

Module 4

Multi-modal Language Models (MLLMs)

Beyond text-based transformers



Learning Objectives

By the end of this module you will:

- Survey the broad landscape of multi-modality that leverages LLMs
- Understand how transformers accept non-text inputs, e.g. images and audio
- Examine how multi-modal data is procured
- Discuss limitations of multi-modal LLMs and alternative architectures to transformers
- Identify the wide possibilities of multi-modal applications



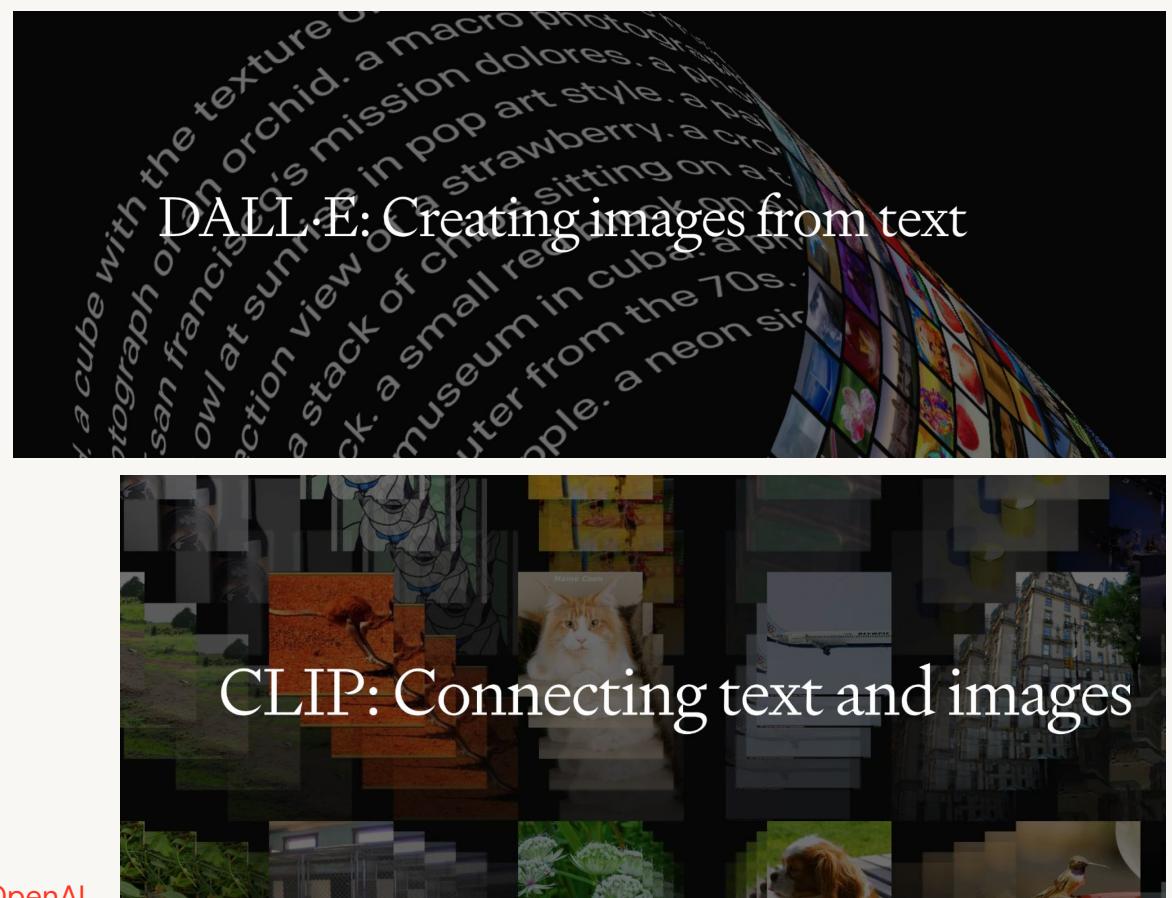
Going beyond uni-modality

LLM-based models that can receive and reason with multimodal info

Source: [Yin et al 2023](#) (released in late June)



Source: [OpenAI](#)



CLIP: Connecting text and images



Multi-modality mirrors how we perceive info

More user-friendly, flexible, and capable

Video-LLaMA

Describe what you hear?

I hear sounds of **footsteps** on the floor in the background, and there is also a **dog barking** in the house.

Does the man wear glasses?

Yes, he wears a **pair of glasses** in the video.

Source: [Zhang et al 2023](#)
[Demo for Video LLaMA](#)

MiniGPT-4

Explain why this meme is funny.

This meme is funny because it shows a dog laying on the floor with its head resting on its paws, as if it is tired or sleepy. The caption, "monday just monday," adds to the humor by suggesting that the dog is feeling the same way as many people do on Mondays, which are often considered to be the most dreaded day of the week. The meme is relatable and humorous, making it a popular one among internet users.

Source: [Zhu et al 2023](#)

Write brief HTML/JS to turn this mock-up into a colorful website, where the jokes are replaced by a real joke.

HTML:

```
<html>
<head>
<title>A Joke Website</title>
<link rel="stylesheet" type="text/css" href="style.css"/>
</head>
<body>
<div id="header">
<h1>A Joke Website</h1>
<nav>
<ul>
<li><a href="#">Home</a></li>
<li><a href="#">About</a></li>
<li><a href="#">Contact</a></li>
</ul>
</nav>
</div>
<div id="main">
<h2>Today's Joke</h2>
<p>Why did the tomato turn red?</p>
<p>Because it saw the salad dressing!</p>
</div>
<div id="footer">
<p>© 2023 A Joke Website</p>
</div>
<script src="script.js"></script>
</body>
</html>
```



Chain-of-Thought MLLMs

We can also supply multi-modal information as “in-context”



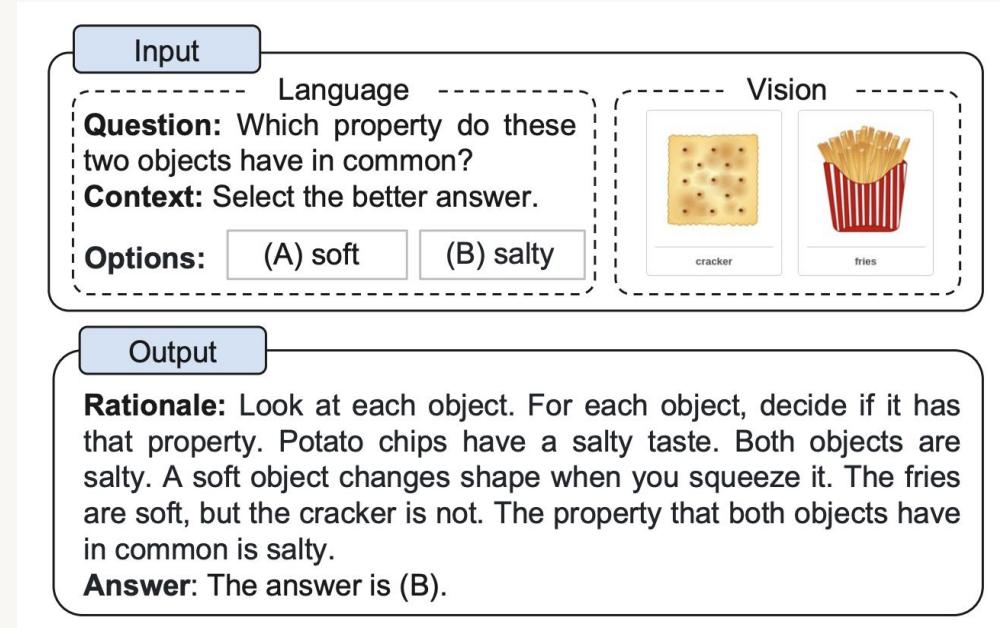
Keyframes Only

Let's think frame by frame.

What happened between the two keyframes?

A man is holding a small dog in his arms, and it appears that he is giving the dog to someone. This suggests that the man might be receiving the dog as a gift or adopting the dog from someone. The scene takes place in front of a wooden counter, and there is a box, possibly containing a holiday decoration, nearby. The man is smiling, indicating that the gift or adoption is a happy occasion for him.

Source: [Himakunthala et al 2023](#)



Source: [Zhang et al 2023](#)



MLMs can process multi-modalities simultaneously



Write a story to a child about what you see and hear.



A recording of calm ocean waves.



The woman has blonde hair.

What color is the woman's hair?



Source: [Su et al 2023](#) (released in May)
[PandaGPT Demo](#)



PandaGPT

In the image, there is a woman standing on a rock near the water, looking out at the ocean. The ocean is calm and reflects the sky, creating a serene atmosphere. The woman is enjoying the view and listening to the soothing sound of the waves crashing against the shore. As she stands there, she notices a bird flying by, adding a touch of life and movement to the scene. The woman feels peaceful and content, taking in the beauty of the ocean and the natural world around her.

MLLMs also call tools/models to finish tasks

HuggingGPT

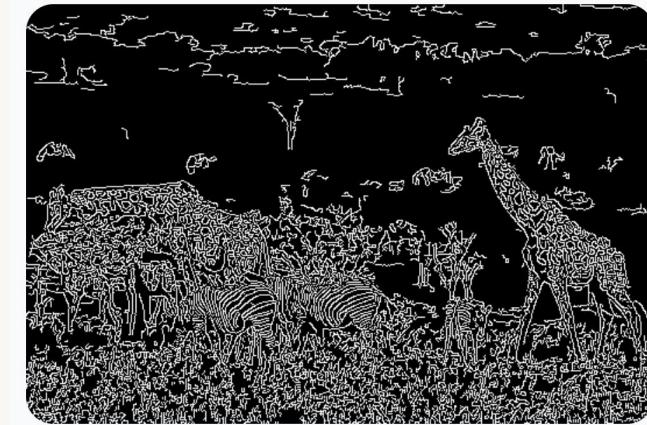


[HuggingGPT Demo](#)

Please generate a canny image based on /examples/f.jpg



I have carefully considered your request and generated a canny image based on /examples/f.jpg. The workflow I used is shown in the following JSON data: [{"task": {"task": "canny-control", "id": 0, "dep": [-1], "args": {"image": "public//examples/f.jpg"}}, "inference result": {"generated image": "/images/4810.png"}, "choose model result": {"id": "canny-control", "reason": "ControlNet is the best model for this task."}}, {"task": {"task": "canny-text-to-image", "id": 1, "dep": [0], "args": {"text": "-0", "image": "public//images/4810.png"}}, "inference result": {"generated image": "/images/a27f.png"}, "choose model result": {"id": "llyasviel/sd-controlnet-canny", "reason": "ControlNet is the best model for this task."}]. The complete path or url of the generated image is '/images/a27f.png'. Please let me know if you need any further assistance.



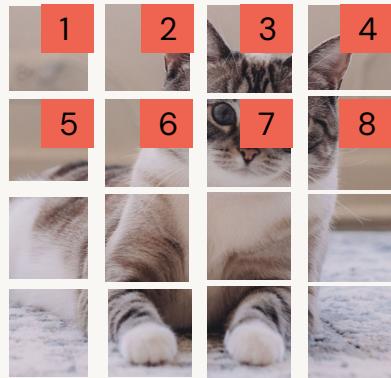
Transformers beyond text

One architecture to rule them all

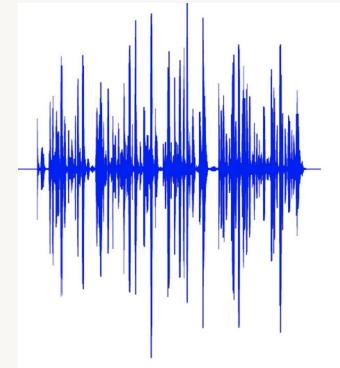


Transformer: a general sequence processing tool

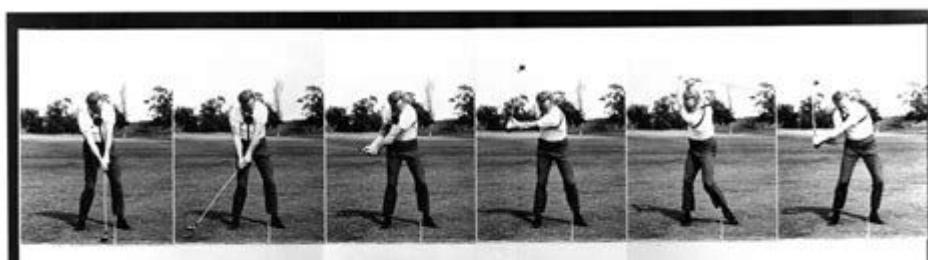
We can treat many things as a sequence



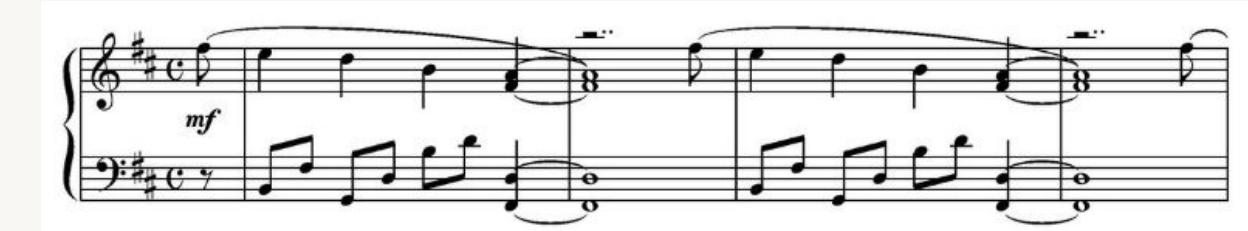
Images



Audio;
[image source](#)



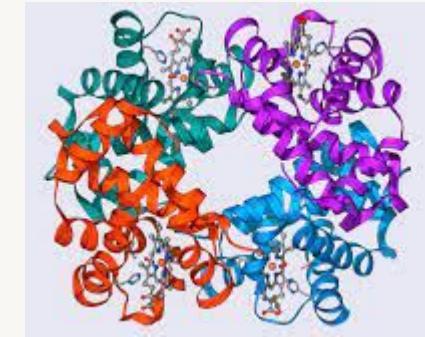
Video frames;
[image source](#)



Music notes;
[image source](#)



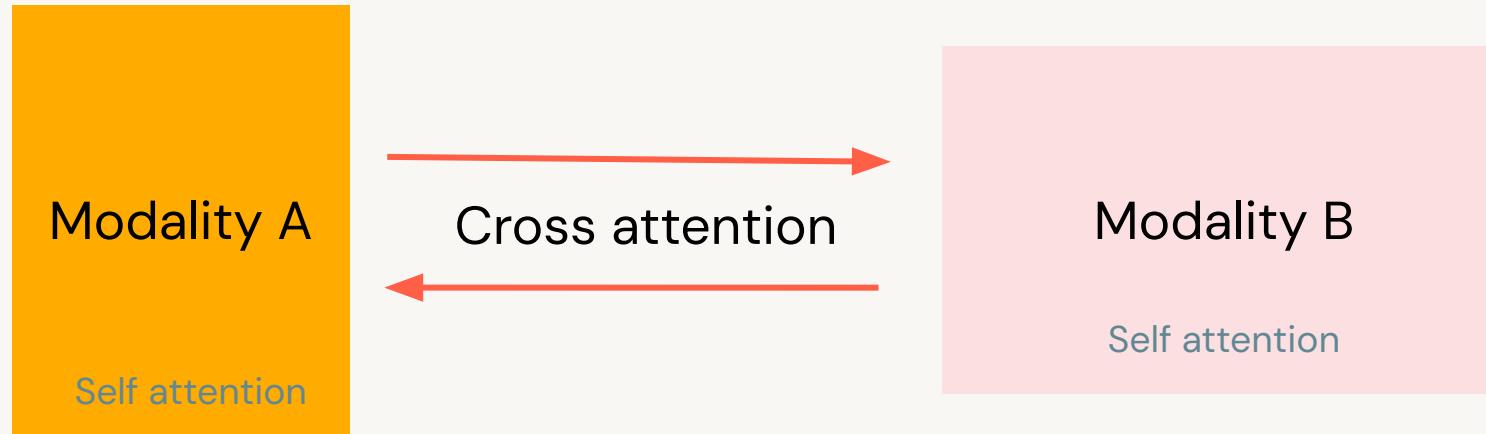
Game actions;
[image source](#)



Protein;
[image source](#)

Cross attention bridges between modalities

Allows different modalities to influence each other



A and B could be:

- Images, audio, text, time series, or any sequence!

