

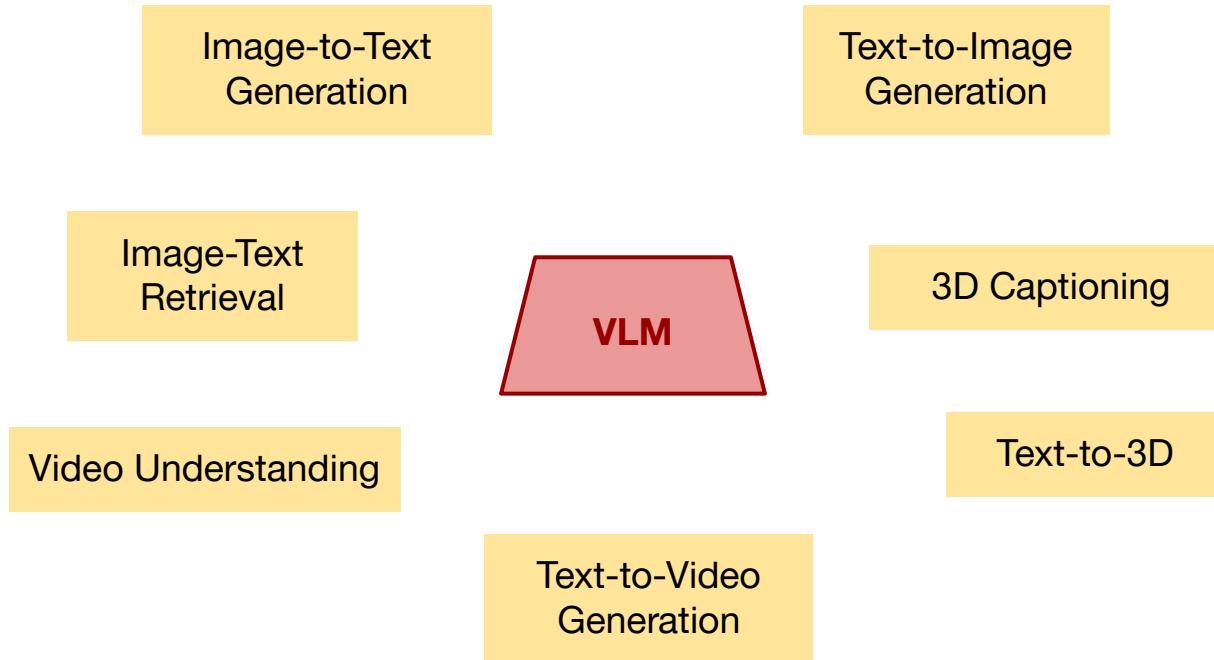
# Multimodal Language Models

**10-707**

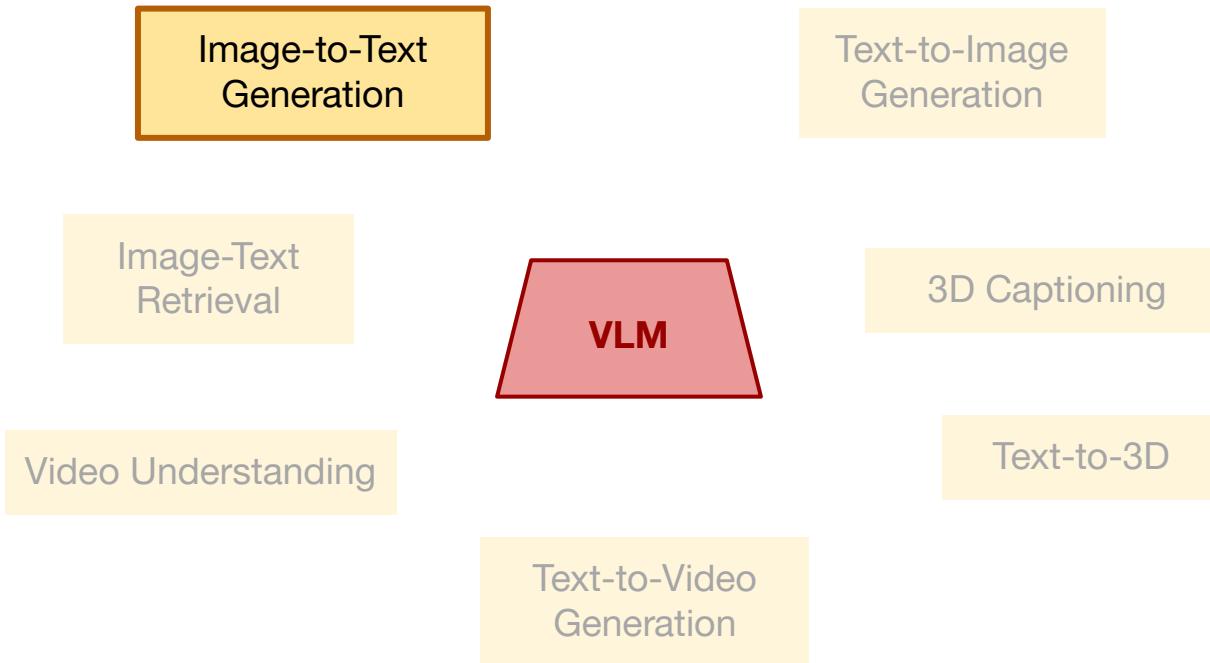
Jing Yu Koh

Carnegie  
Mellon  
University

# Vision-Language Tasks

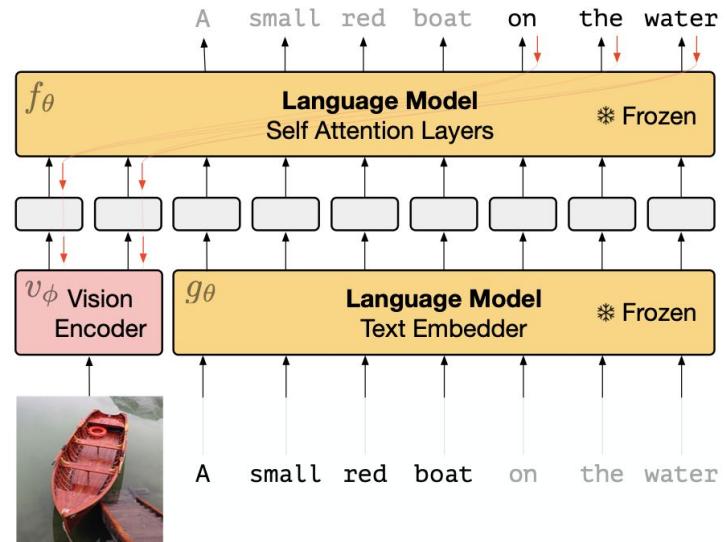


# Vision-Language Tasks

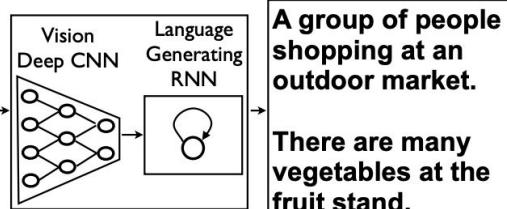


# Vision-Language Models (VLMs)

- **Image + text inputs → text outputs**
  - Much more general than text-only LMs
- **Early on (2014 - 2020):** Finetune on paired image-text datasets for specific tasks
- **More recently:** Map a strong vision model to an LLM
  - Use prompting rather than finetuning
  - Leverage the abilities of pretrained LLMs



# Image-to-Text Tasks



## Image Captioning

Show and Tell: A Neural Image Caption Generator ([Vinyals et al., 2014](#))  
VQA: Visual Question Answering ([Agrawal et al., 2016](#))



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



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



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



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

## VQA: Visual Question Answering

# Image-to-Text Tasks



## Visual Dialog

Q: What is the gender of the one in the white shirt ?

A: She is a woman

Q: What is she doing ?

A: Playing a Wii game

Q: Is that a man to her right ?

A: No, it's a woman

## Visual Dialogue



What does it say near the star on the tail of the plane?

Ground Truth      Prediction

jet      nothing

(a)



What is the time on bottom middle phone?

Ground Truth      Prediction

15:20      12:00

(b)



What is the top oz?

Ground Truth      Prediction

16      red

(c)



What is the largest denomination on table?

