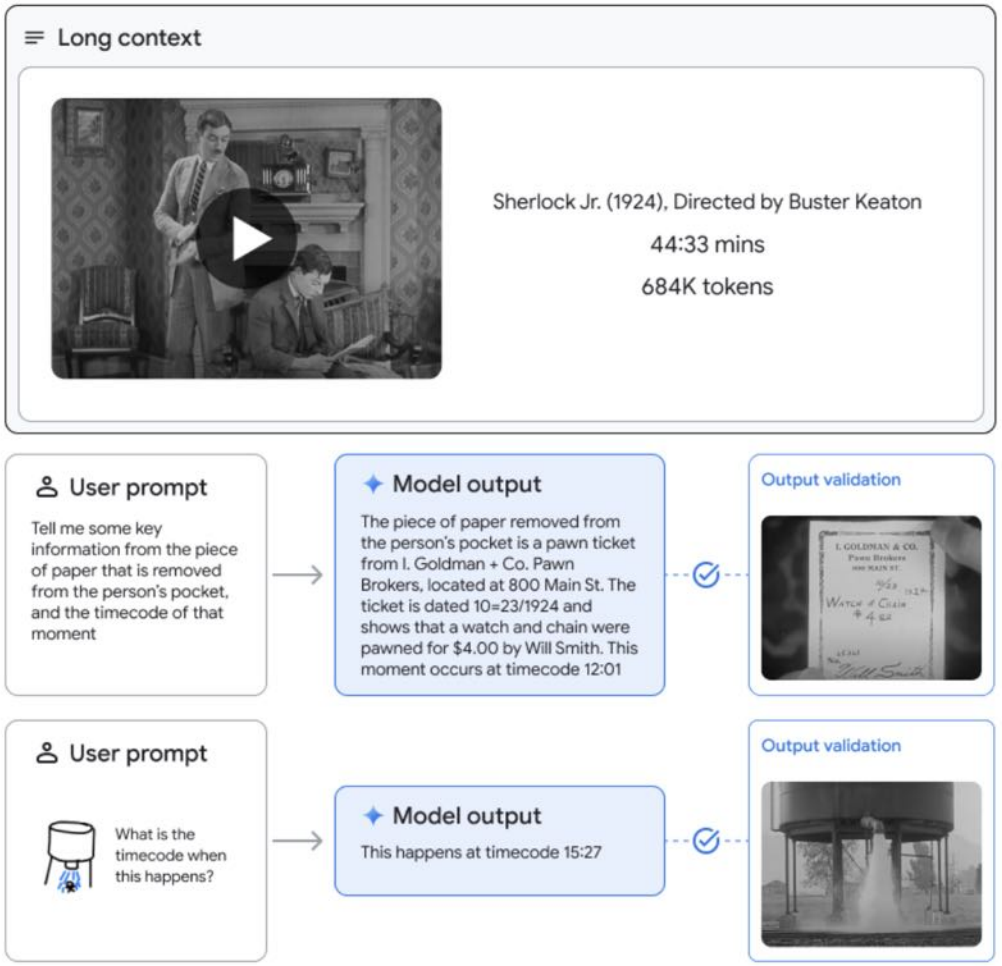
Google DeepMind



# When O(1M) context length matters

slide credit: Antoine Miech

Google DeepMind



# Measuring the long-context ability
## Needle in the Haystack across Vision, Text and Audio

slide credit: Antoine Miech

Google DeepMind

1. **The haystack:** Take a long document (e.g. the 1h30 hour long AlphaGo DeepMind movie)