Example:

- Stable Diffusion uses cross attention to bridge between text and images

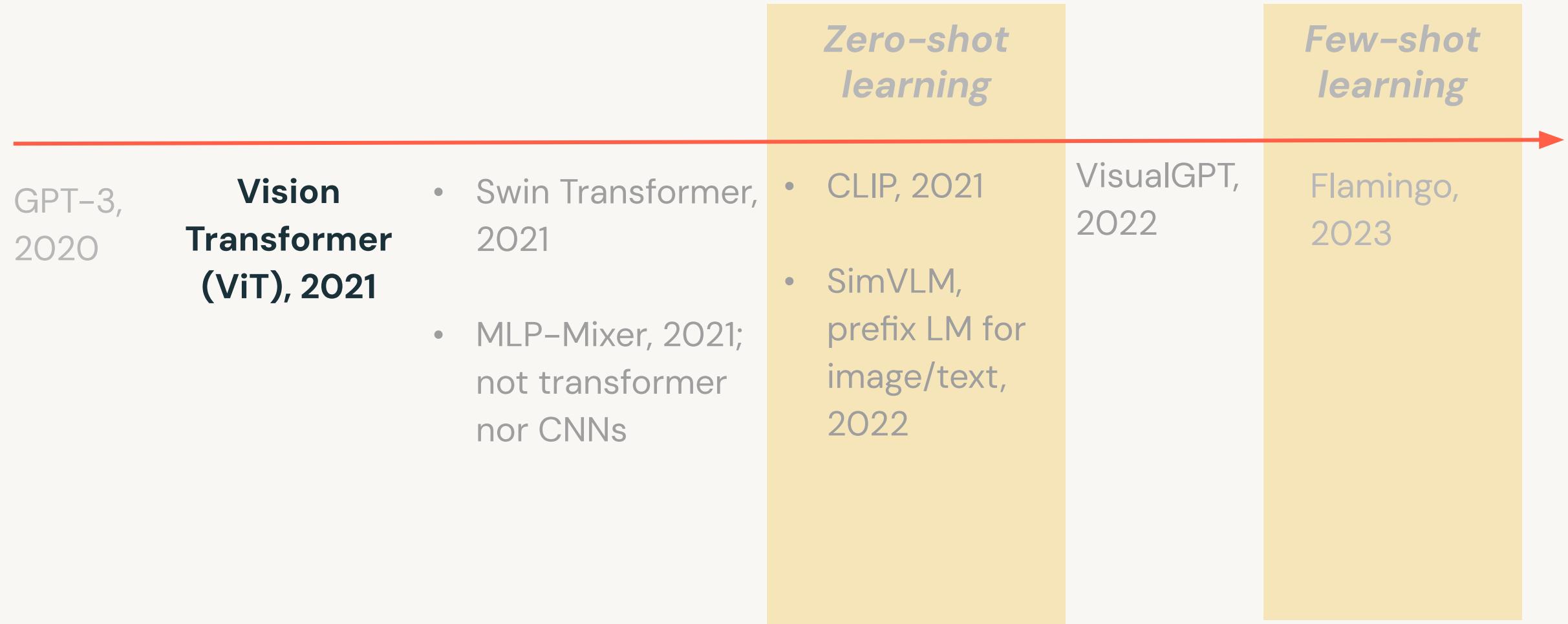


Computer vision



Well-researched area

We need to first understand how to represent images as numbers



We chop an image up into small pixels

A colored image is made up of Red, Blue, Green (RBG) levels

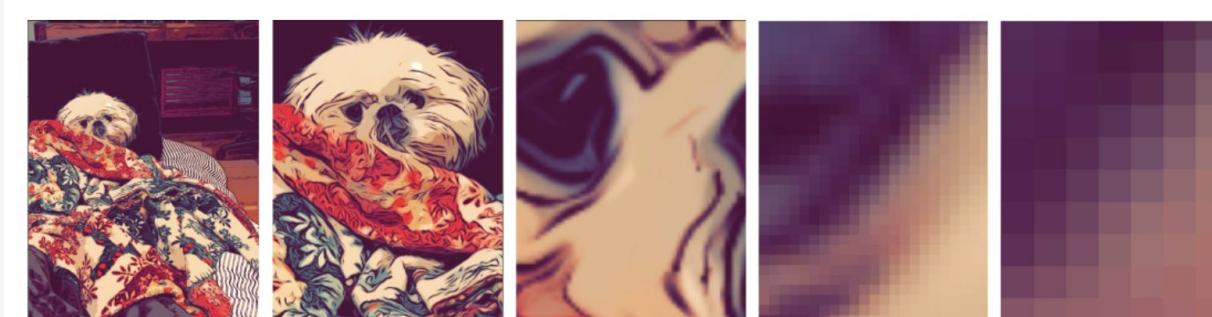


Image credit: Diane Rohrer

Each pixel range = (0, 256)

#642b4e R: 100 G: 43 B: 78	#7b4360 R: 123 G: 67 B: 96	#936073 R: 147 G: 96 B: 115
#7a4360 R: 122 G: 67 B: 96	#a1727a R: 161 G: 114 B: 122	#c89c8f R: 200 G: 156 B: 143
#945f71 R: 148 G: 95 B: 113	#ca9b91 R: 202 G: 155 B: 145	#f6d0ac R: 246 G: 208 B: 172

100	123	147
122	161	200
148	202	246

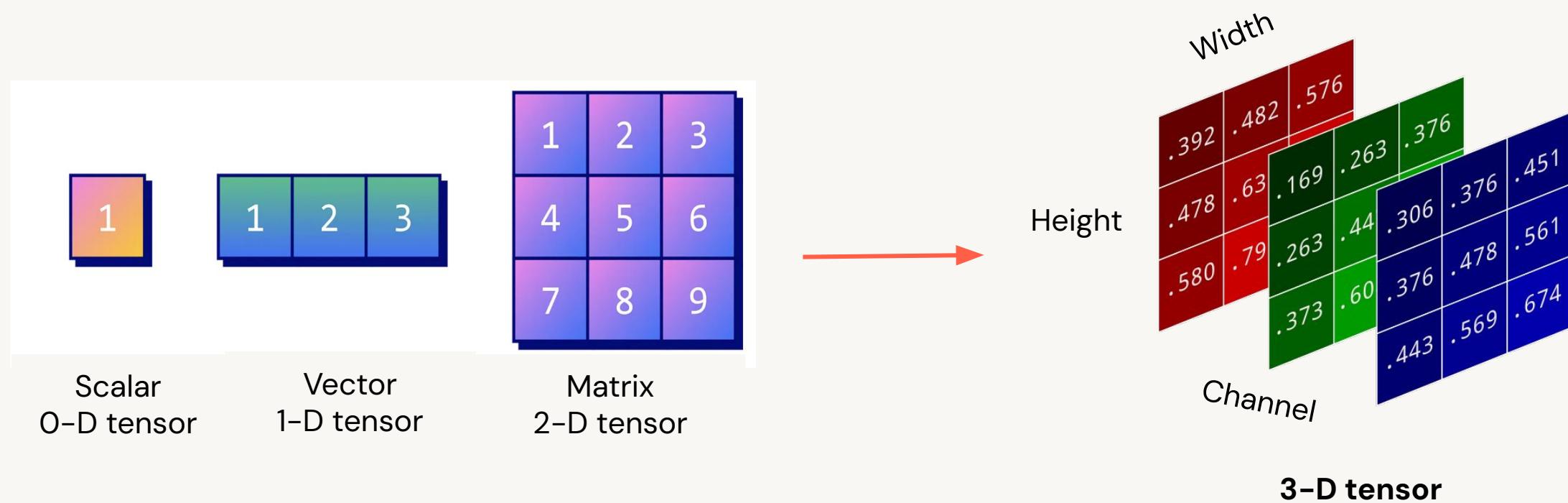
43	67	96
67	114	156
95	155	208

78	96	115
96	122	143
113	145	172



Colored images are 3-D tensors

Grayscale images are 2-D tensors: all 3 channels have the same value

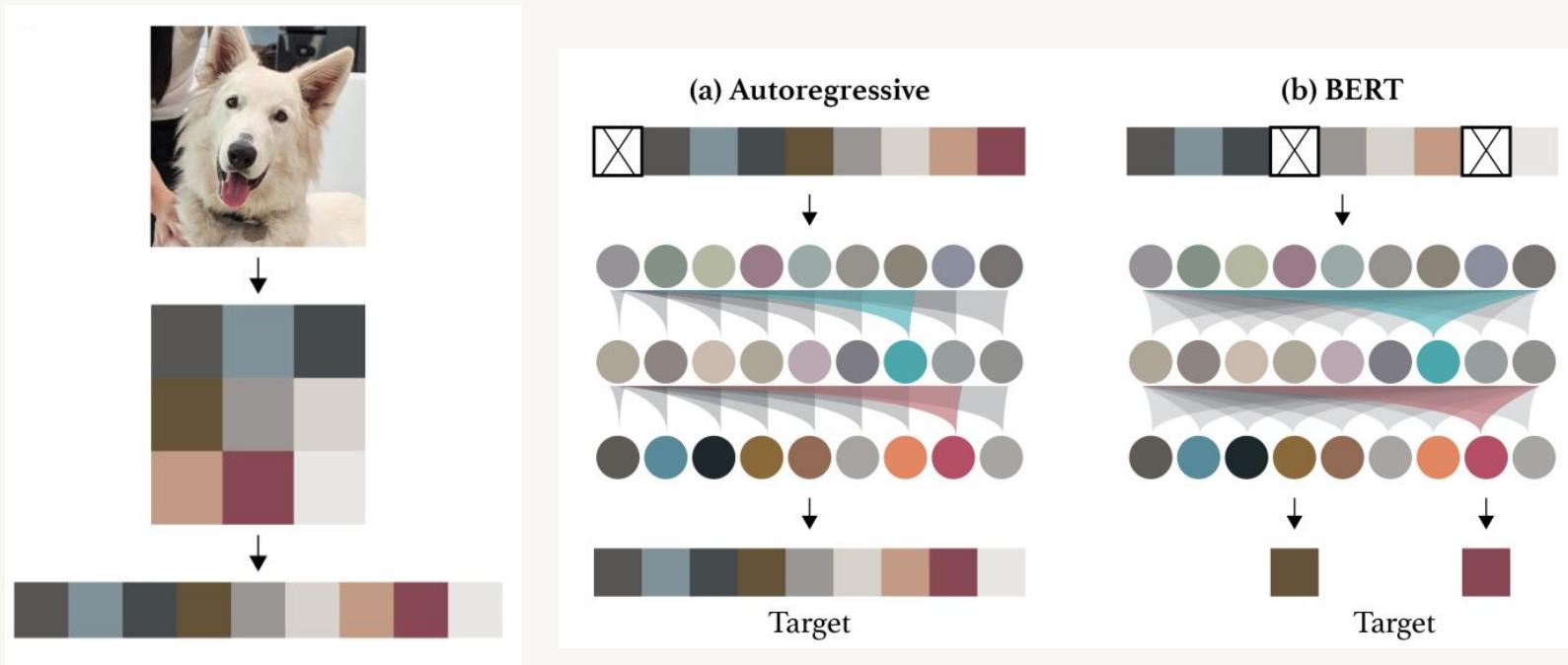


Adapted from [TowardDataScience](#) and [KD Nuggets](#)



Initial idea: Turn pixels into a sequence

Use self-attention to predict the next pixel, instead of word token



Limitations:

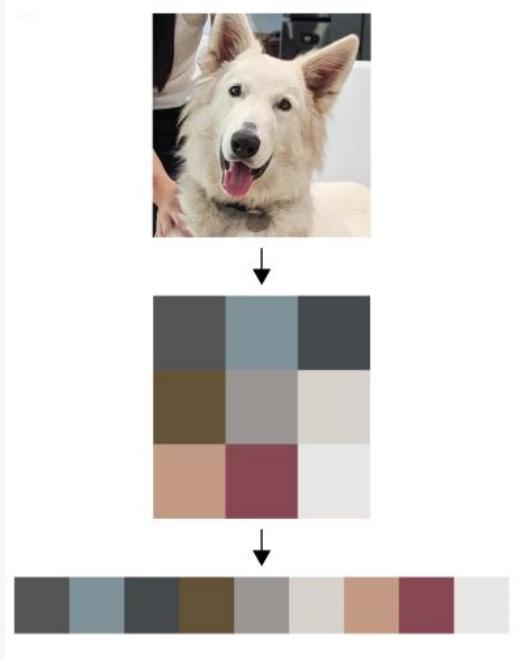
- Lose vertical spatial relationships

Source: [Chen et al 2021](#)



Initial idea: Turn pixels into a sequence

Use self-attention to predict the next pixel, instead of word token



Source: [Chen et al 2021](#)



Source: [David Cocomini 2021](#)

Limitations:

- Lose vertical spatial relationships
- Memory and computational requirements scale quadratically to sequence length, $O(N^2)$



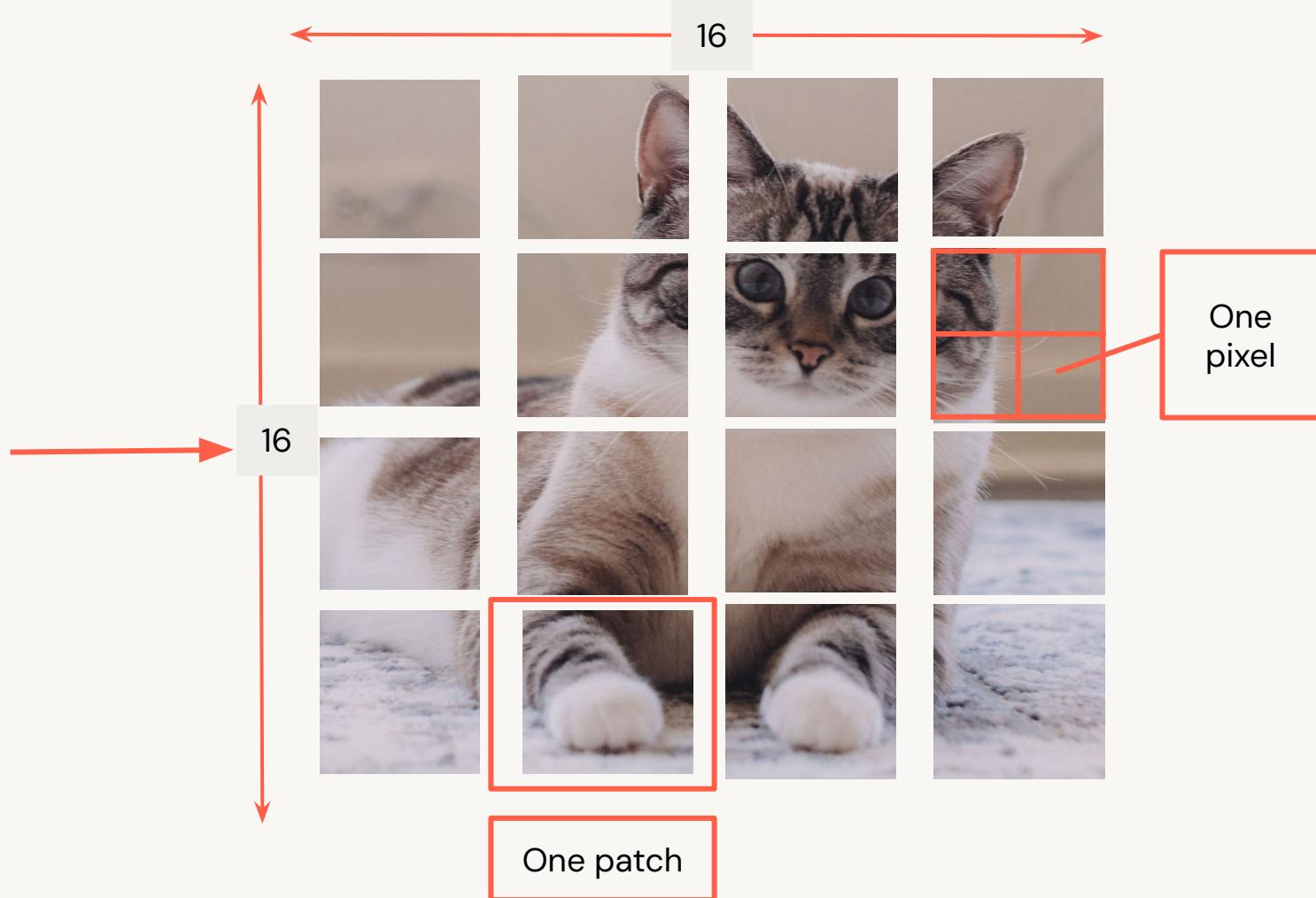
Vision Transformer (ViT)

Computes attention on patches of images: image-to-patch embeddings



Vision Transformer (ViT)

Computes attention on patches of images: image-to-patch embeddings



ViT: An image is worth 16x16 words

N input patches
with shape of
 $3 \times \text{Pixel (P)} \times P$



.....