Ground Truth      Prediction

500      unknown

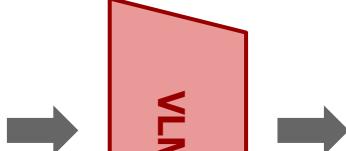
(d)

## TextVQA

# Image-to-Text Applications: Image Identification

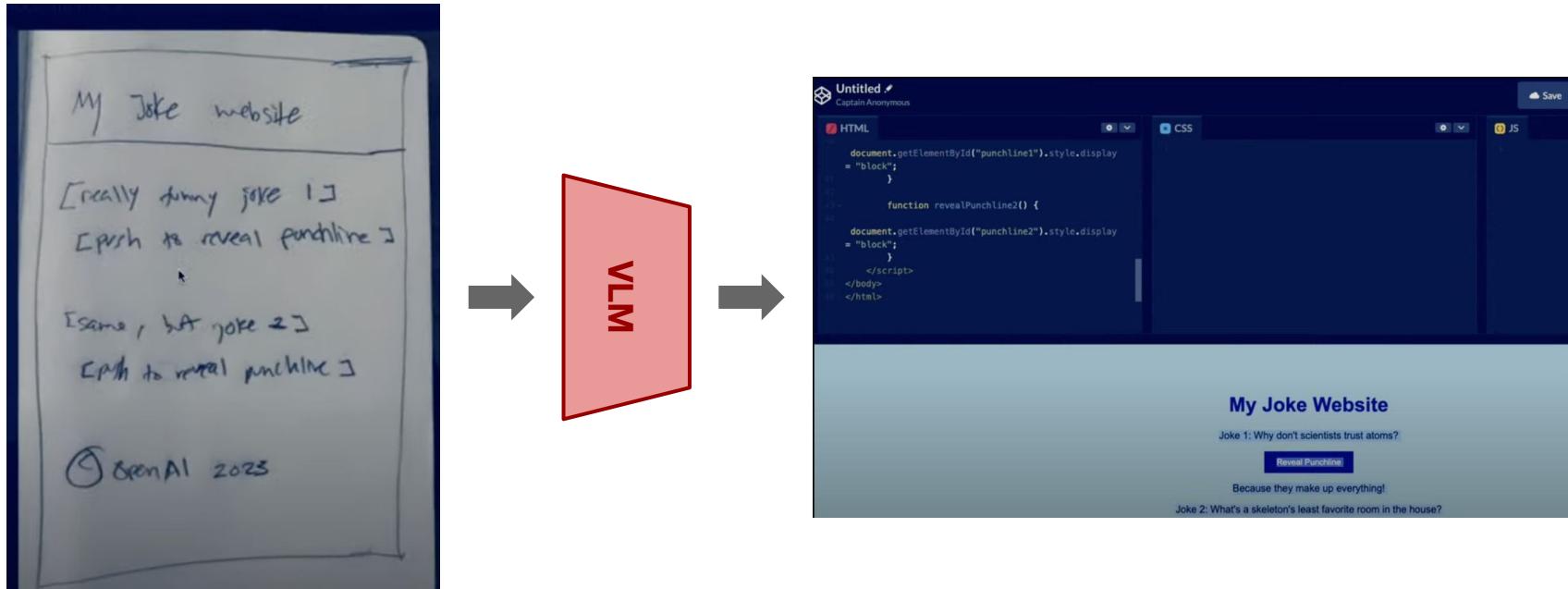


Q: Please  
can you tell  
me what  
this item is?



A: butternut  
squash red  
pepper soup

# Image-to-Text Applications: Sketch-to-Website



# Image-to-Text Applications: Image-to-LaTeX

$$\zeta(s) = \prod_{p \text{ prime}} \frac{1}{1 - p^{-s}} = \frac{1}{1 - 2^{-s}} \cdot \frac{1}{1 - 3^{-s}} \cdot \frac{1}{1 - 5^{-s}} \cdot \frac{1}{1 - 7^{-s}} \cdots$$



Can you  
write this in  
LaTeX?



```
$\zeta(s) =  
\prod_{p \text{ prime}} \text{prime}  
\frac{1}{1-p^{-s}} =  
\frac{1}{1-2^{-s}} \cdot  
\frac{1}{1-3^{-s}} \cdot  
\frac{1}{1-5^{-s}} \cdot  
\frac{1}{1-7^{-s}} \cdots
```