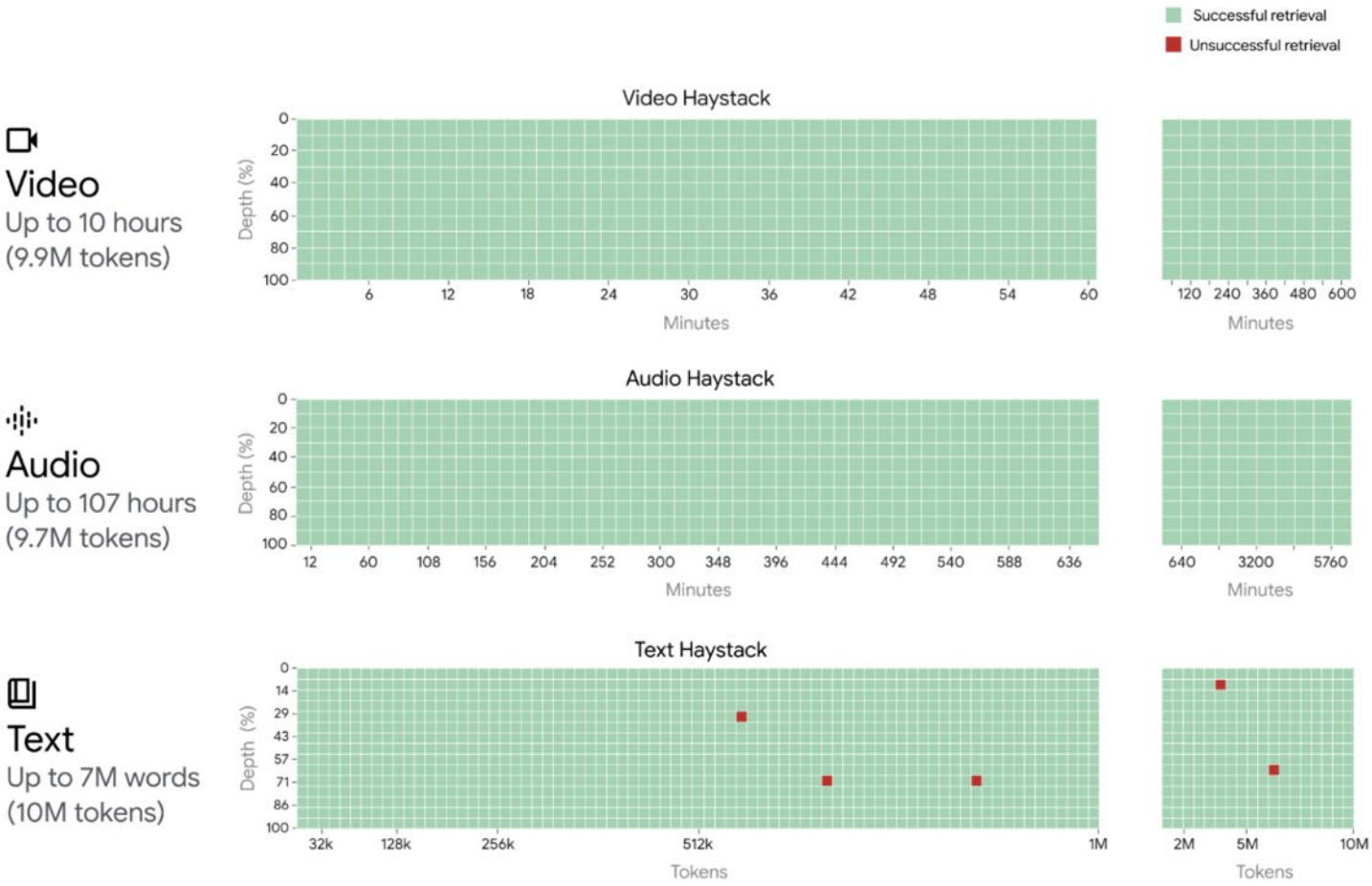
2. **The needle:** Insert a “needle” (e.g. a word written on a random frame inside the video)

3. **Verify whether the model is able to find the needle at different timecodes in the video**

slide credit: Antoine Miech

# Needle in the Haystack across Vision, Text and Audio

Google DeepMind



Google DeepMind

# What's next?

# The progression so far of Transformers for VLMs

slide credit: Antoine Miech

Google DeepMind


- Specialized Vision-Language transformers
  - BERT-style
  - CLIP-style
- Great at specialized tasks
- Not versatile enough

Single unified Large Vision language model architectures for a large range of tasks:

- VQA, Captioning, Summarisation, Temporal localization
- Richer I/O interface
  - Multi Image In / Out
  - Audio In / Out
  - Long (+1h) video
- Various image domains: Photographs, documents, charts, tables, screenshots
- To some extent, versatile to new tasks

- More agentic behaviours?
  - Control webs browser
  - Play video game
  - Robot actions
- Even richer I/O interface
  - Video out
  - 3D in / out
  - Low level computer vision tasks