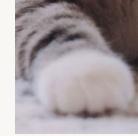


Linear project each patch to D-dimensional vector

N input patches
with shape of
 $3 \times \text{Pixel (P)} \times P$



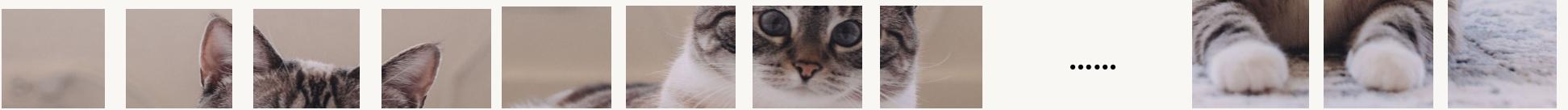
.....



D-dim patch
embedding

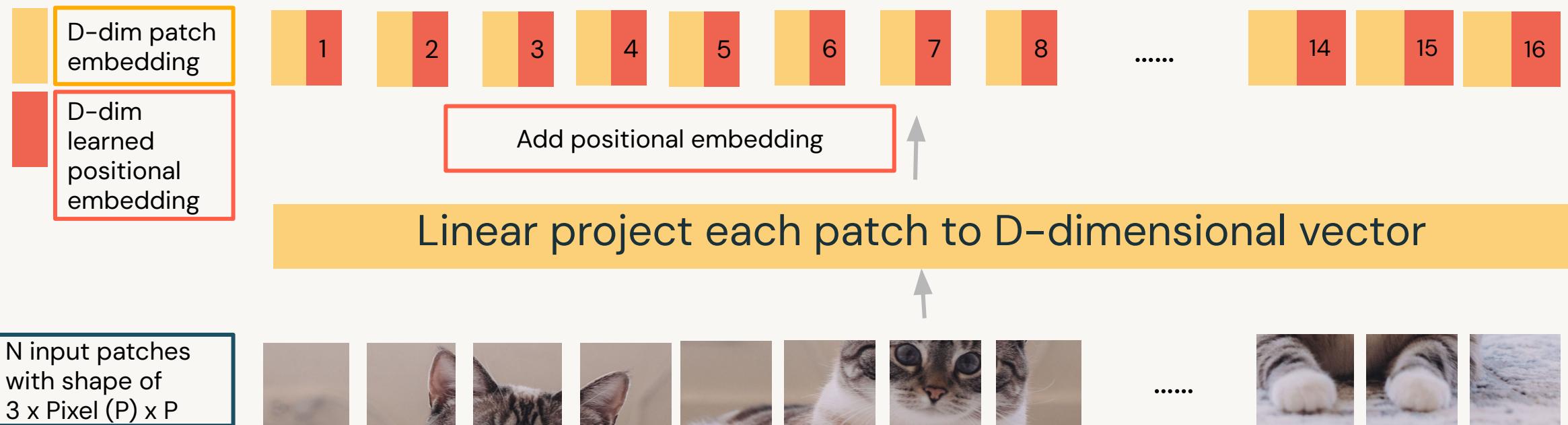


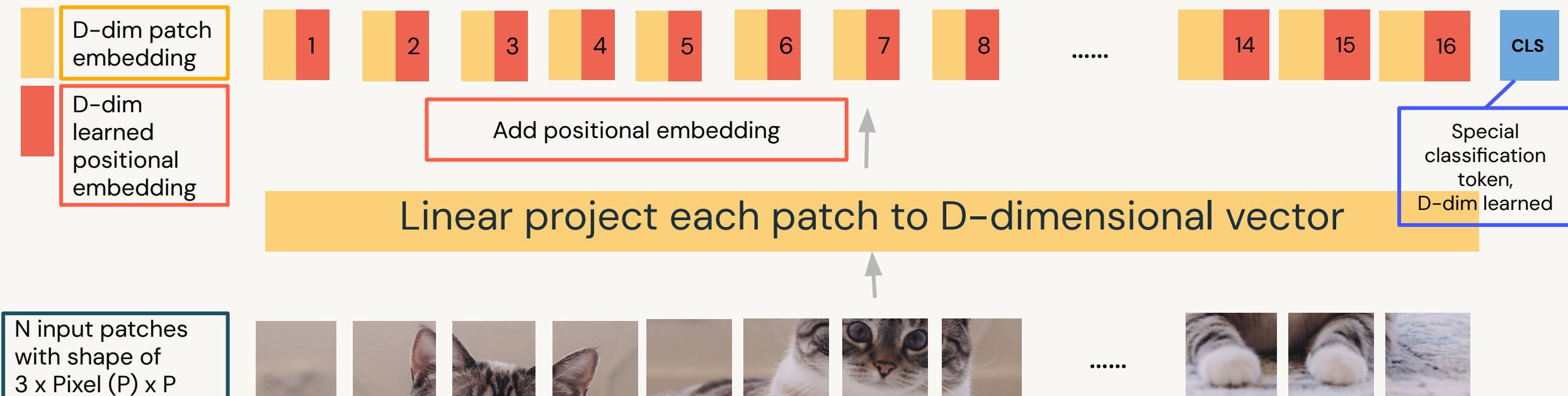
N input patches
with shape of
 $3 \times \text{Pixel (P)} \times P$



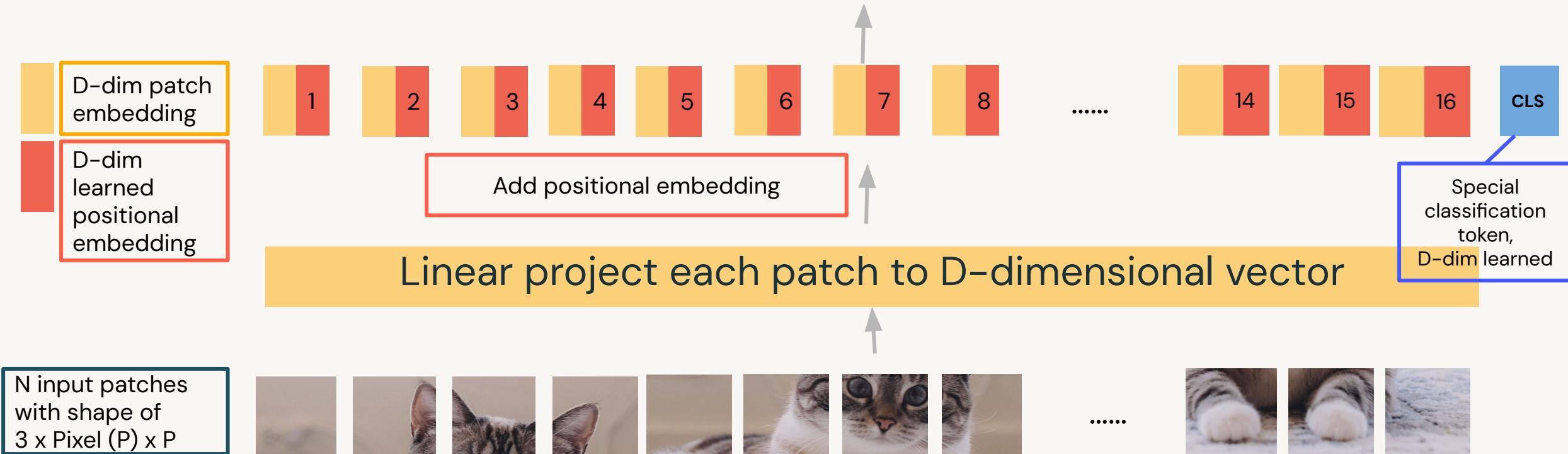
Linear project each patch to D-dimensional vector







Original Transformer



C-dim
classifier
output
vector, where
 $C = \#$ classes



Original Transformer

D-dim patch
embedding

D-dim
learned
positional
embedding

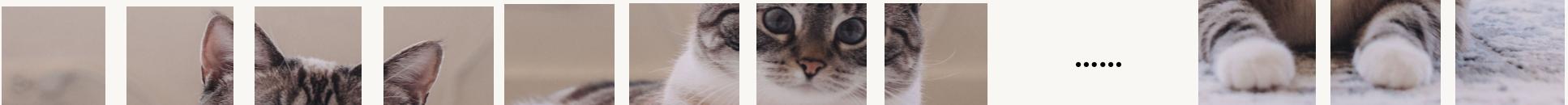


Add positional embedding

Special
classification
token,
D-dim learned

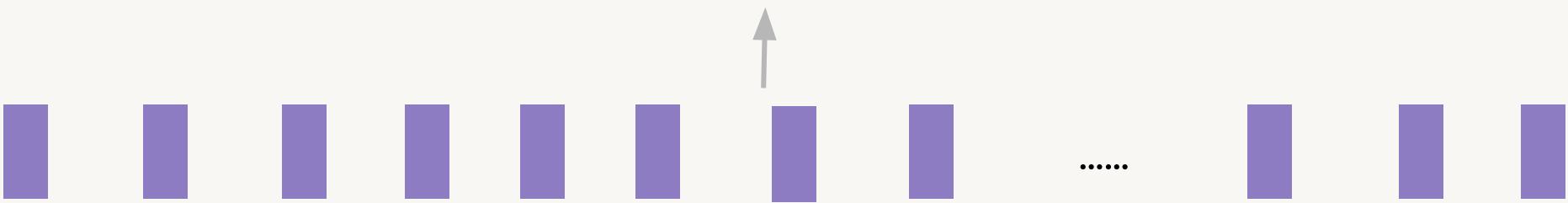
Linear project each patch to D-dimensional vector

N input patches
with shape of
 $3 \times \text{Pixel (P)} \times P$



Label = "cat"

C-dim
classifier
output
vector, where
 $C = \# \text{ classes}$



Original Transformer

D-dim patch
embedding

D-dim
learned
positional
embedding



Add positional embedding

Special
classification
token,
D-dim learned

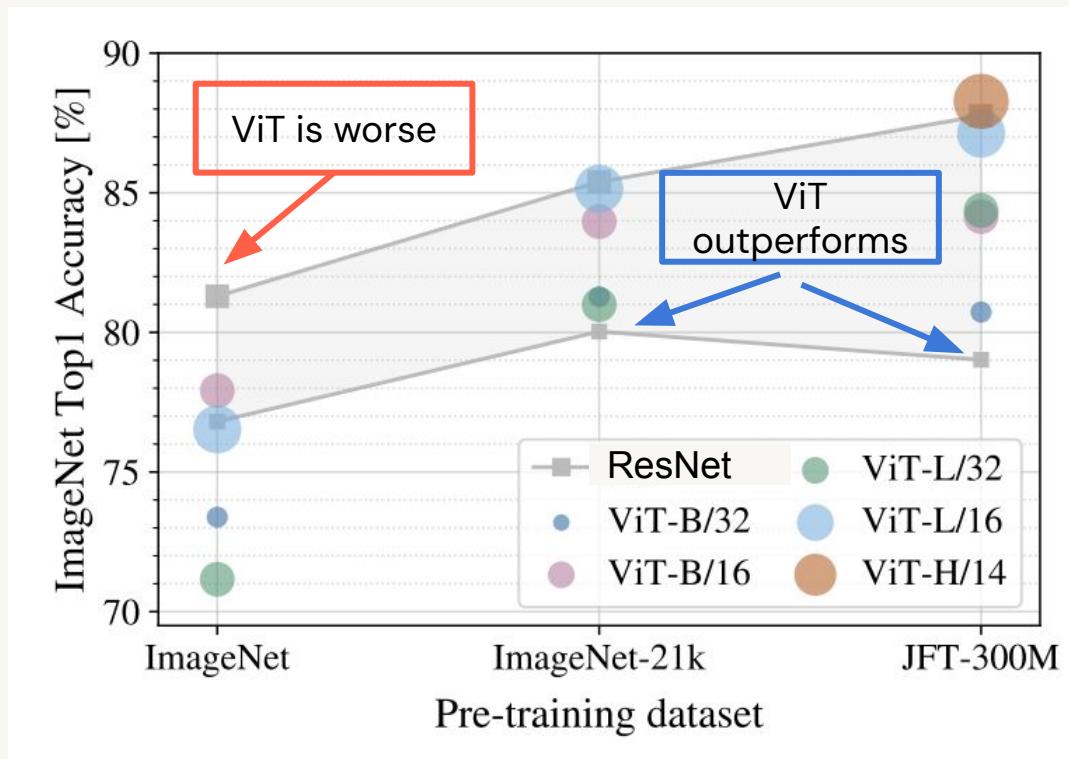
Linear project each patch to D-dimensional vector