# Image-to-Text Applications: Image-to-LaTeX

LaTeX formatting for Riemann zeta function Euler product formula ✓

$$\zeta(s) = \prod_{p \text{ prime}} \frac{1}{1 - p^{-s}} = \frac{1}{1 - 2^{-s}} \cdot \frac{1}{1 - 3^{-s}} \cdot \frac{1}{1 - 5^{-s}} \cdot \frac{1}{1 - 7^{-s}} \cdots$$

JY

Can you write this in LaTeX?

Here is the LaTeX code for the given mathematical expression:

```
\zeta(s) = \prod_{p \text{ prime}} \frac{1}{1 - p^{-s}} = \frac{1}{1 - 2^{-s}} \cdot \frac{1}{1 - 3^{-s}} \cdot \frac{1}{1 - 5^{-s}} \cdot \frac{1}{1 - 7^{-s}} \cdots
```

This expression represents the Euler product formula for the Riemann zeta function, where the product is taken over all prime numbers  $p$ . Each factor in the product is of the form  $(1 - p^{-s})^{-1}$ , where  $s$  is a complex variable.

Copy Retry

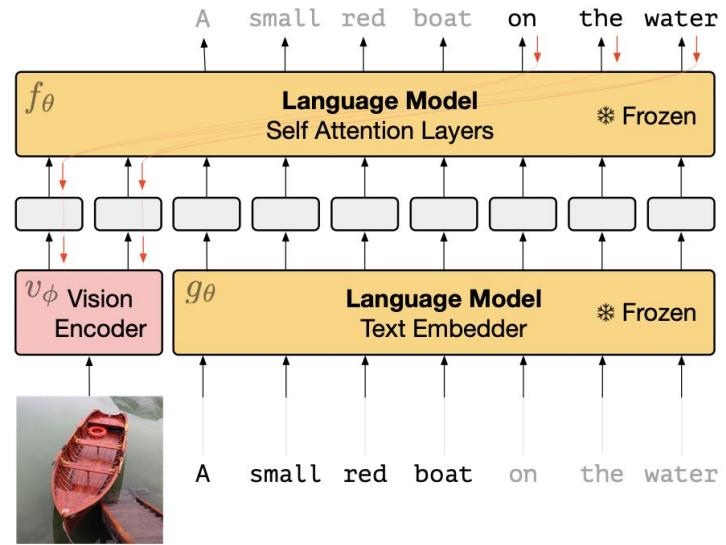


Claude can make mistakes. Please double-check responses.

# **How do you train VLMs?**

# Training Multimodal LMs

- Image + text inputs → text outputs
  - Very general!
- **Early on (2014 - 2020):** Finetune on paired image-text datasets to target specific tasks (e.g., VQA)
- **More recently:** Staple a pretrained vision model to a pretrained LLM



# LLMs are impressive general models

## Chain of Thought Prompting

### Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls.  $5 + 6 = 11$ . The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

### Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had  $23 - 20 = 3$ . They bought 6 more apples, so they have  $3 + 6 = 9$ . The answer is 9. ✓

### (d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: Let's think step by step.

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

I'm working on a new character for a game of mine, and I need him to have the most ridiculous backstory ever. Like, literally so ridiculous that it arguably doesn't make sense. I need it to include the following themes: royalty, abandonment, hell dimensions, seagulls, loss of limb, and enlightenment. Can you think of a ridiculous backstory for him?

Sure, here's an idea for a ridiculous backstory for your character:

Your character was once a royal prince, born into a wealthy and powerful family. He lived a life of luxury and privilege, surrounded by opulence and excess. But one day, everything changed. His parents abandoned him, casting him out of the royal court and leaving him to fend for himself on the streets.

At first, your character was lost and confused. But he quickly realized that he was not alone. A group of seagulls had taken him under their wing, and they taught him how to survive in the harsh and unforgiving world. Together, they roamed the streets, scavenging for food and shelter.

One day, while exploring a dark and dangerous alley, your character stumbled upon a portal to a hell dimension. Without hesitation, he jumped through the portal, determined to face whatever challenges lay ahead. Inside the hell dimension, your character was tested to his limits. He was forced to fight for his life against demonic creatures and malevolent spirits.

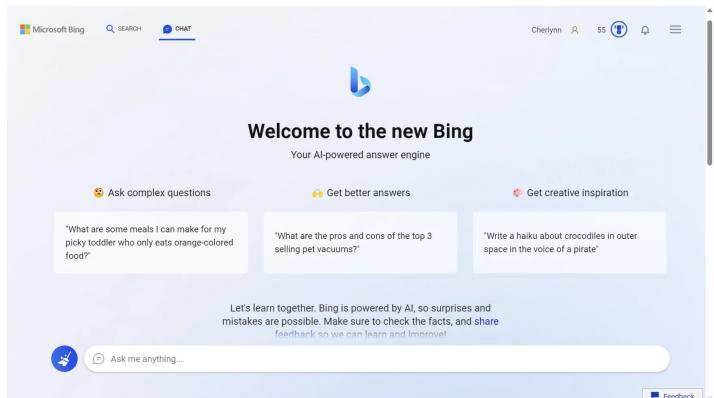
Despite the odds, your character managed to survive and even thrive in the hell dimension. He became a fierce warrior, mastering the art of combat and gaining powerful magical abilities. But in the heat of battle, he suffered a terrible injury and lost one of his limbs.

Despite this setback, your character refused to give up. He continued to fight, using his remaining limb and his magical powers to defeat his enemies. And eventually, he emerged from the hell dimension victorious, having achieved a state of enlightenment and inner peace.

Now, your character returns to the world he once knew, a changed man. He is no longer the royal prince he once was, but rather a powerful and wise warrior, ready to face whatever challenges may come his way.

A screenshot of a terminal window titled "sentiments.js". It shows a snippet of Node.js code using the "fetch" module to check the sentiment of text. The code includes imports, a function named "isPositive", and a main block that sends a POST request to a sentiment analysis API and checks the response. A "Copilot" button is visible at the bottom right of the code area.

```
1 //!/usr/bin/env ts-node
2
3 import { fetch } from "fetch-h2";
4
5 // Determine whether the sentiment of text is positive
6 // Use a web service
7 async function isPositive(text: string): Promise<boolean> {
8   const response = await fetch(`http://text-processing.com/api/sentiment/`, {
9     method: "POST",
10    body: `text=${text}`,
11    headers: {
12      "Content-Type": "application/x-www-form-urlencoded",
13    },
14  });
15  const json = await response.json();
16  return json.label === "pos";
17}
```



Chain of Thought Prompting Elicits Reasoning in Large Language Models (Wei et al., 2022)

Large Language Models are Zero-Shot Reasoners (Kojima et al., 2022)

ChatGPT (OpenAI, 2022)

Copilot (GitHub, 2021)

Bing (Microsoft, 2023)

# LLMs are impressive general models

## Chain of Thought Prompting

### Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls.  $5 + 6 = 11$ . The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

### Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had  $23 - 20 = 3$ . They bought 6 more apples, so they have  $3 + 6 = 9$ . The answer is 9. ✓

### (d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: Let's think step by step.

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

The screenshot shows a conversation between a user and a large language model. The user asks for a ridiculous backstory for a character. The model responds with a detailed narrative involving a royal prince who becomes homeless and fights his way through hell to become a warrior. This narrative is presented as a chain of thought, explaining the steps from the initial premise to the final outcome.

I'm working on a new character for a game of mine, and I need him to have the most ridiculous backstory ever. Like, literally so ridiculous that it arguably doesn't make sense. I need it to include the following themes: royalty, abandonment, hell dimensions, seagulls, loss of limb, and enlightenment. Can you think of a ridiculous backstory for him?

Sure, here's an idea for a ridiculous backstory for your character:

Your character was once a royal prince, born into a wealthy and powerful family. He grew up surrounded by luxury and privilege, surrounded by opulence and excess. But one day, his parents abandoned him, casting him out of the royal court and leaving him to scavenge for food on the streets.