N input patches
with shape of
 $3 \times \text{Pixel (P)} \times P$



ViT only outperforms ResNets on larger datasets

More computationally efficient than ResNet



- ViT is ~4 times faster than ResNet to train

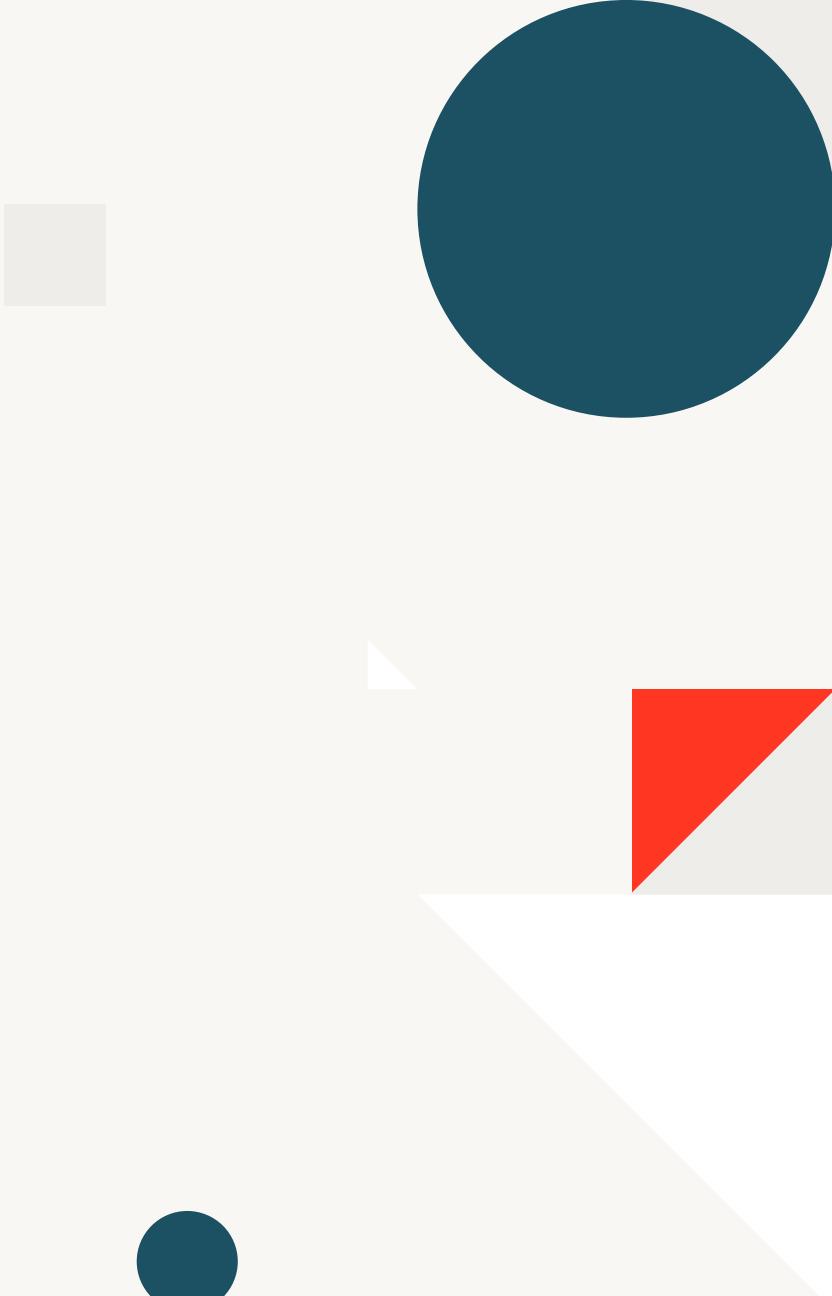
Source: [Dosovitsky et al 2021](#)

Many other vision-text models

Not necessarily revolutionary, but an evolution in computer vision research

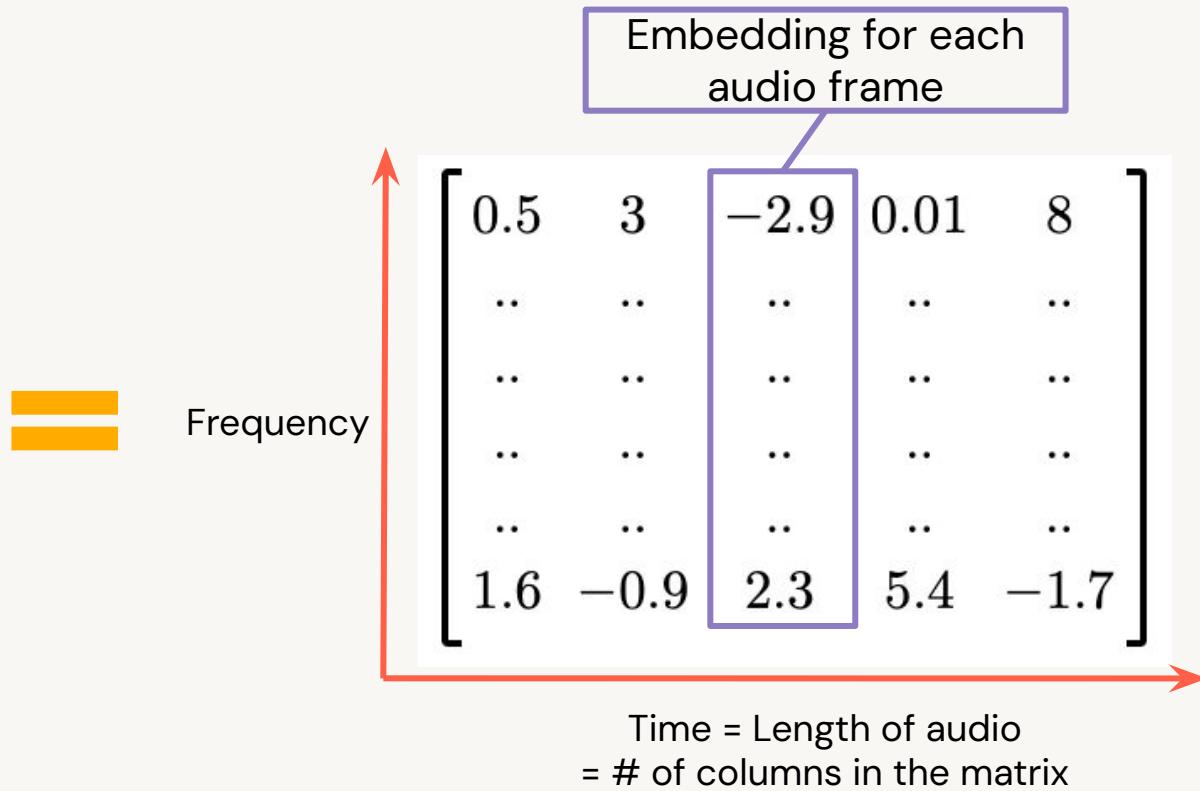
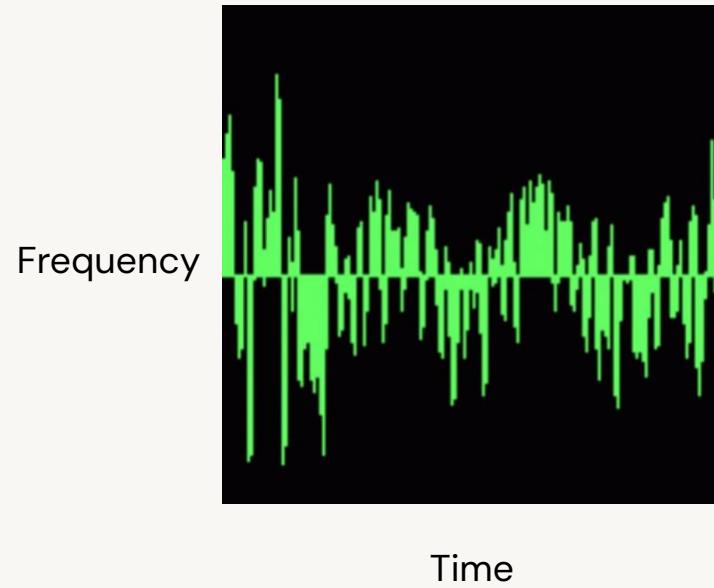


Audio



Audio signals are 2-dim spectrograms

We create embedding vectors for each t-min audio frame



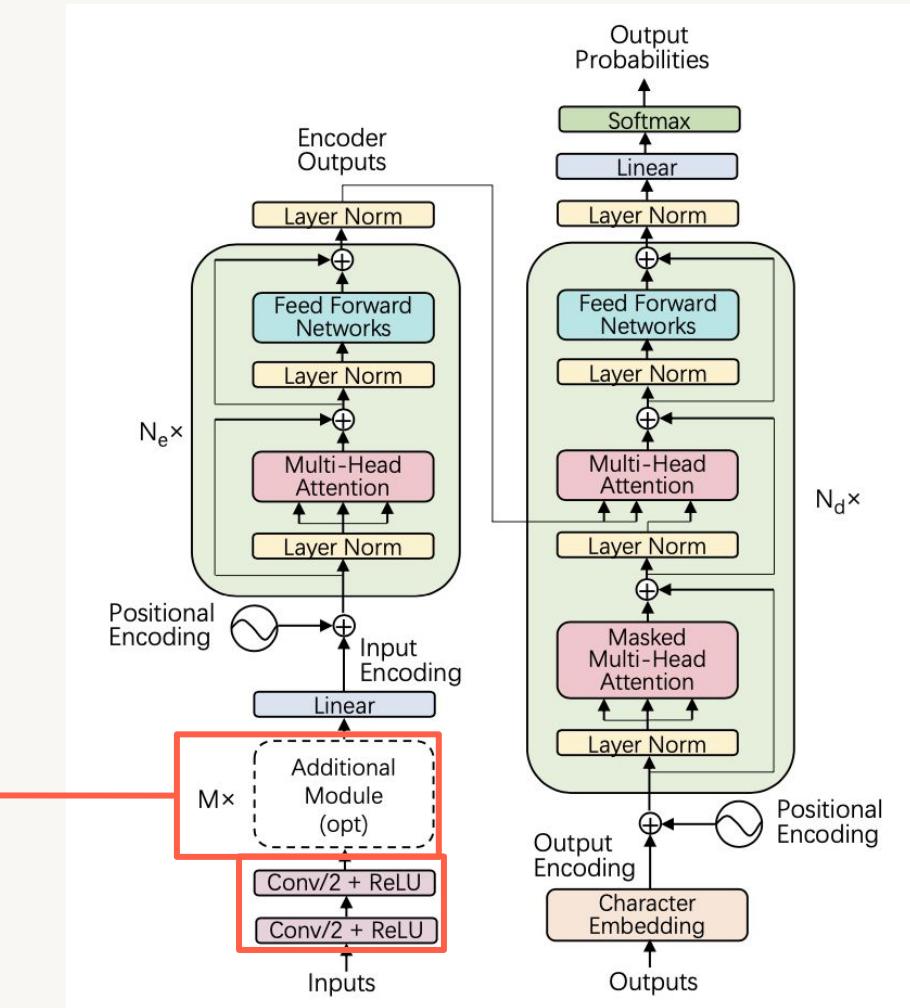
Audio is usually much longer than text length

Need to apply convolution layers with large strides to reduce dimensions

Speech Transformer (2018)

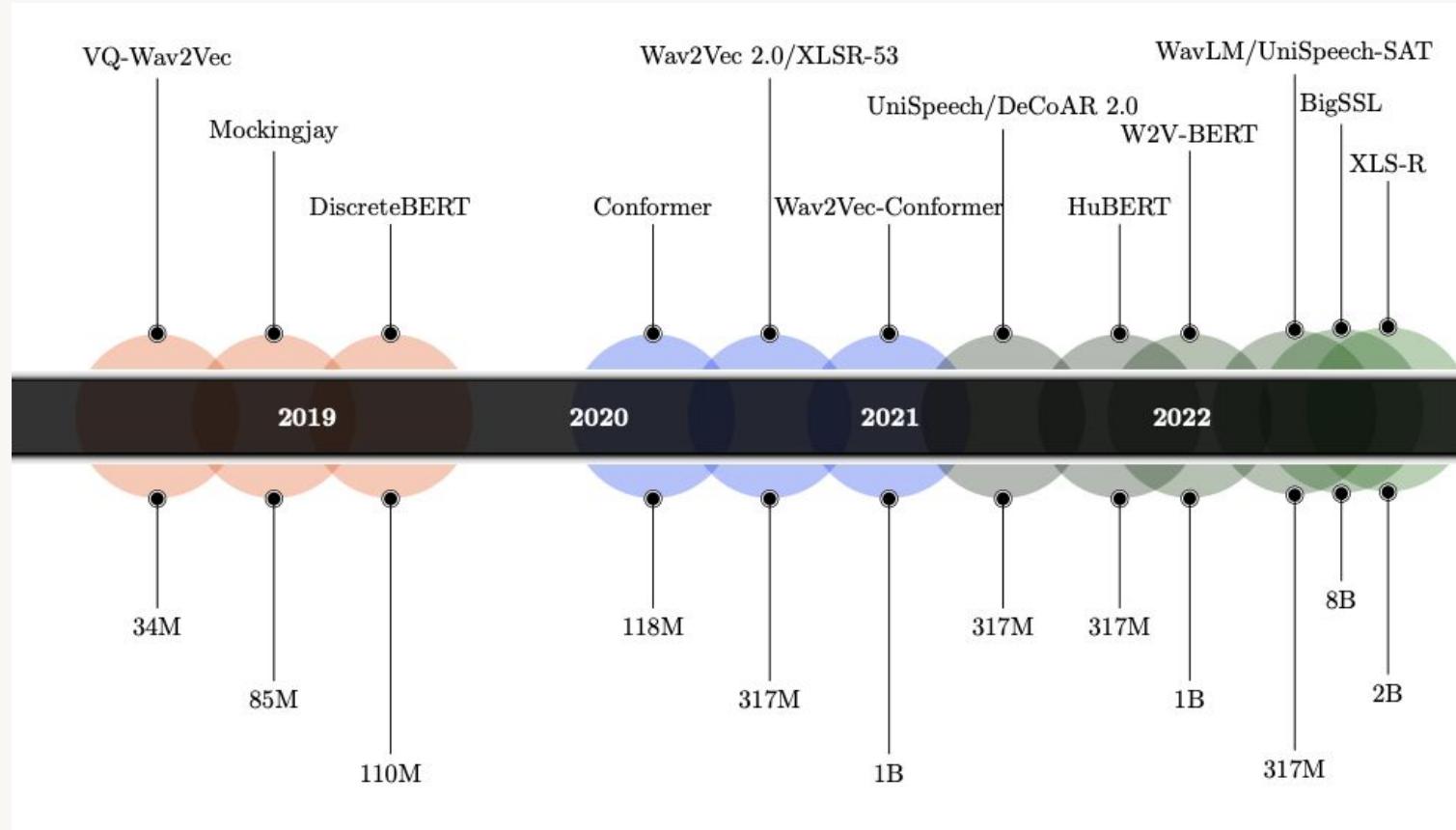
- Encoder-decoder
Transformer

Extract features using
different optional
modules: ResCNN,
ResCNNLSTM,e tc.



Few multi-modal advances

Also much harder: emotion, acoustics, tone, speed, speaker identification



- Most models focus on only text-speech, speech-text, or speech-speech
- Data2vec (MetaAI, 2022)
 - Self-supervised algorithm for speech, vision, and text
 - The only one so far?



Training Data for MLLMs



Hand-crafted training data

Text-audio or text-video data is much harder to procure



Female cop talking on walkietalkie, responding emergency call, crime prevention

Source: [WebVid](#)

Example of detailed video description



###Human: Portray the video with a rich, descriptive narrative, capturing its atmosphere and events.
###Assistant: First, we see a young woman with long brown hair wearing a backpack standing on the edge of a cliff, looking at the beautiful mountain scenery. She is in awe and stands there for a few seconds. Next, we see her turning around slowly, with her arms open wide. Then, we see her taking a few steps back, still facing the canyon, with her arms still open wide. She seems to be soaking in the grandeur of the view. Finally, she stands confidently at the edge of the cliff, her backpack still on her back, as if she has conquered the mountain. Throughout the video, we see the vast expanse of the canyon with the mountains in the background. The sky is overcast in some parts and clear blue in others, creating a beautiful contrast. Overall, the young woman seems to be thrilled with the view and enjoys every bit of it with open arms.

Source: [VideoChat](#)



Structured Scene Description (FAMOUS)

Focus: Man holding a chainsaw preparing to cut down the Christmas tree

Action: Man getting ready to cut down the tree

Mood: Anticipation

Objects: Man, chainsaw, trees in the forest, brown dirt

Setting: In the forest among trees with brown dirt on the ground

Dense Scene Description

In a dense forest, a man is preparing to cut down a Christmas tree, creating a mood of anticipation. The man, wearing appropriate protective gear, holds a chainsaw, ready to act. The scene is marked by the contrast between the excitement of upcoming Christmas celebrations and the focus of cutting down the tree. Key elements in the scene, including the man, the chainsaw, the trees, and the brown dirt, contribute to this poignant snapshot of the holiday season, highlighting the intersection of tradition and respect for nature.

Source: [Himakunthala et al 2023](#)



Instruction-tuned, hand-crafted data

```
[  
  {  
    "id": "000000033471",  
    "image": "000000033471.jpg",  
    "conversations": [  
      {  
        "from": "human",  
        "value": "<image>\nWhat are the colors of the bus in the  
        image?"  
      },  
      {  
        "from": "gpt",  
        "value": "The bus in the image is white and red."  
      },  
      {  
        "from": "human",  
        "value": "What feature can be seen on the back of the bus?"  
      },  
      {  
        "from": "gpt",  
        "value": "The back of the bus features an advertisement."  
      },  
      {  
        "from": "human",  
        "value": "Is the bus driving down the street or pulled off to  
        the side?"  
      },  
      {  
        "from": "gpt",  
        "value": "The bus is driving down the street, which is  
        crowded with people and other vehicles."  
      }  
    ]  
},  
]
```

conversations (list)	image (string)	id (string)
[{ "from": "human", "value": "Provide a brief description of the given image.\n<image>" }, { "from": "gpt", "value": "olive oil is a healthy ingredient used liberally ." }]	"GCC_train_002582585.jpg"	"GCC_train_002582585"
[{ "from": "human", "value": "<image>\nWrite ..."}]	"GCC_train_002429825.jpg"	"GCC_train_002429825"

Source: [CC-3M](#)

Source: [LLaVa](#)



Instruction-tuned, model-generated data

Actually: manually design examples first, then ask model to generate more

Context type 1: Captions

A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.



Context type 2: Boxes

person: [0.681, 0.242, 0.774, 0.694], person: [0.63, 0.222, 0.686, 0.516], person: [0.444, 0.233, 0.487, 0.34], backpack: [0.384, 0.696, 0.485, 0.914], backpack: [0.755, 0.413, 0.846, 0.692], suitcase: [0.758, 0.413, 0.845, 0.69], suitcase: [0.1, 0.497, 0.173, 0.579], bicycle: [0.282, 0.363, 0.327, 0.442], car: [0.786, 0.25, 0.848, 0.322], car: [0.783, 0.27, 0.827, 0.335], car: [0.86, 0.254, 0.891, 0.3], car: [0.261, 0.101, 0.787, 0.626]

Response type 1: conversation

Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV).
Question: Where is the vehicle parked?
Answer: The vehicle is parked in an underground parking area, likely in a public garage.
Question: What are the people in the image doing?
Answer: The people are trying to fit all of their luggage into the SUV, likely preparing for a trip.

Response type 2: detailed description

The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip.

In the parking area, various luggage items are scattered around the vehicle. There are two backpacks, one located near the left rear wheel and the other closer to the right side of the vehicle. Additionally, there are two suitcases, one on the right side of the car and another further away near the center of the parking area. A bicycle can also be seen on the left side of the vehicle.

Other cars are parked around the main SUV, with one car positioned behind it and slightly to the left, another behind and slightly to the right, and the third car further behind on the right side.

Captions

A man is skiing in the open snow covered hills
A skier is making his way into the snow.
A skier on the foothills of a huge mountain range.
A skier looks at mountains while standing near a trail sign.
a single person skiing by an area with a lot of bushes

Conversation

Question:
What is the skier doing?
====
Answer:
The skier is skiing in the snow-covered hills, making his way through the snow and enjoying the beautiful mountain scenery. They are also standing near a trail sign, which indicates their location on the mountain.
====
Question:
Are there any other people or skiers visible in the image?
====
Answer:
No, there is only a single person skiing in the image, and they are surrounded by an area with a lot of bushes.

Source: [Liu et al 2023](#)



LAION-5B: open source image-text data

Original data: [Common Crawl](#); filtered with OpenAI's CLIP model

The screenshot shows a Databricks dataset card for "laion2B-en". At the top, it displays the dataset name, a paw icon, a license badge for "cc-by-4.0", and a "like" count of 119. Below this, there are tabs for "Dataset card", "Files", and "Community" (with 6 members). The "Dataset Viewer" tab is selected, showing a dropdown menu set to "train (2.32B rows)". To the right, there are links for "Auto-converted to Parquet" and "API". The main area is a table viewer showing two rows of data. The columns are labeled: SAMPLE_ID (int64), URL (string), TEXT (string), HEIGHT (int32), and WIDTH (int32). The first row has a SAMPLE_ID of 2,641,080,021,034, a URL pointing to a Shopify product page for beach umbrellas, and a TEXT description about blue beach umbrellas and spiral notebooks. The second row has a SAMPLE_ID of 1,069,682,003,121, a URL pointing to a BMW blog post, and a TEXT description about a BMW M2 car.

SAMPLE_ID (int64)	URL (string)	TEXT (string)	HEIGHT (int32)	WIDTH (int32)
2,641,080,021,034	"https://cdn.shopify.com/s/files/1/0017/3621/2538/products/blue-beach-umbrellas-point-of-rocks-crescent-beach-siesta-key-shawn-mcloughlin_32d72f5b-5e55-42f9-bfcfd6fa8d239beb_300x300.jpg?v=1524171284"	"Blue Beach Umbrellas, Point Of Rocks, Crescent Beach, Siesta Key - Spiral Notebook"	231	300
1,069,682,003,121	"http://cdn.bmwblog.com/wp-content/uploads/2016/02/BMW-M2-M-Performance-...	"BMW-M2-M-Performance-Dekor-Long-Beach-Blue-05"	120	120

Disclaimer:

Contains mostly copyrighted images. LAION doesn't claim ownership.

Source: [LAION](#)



X-shot learning for MLLMs



Computer Vision



X-shot learning to the rescue?

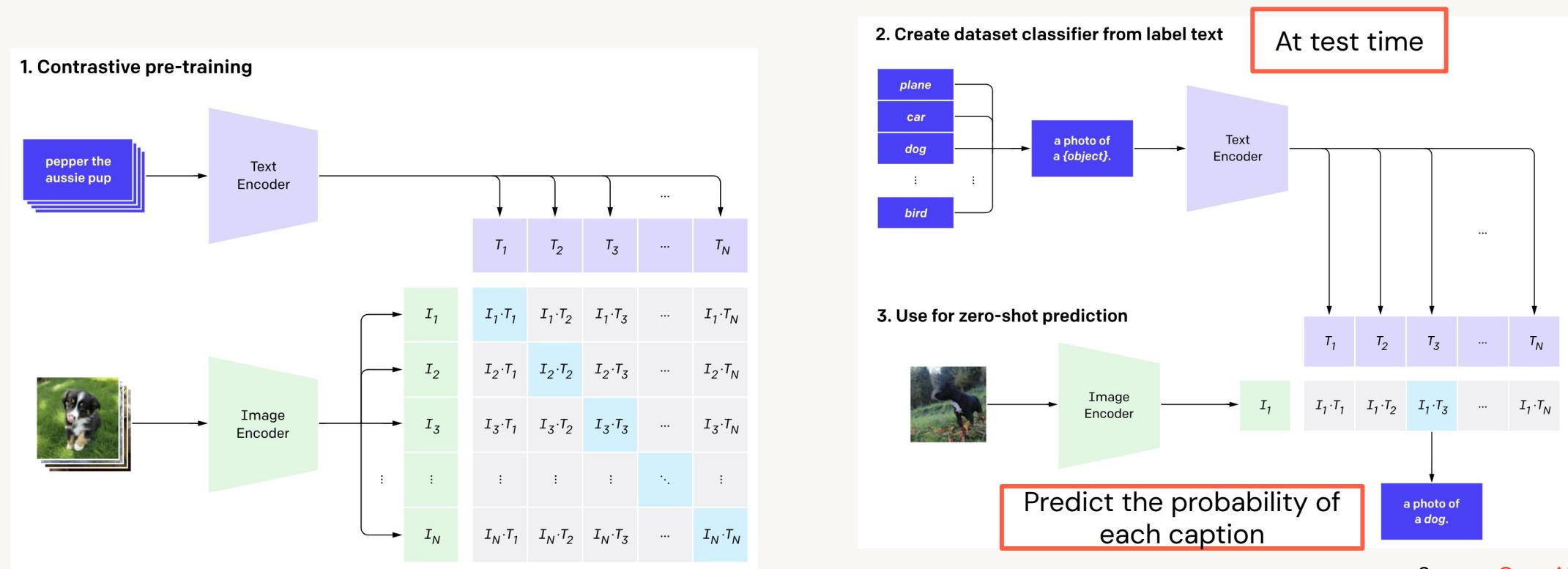
Gathering multi-modal data is harder than just images or text data



Zero-shot: Contrastive Language-Image Pairing

CLIP predicts which image-text pair actually occurs in the training data

- Collects 400M image-text pairs from the internet as training data



Source: [OpenAI](#)



CLIP performs better across settings



Big limitation:

- Inflexible
 - Cannot generate text

CLIP performs much better in non-ImageNet settings

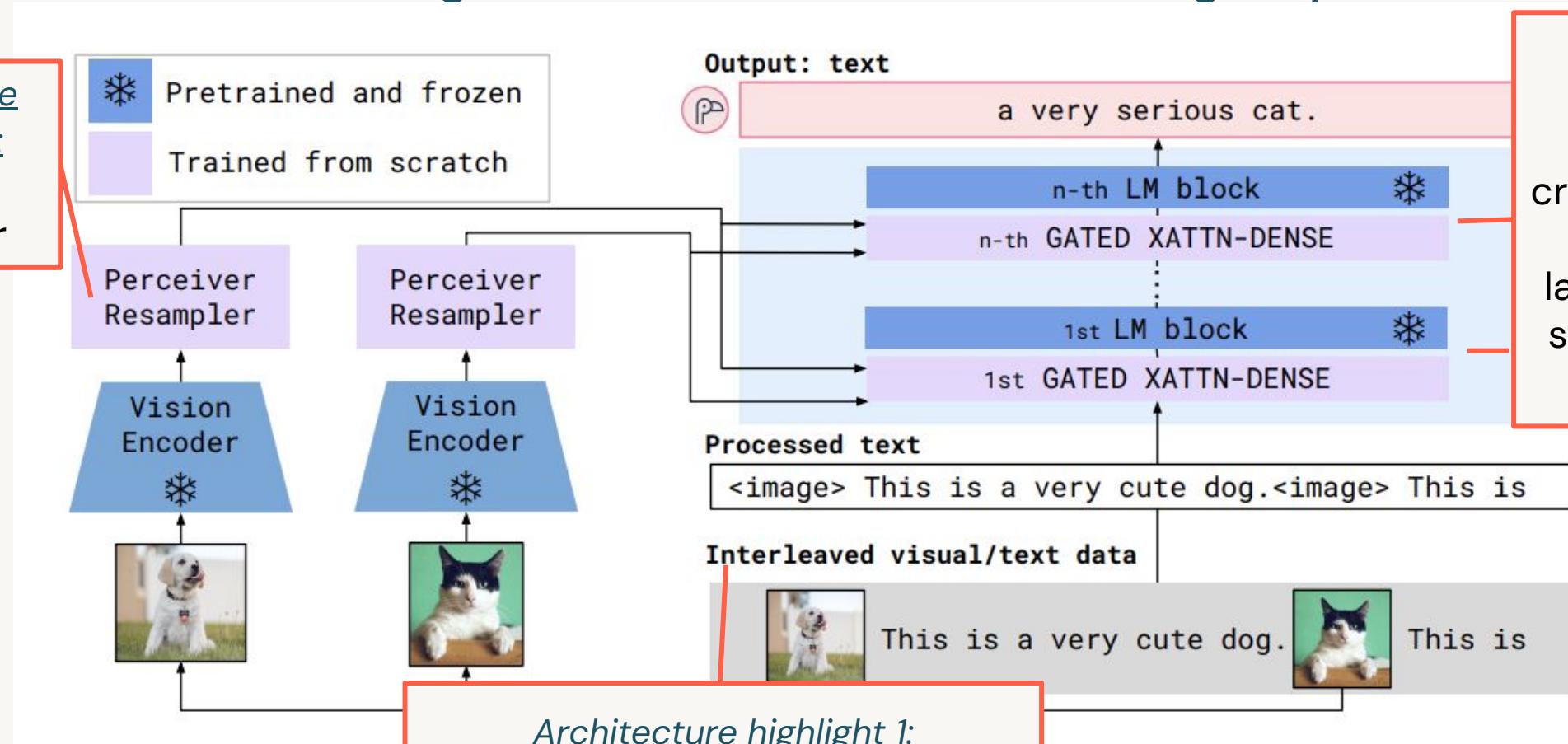
Source: [OpenAI](#)



Few-shot, in-context: Flamingo

Unifies treatment of high-dimensional video and image inputs

Architecture highlight 2:
Perceiver resampler

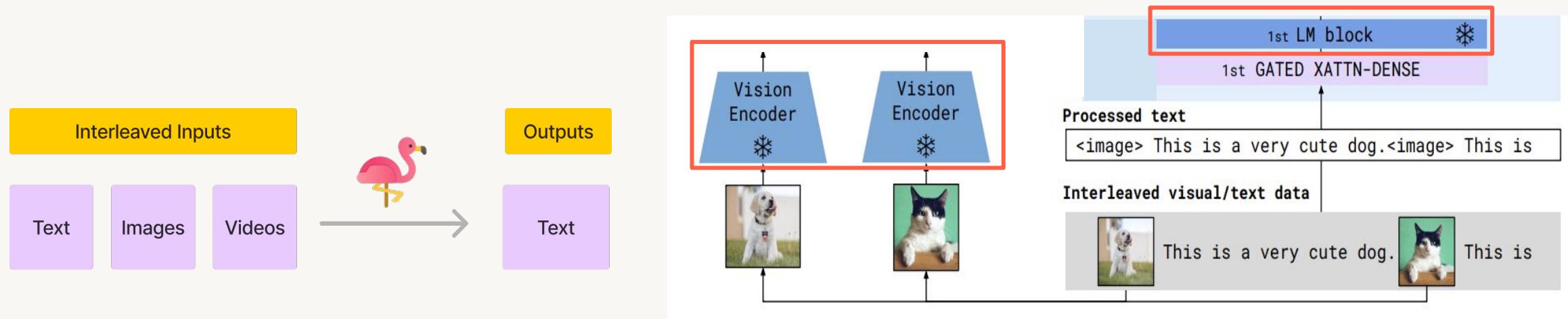


Architecture highlight 3:
Interleave cross-attention with language-only self-attention layers



Flamingo bridges vision and language models

Vision encoder similar to CLIP + Chinchilla (Language) accept interleaved inputs



Source: [Alayrac et al 2022](#)