He quickly realized that he was no longer welcome in the palace, so he began to scavenge for food on the streets. He learned to fight to protect himself from other scavengers and malevolent spirits. Despite the odds, your character managed to survive and even thrive in the hell dimension. He became a fierce warrior, mastering the art of combat and gaining powerful magical abilities. But in the heat of battle, he suffered a terrible injury and lost one of his limbs.

Despite this setback, your character refused to give up. He continued to fight, using his remaining limb and his magical powers to defeat his enemies. And eventually, he emerged from the hell dimension victorious, having achieved a state of enlightenment and inner peace.

Now, your character returns to the world he once knew, a changed man. He is no longer the royal prince he once was, but rather a powerful and wise warrior, ready to face whatever challenges may come his way.

The screenshot shows a code editor with a file named 'sentiments.ts'. The code is a Node.js script that uses the 'fetch' module to send a POST request to a sentiment analysis API at 'http://text-processing.com/api/sentiment/'. It expects a JSON response with a 'label' field indicating whether the text is positive or negative.

```
#!/usr/bin/env ts-node
import { fetch } from "fetch-h2";
// Determine whether the sentiment of text is positive
// Use a web service
async function isPositive(text: string): Promise<boolean> {
  const response = await fetch(`http://text-processing.com/api/sentiment/`, {
    method: "POST",
    body: `text=${text}`,
    headers: {
      "Content-Type": "application/x-www-form-urlencoded",
    },
  });
  const json = await response.json();
  return json.label === "pos";
}
```

The screenshot shows the new Bing AI-powered answer engine interface. It features a search bar at the top with the text 'come to the new Bing' and 'Your AI-powered answer engine'. Below the search bar are three sections: 'Ask complex questions', 'Get better answers', and 'Get creative inspiration'. Each section includes a representative icon and a brief description of the feature. At the bottom, there is a feedback button labeled 'Feedback'.

Ask complex questions

Get better answers

Get creative inspiration

"What are some meals I can make for my picky toddler who only eats orange-colored food?"

"What are the pros and cons of the top 3 selling pet vacuums?"

"Write a haiku about crocodiles in outer space in the voice of a pirate"

Let's learn together. Bing is powered by AI, so surprises and mistakes are possible. Make sure to check the facts, and share feedback so we can learn and improve!

Ask me anything...

Feedback

Chain of Thought Prompting Elicits Reasoning in Large Language Models (Wei et al., 2022)

Large Language Models are Zero-Shot Reasoners (Kojima et al., 2022)

ChatGPT (OpenAI, 2022)

Copilot (GitHub, 2021)

Bing (Microsoft, 2023)

# LLMs are impressive general models

## Chain of Thought Prompting

### Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls.  $5 + 6 = 11$ . The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

### Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had  $23 - 20 = 3$ . They bought 6 more apples, so they have  $3 + 6 = 9$ . The answer is 9. ✓