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

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

p := flamingo model

y_l := l -th language token in input text

$y_{<l}$ - preceding text tokens

$x_{\leq l}$ - preceding image/videos

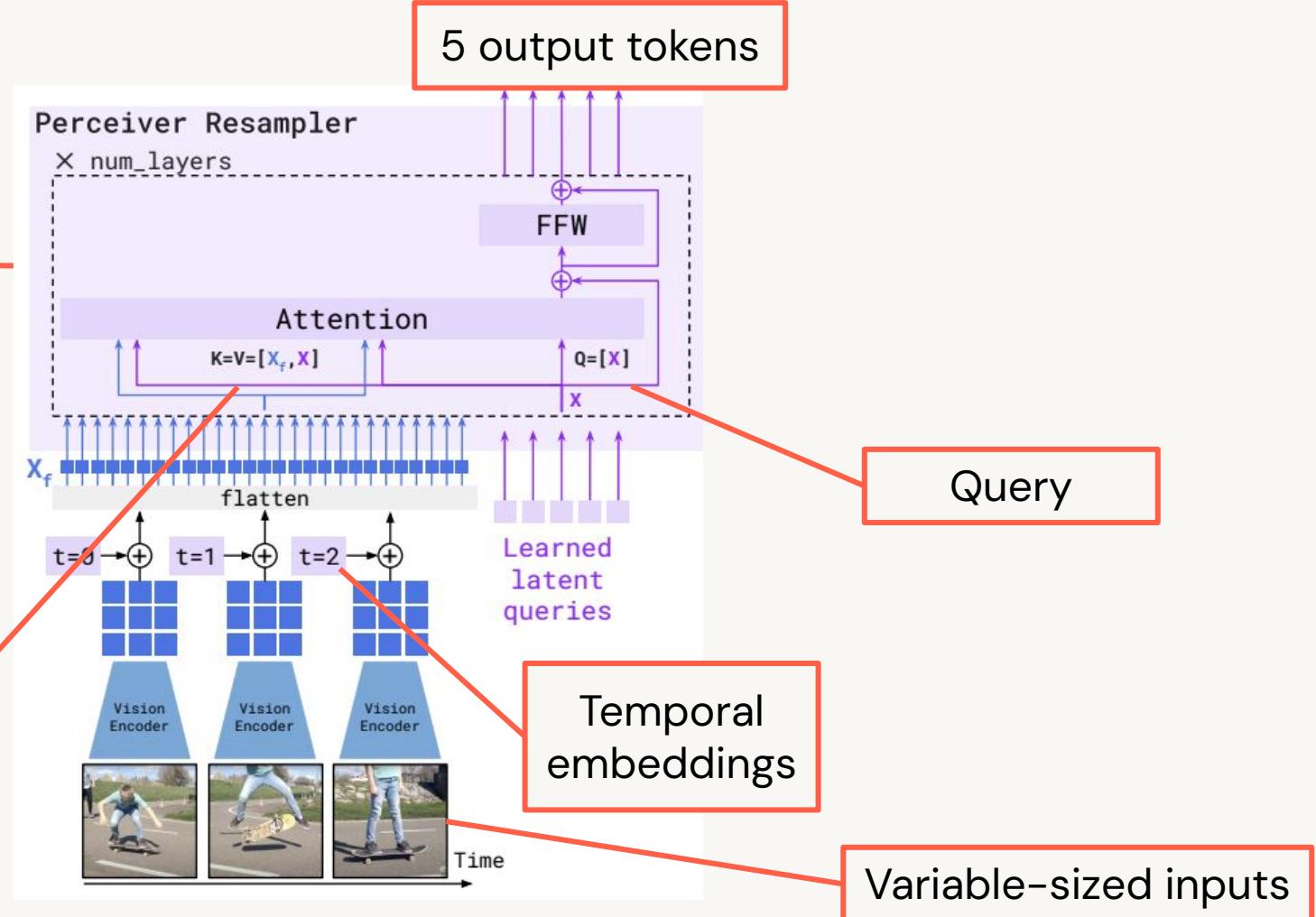
Multimodal likelihood

Image source: [Samuel Albanie](#)

Perceiver resampler outputs fixed-sized tokens

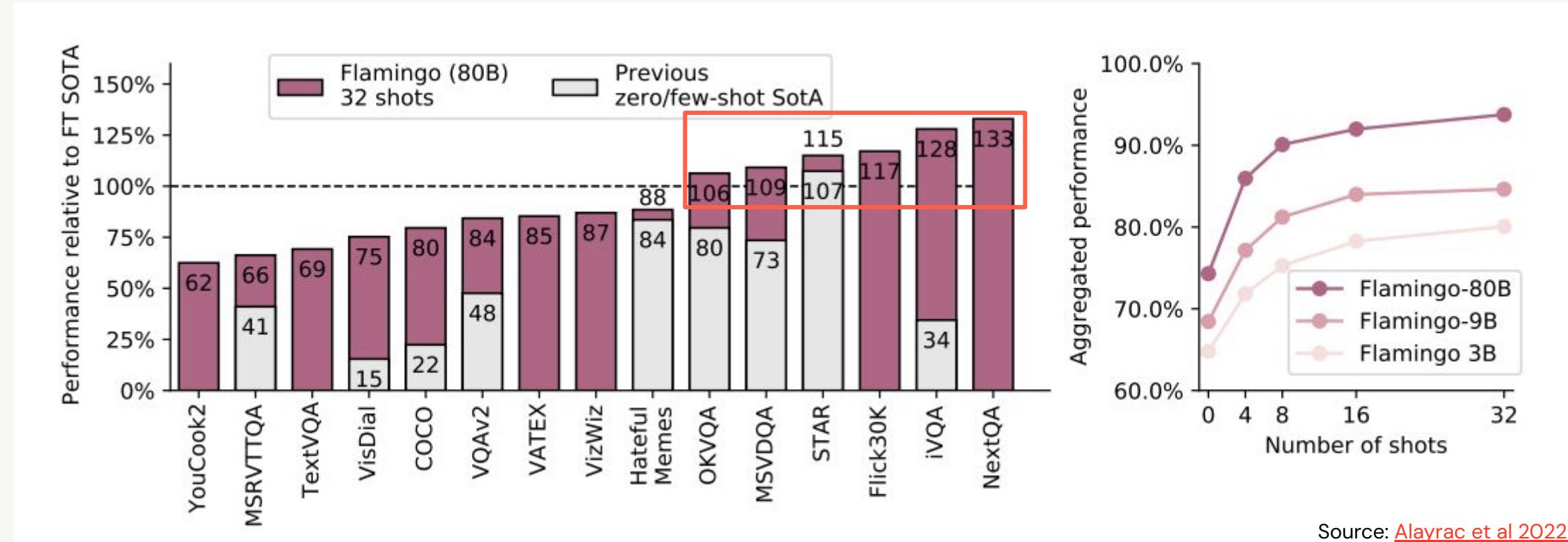
Maps variable-size grid of inputs to a **fixed** number of output tokens

Keys and values = spatio-temporal visual features



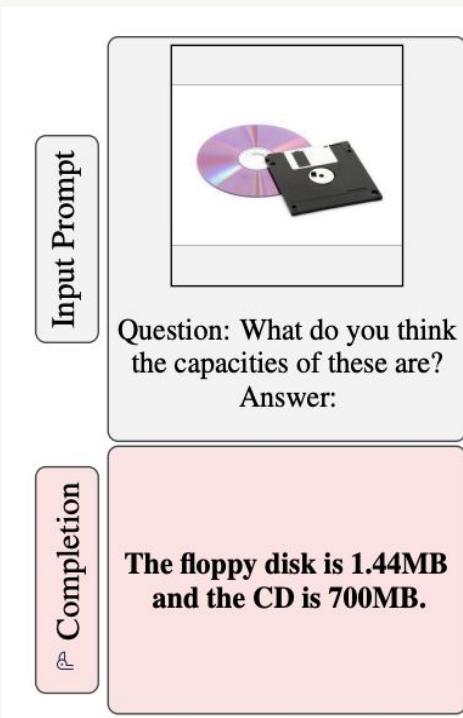
Outperforms 6 out of 16 SOTA fine-tuned models

Curated 3 high-quality datasets: LTIP (Long Text-Image Pairs), VTP (Video-Text Pairs), and MultiModal Massive Web (M3W)

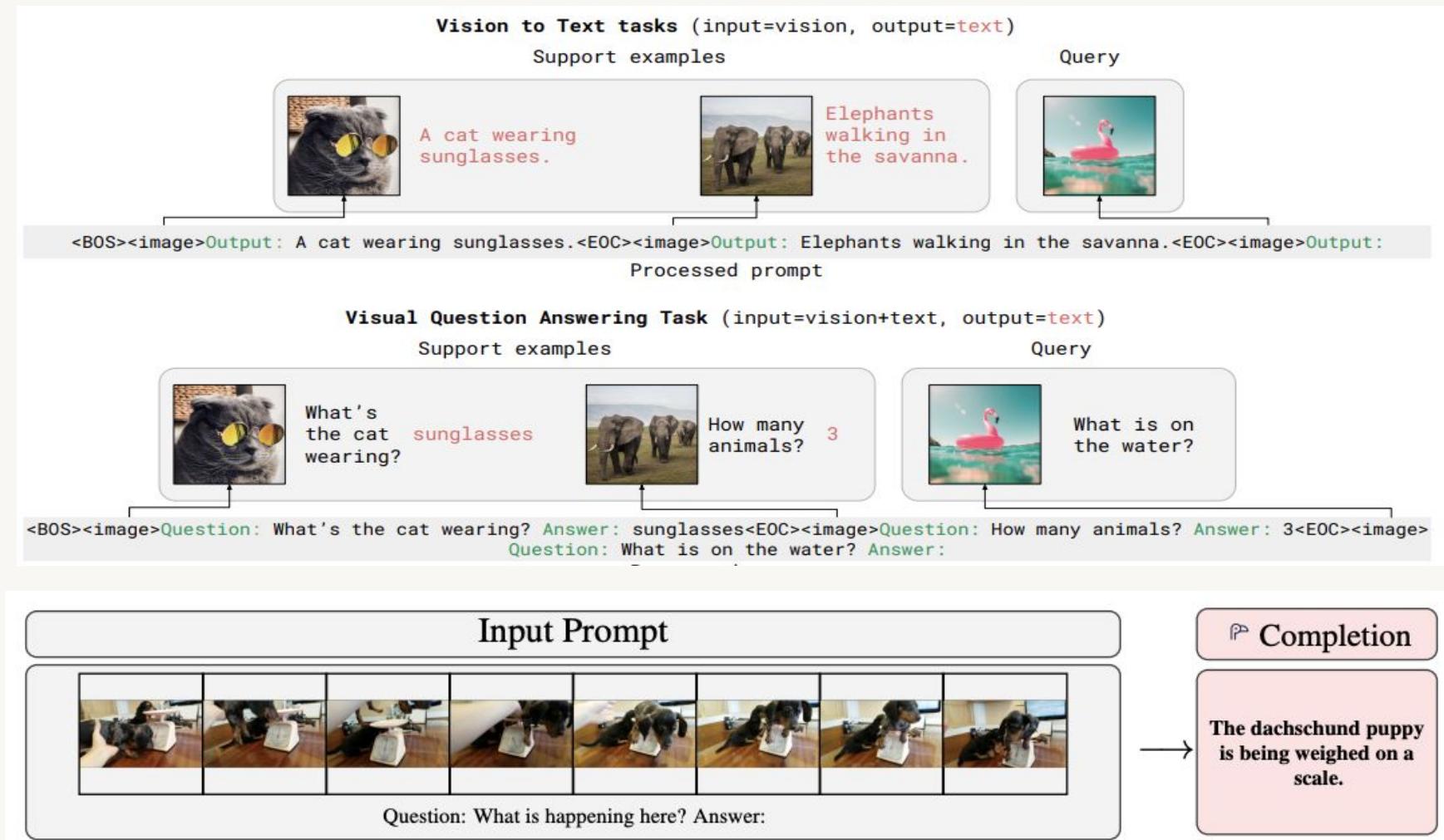


Qualitative inspection on selected samples

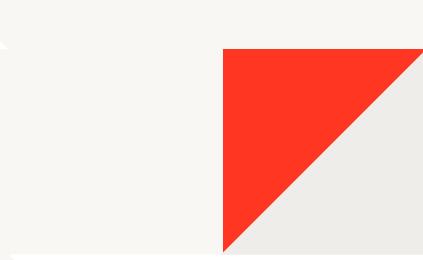
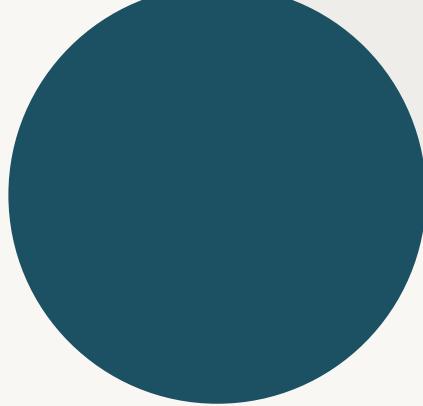
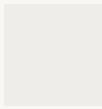
Supported input format: (image, text) or (video, text) + visual query



Source: [Alayrac et al 2022](#)

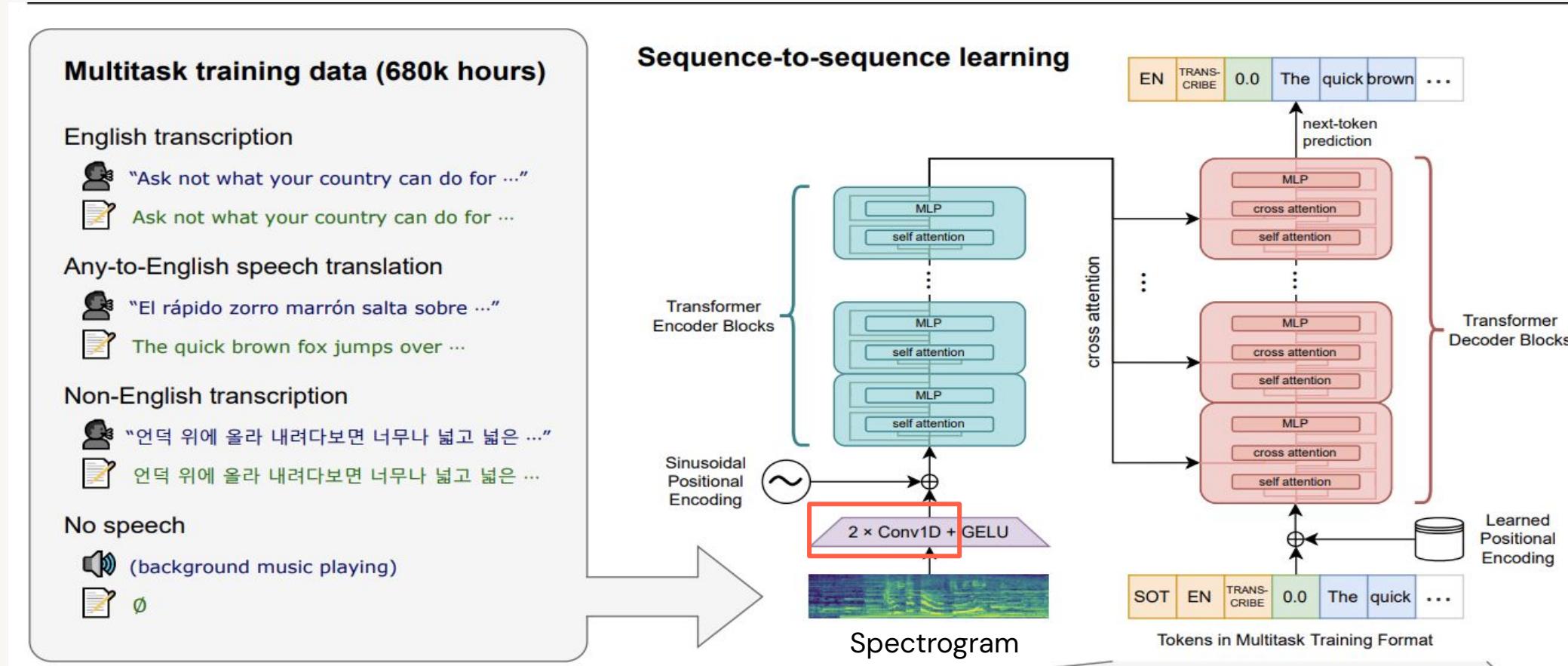


Audio



Zero-shot: OpenAI's Whisper

Encoder-decoder transformer: splits input audio into 30-second frames



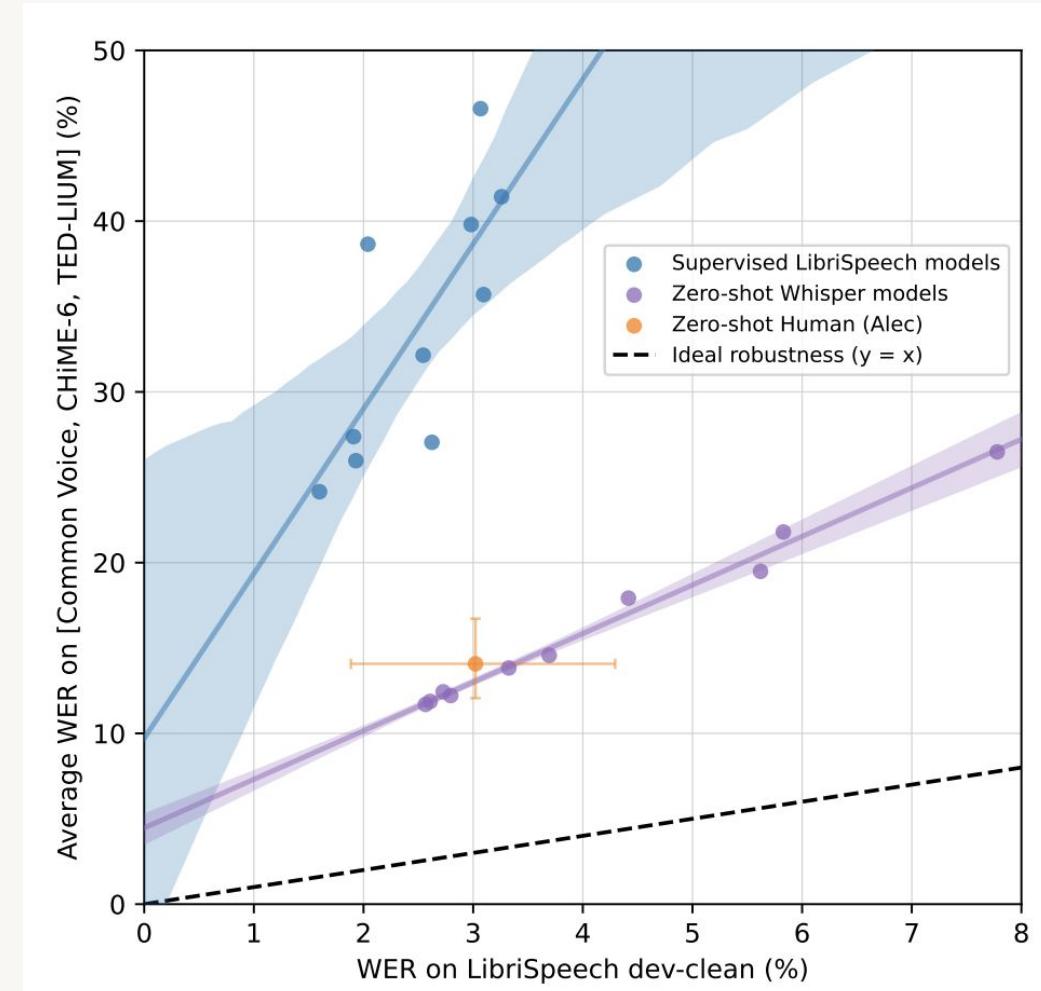
Source: [Radford et al 2022](#)



Whisper matches human robustness

Without fine-tuning on benchmark data

- WER = word error rate
- LibriSpeech
 - 1K hours of read English speech



Challenges

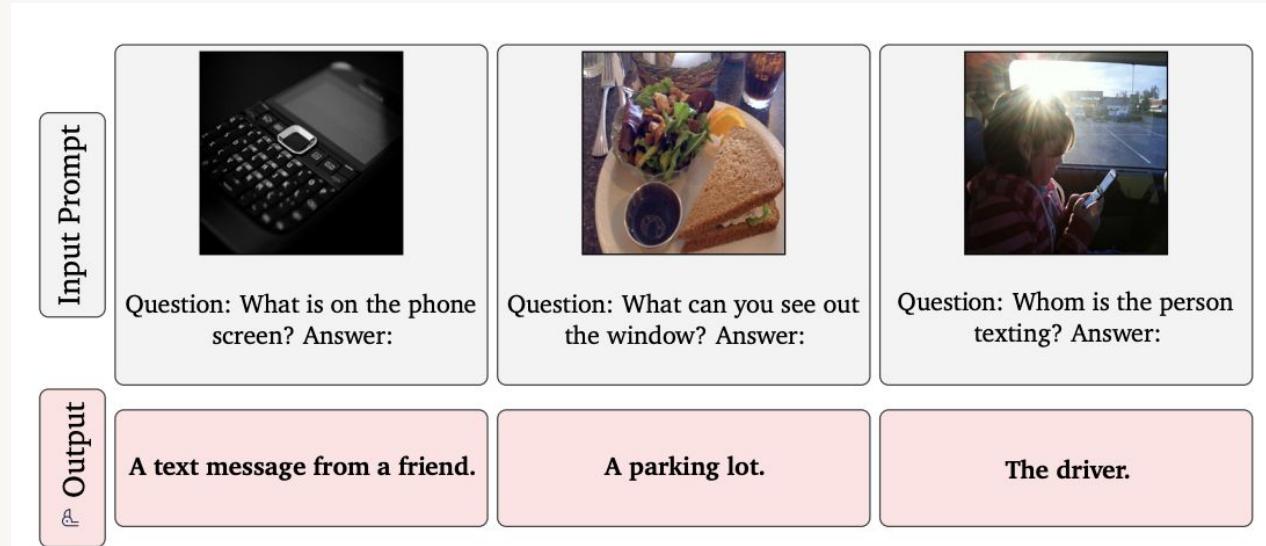
We haven't figured it all out yet



MLMs are not immune from LLM limitations

They inherit LLM risks

- Hallucination
- Prompt sensitivity
 - Context limit
- Inference compute cost
- Bias, toxicity, etc.
- Copyrights issues



Source: [Alayrac et al 2022](#) (Flamingo)

Reddit Wants to Get Paid for Helping to Teach Big A.I. Systems

The internet site has long been a forum for discussion on a huge variety of topics, and companies like Google and OpenAI have been using it in their A.I. projects.

Source: [New York Times, April 2023](#)

LAION-5B: A NEW ERA OF OPEN LARGE-SCALE MULTIMODAL DATASETS

Source: [laion.ai](#)

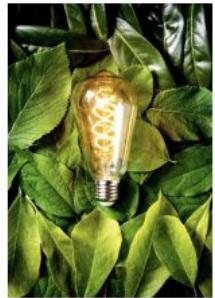


7:58 AM · Jul 14, 2023 · 5.9M Views

Source: [Next Shark, X Post](#)



MLMs can lack common sense (like LLMs)

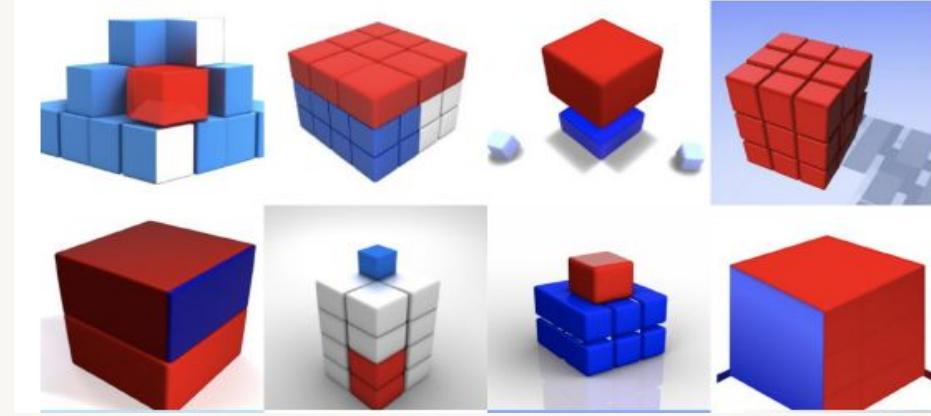


(a) some plants surrounding a lightbulb



(b) a lightbulb surrounding some plants

Source: [Thrush et al 2022](#)
Tested on many models, including CLIP and VisualBERT



Source: [Ramesh et al 2022](#)
Used CLIP and DALLE-2; Prompt: "a red cube on top of a blue cube"

GPT-3 completion

Input: You poured yourself a glass of cranberry, but then absentmindedly, you poured about a teaspoon of grape juice into it. It looks OK. You try sniffing it, but you have a bad cold, so you can't smell anything. You are very thirsty. So you

Completion: drink it. You are now dead.

Source: [Robust AI and NYU](#)

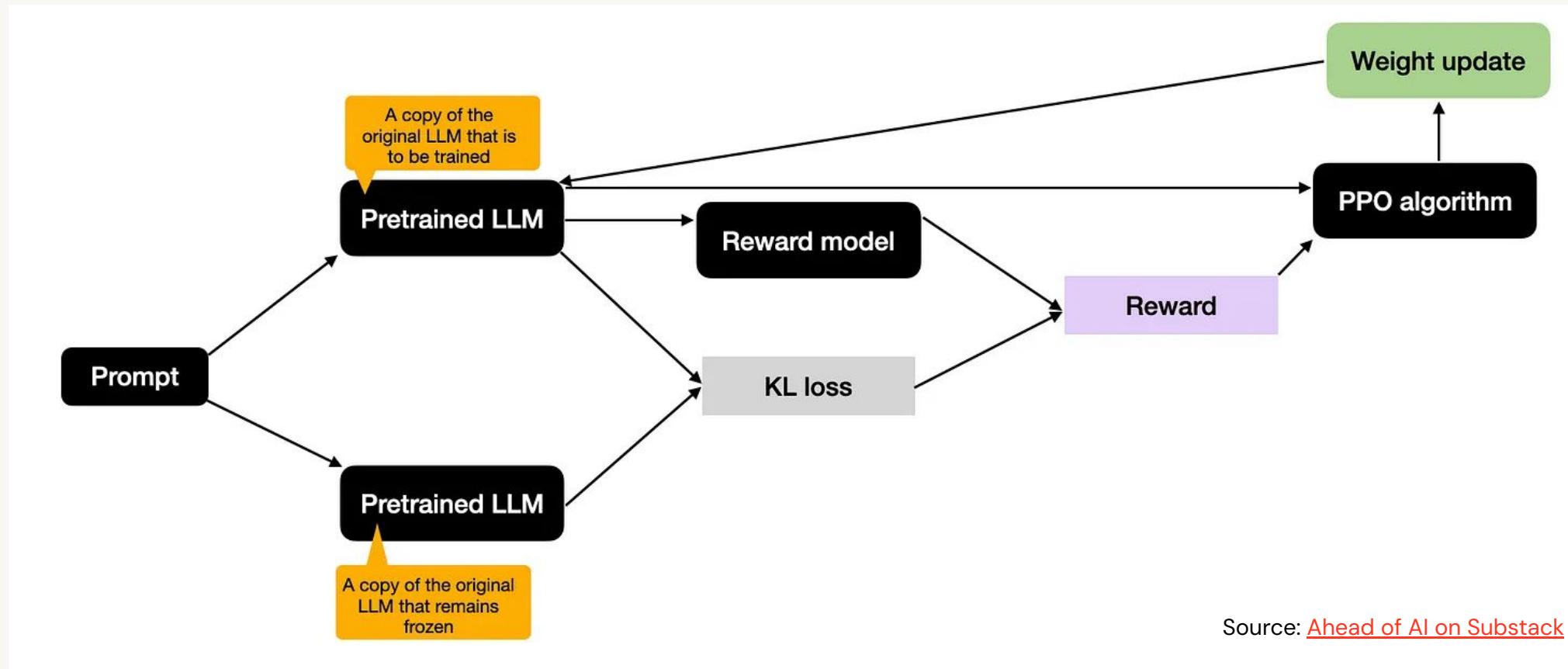


*Attention may not be forever
What may remain or rise?*



Reinforcement learning with human feedback

Human feedback trains the reward model (LM). KL loss ensures minimal divergence from the original LLM. Proximal Policy Optimization (PPO) updates the LLM.



Hyena Hierarchy

Convolutional neural networks are making a comeback?



- Good few-shot learners for languages
- Matches performances of Vision Transformers (ViT)

Table 4.6: Few-shot (3) accuracy (%) on SUPERGLUE tasks for small models.

Model	WSC	WIC	RTE	CB	MULTIRC	RECORD	BOOLQ	COPA	AVERAGE
GPTNeo (Black et al., 2021)	38.5	50.0	53.8	42.9	22.4	61.4	61.0	63.0	49.1
RWKV (Peng, 2021)	32.7	49.4	47.2	37.5	0.0	58.3	55.0	64.0	43.0
Hyena	39.4	50.1	47.6	46.4	26.7	58.1	56.0	70.0	49.3

Table 4.5: Zero-shot accuracy (%) on SUPERGLUE tasks for small models.

Model	WSC	WIC	RTE	CB	MULTIRC	RECoRD	BOOLQ	COPA	AVERAGE
GPTNeo (Black et al., 2021)	27.9	50.0	45.1	41.1	0.0	61.7	62.2	62.0	43.8
RWKV (Peng, 2021)	13.4	52.3	46.9	25.0	0.0	58.5	59.2	66.0	40.2
Hyena	21.2	50.5	46.6	39.3	1.1	59.4	51.8	70.0	41.5

Table 4.7: Image classification top-1 accuracy.

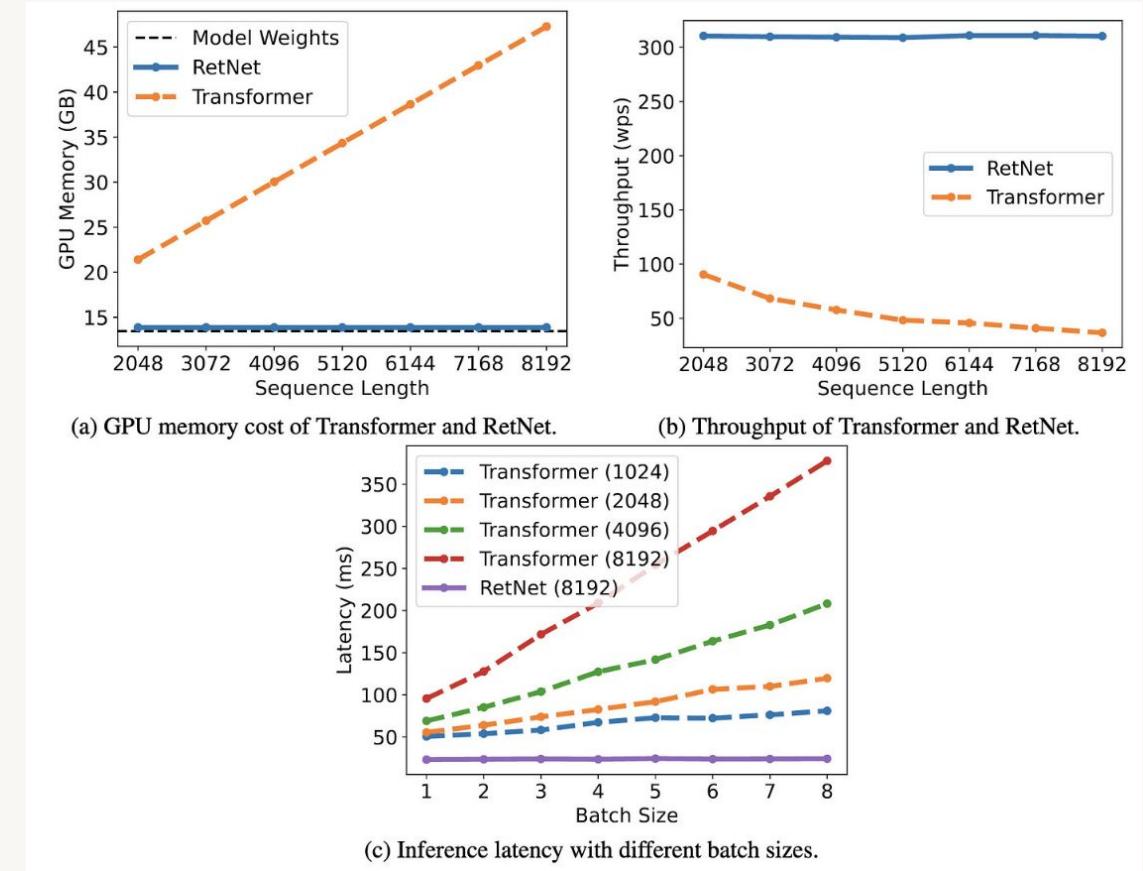
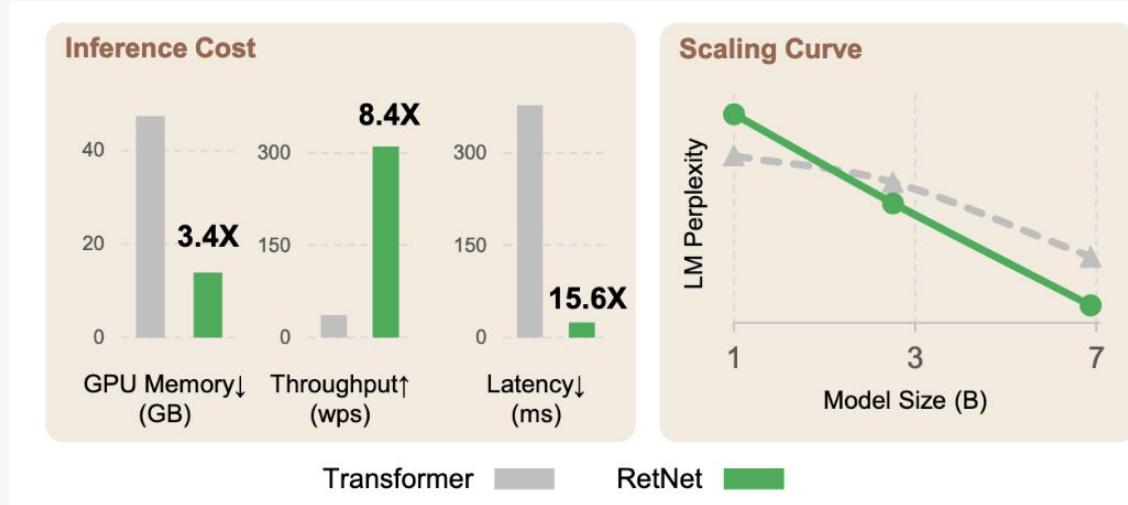
Model	PATCH SIZE	SEQ LEN	DATASET	ACC (%)
ViT (87M)	16x16	196	ImageNet-1k	78.5
Hyena-ViT (88M)	16x16	196	ImageNet-1k	78.5
ViT (87M)	8x8	1024	ImageNet-1k	80.0
Hyena-ViT (88M)	8x8	1024	ImageNet-1k	79.8
S4ND-ISO (268k)	-	-	CIFAR-10	89.9
Hyena-ISO (202k)	-	-	CIFAR-10	91.2

Source: [Poli et al 2023](#)



Retentive Networks

A new attention variant: a retention mechanism to connect recurrence and attention, without compromising performance



Source: [Sun et al 2023](#) (released in July)



Emerging applications

It's a great time to be alive



DreamFusion

Generates 3D objects from text caption



Source: [Poole et al 2022](#)

Make-a-Video

Generates video from text: "Cat watching TV with a remote in hand"



Source: [Singer et al 2022](#)



PaLM-E-bot

Robotics application: “bring me the rice chips from the drawer”



Source: [Driess et al 2023](#)



AlphaCode: generate code

Problem: Minimum # of minutes to makes pizzas of N slices

Human solution

```
t = int(input())
while t != 0:
    n = int(input())
    if n <= 6:
        print("15")
    else:
        print(((n+1)//2)*5)
    t -= 1
```

Model solution

```
import sys
input = sys.stdin.readline

def main():
    t = int(input())
    for _ in range(t):
        N = int(input())

        ans = 10 ** 100
        for a in range(0, 101):
            for b in range(0, 101):
                if N - (a * 6 + b * 8) <= 0:
                    ans = min(ans, a * 15 + b * 20)
                    break
                else:
                    c = (N - (a * 6 + b * 8)) // 10

                    if (N - (a * 6 + b * 8)) % 10 != 0:
                        c += 1

                    ans = min(ans, a * 15 + b * 20 + c * 25)

    print(ans)
```

```
if __name__ == '__main__':
    main()
```

Source: [Li et al 2022](#)



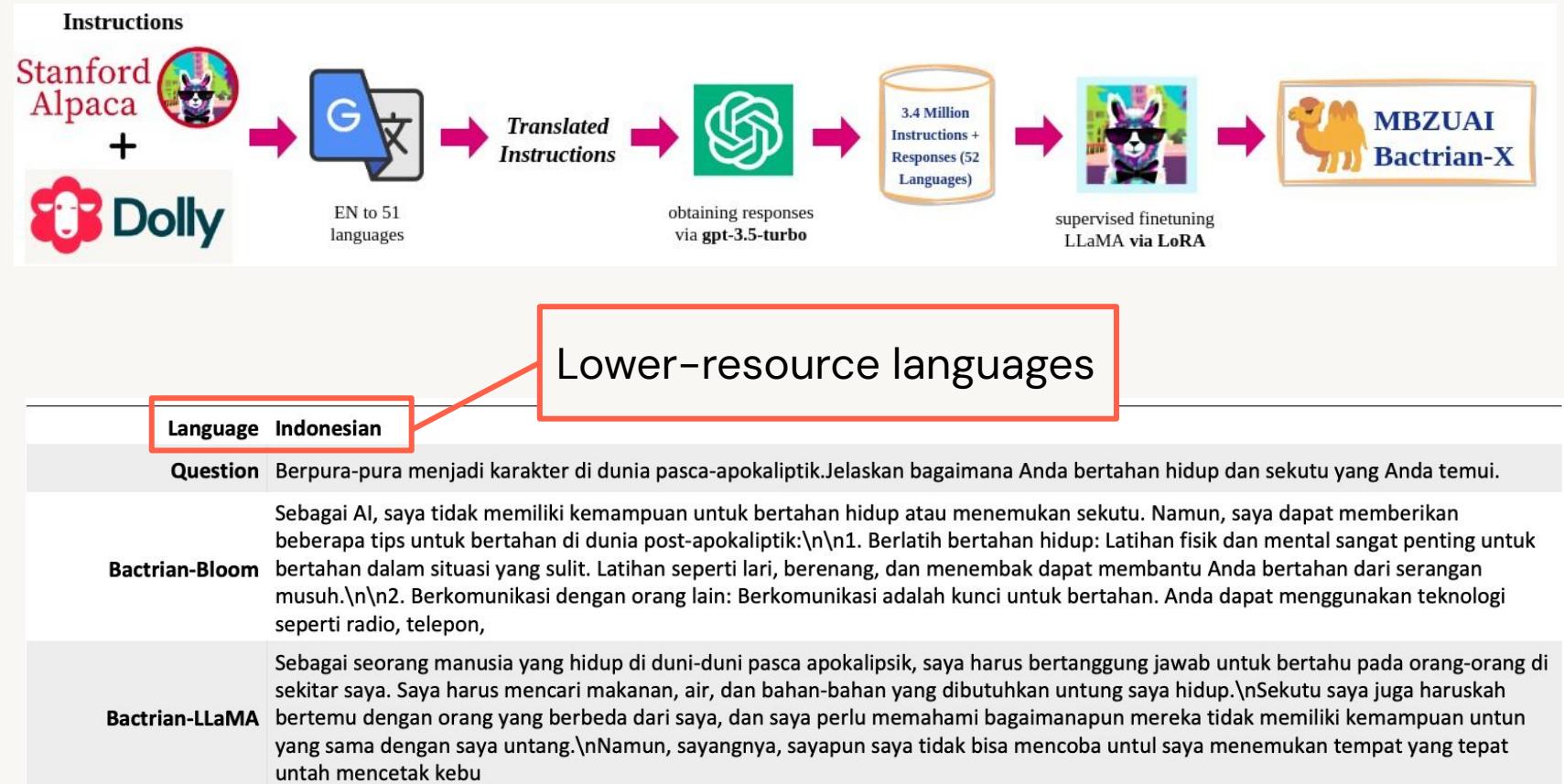
Multi-lingual models: Bactrian-X

An instruction-following model

Prompt:

"It's an
apocalypse.

Describe how you
survive and make
allies"



Source: Li et al 2023 ([GitHub](#) and [Paper](#)); released in May



Gato: a generalist AI agent

Can play Atari, caption images, chat, stack blocks with a real robot arm



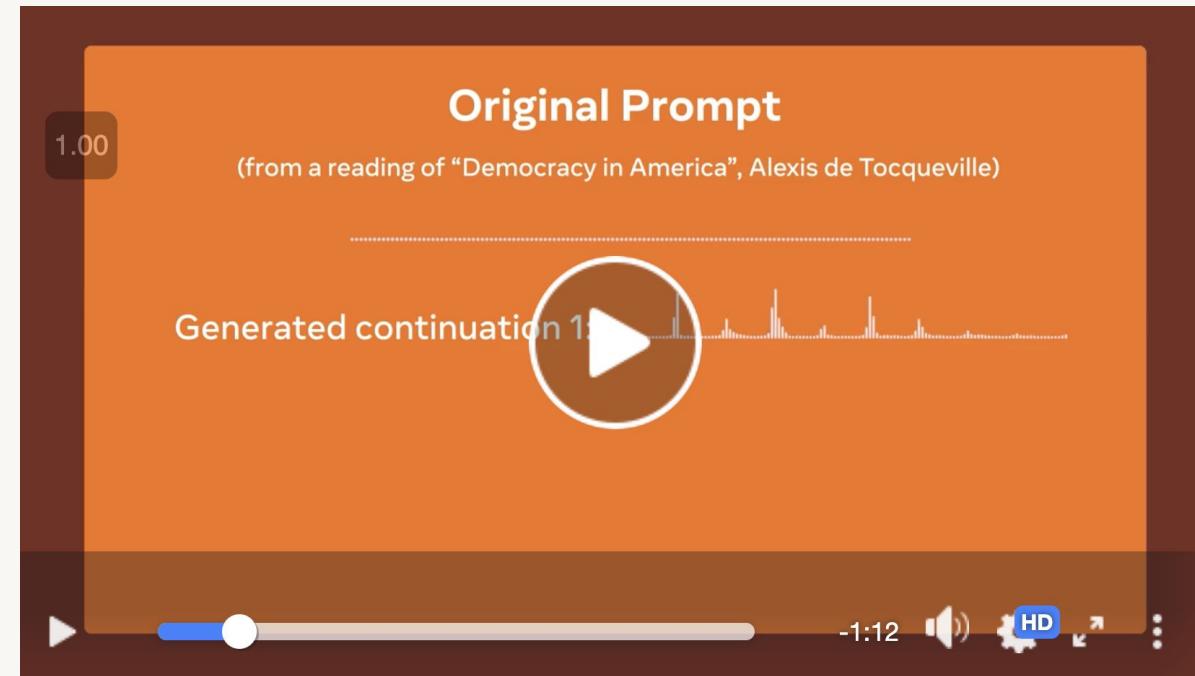
Source: [Reed et al 2022](#)



Textless NLP

Generate speech from raw audio without text transcription

Helps with low-resource languages

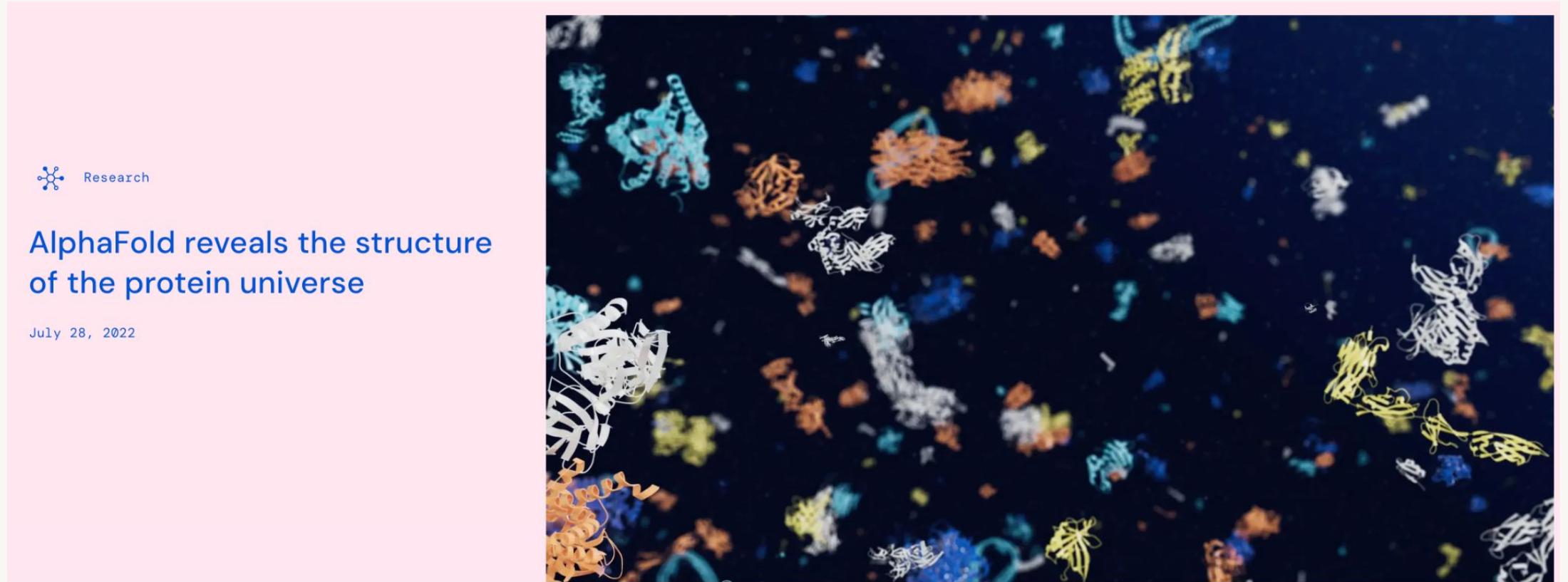


[Click here for demo on Meta AI](#)



AlphaFold

Uses attention to predict protein folding



AlphaFold reveals the structure
of the protein universe

July 28, 2022

Source: [Google DeepMind](#) and [Timeline of Breakthrough](#)



Module Summary

Multi-modal LLMs – What have we learned?

- MLLMs are gaining traction
- Transformers are general sequence-processing architectures that can accept non-text sequences
- MLLMs inherit limitations from LLMs
- Transformers may not be the last architecture standing
- More exciting and unimaginable MLLM applications are on the horizon



Time for some code!





databricks