### (d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: Let's think step by step.

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

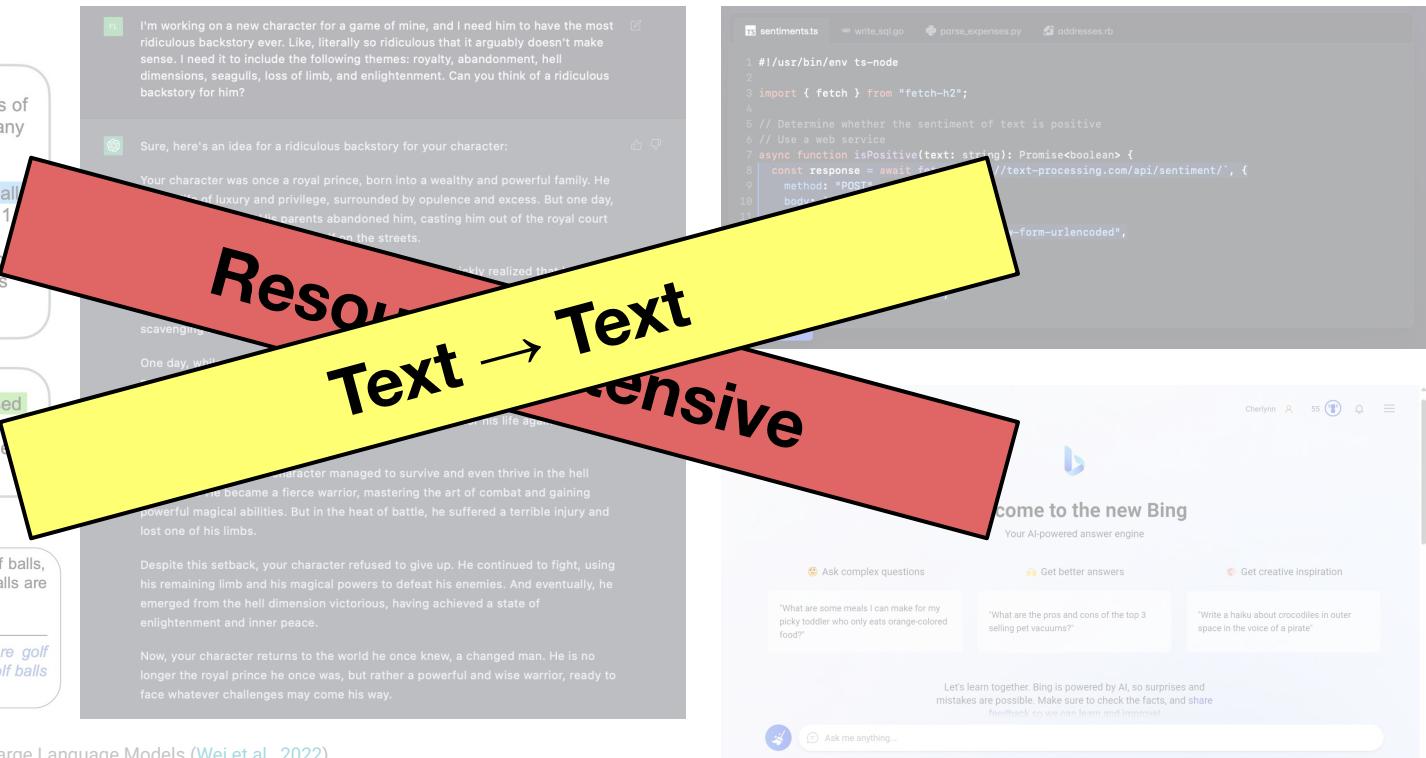
Chain of Thought Prompting Elicits Reasoning in Large Language Models (Wei et al., 2022)

Large Language Models are Zero-Shot Reasoners (Kojima et al., 2022)

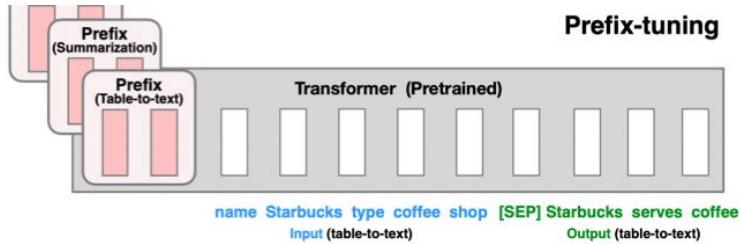
ChatGPT (OpenAI, 2022)

Copilot (GitHub, 2021)

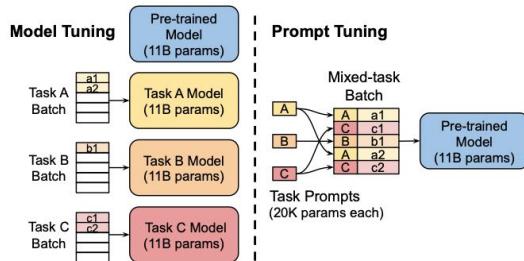
Bing (Microsoft, 2023)



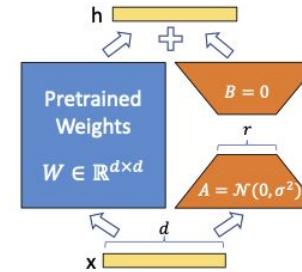
# Parameter Efficient Adaptation



**Prefix Tuning:** Learns a prefix embedding (for each layer) to adapt to new tasks. ~99.9% of the model kept frozen.



**Prompt Tuning:** Similar idea to prefix-tuning, but learns just a single prefix for input embeddings.

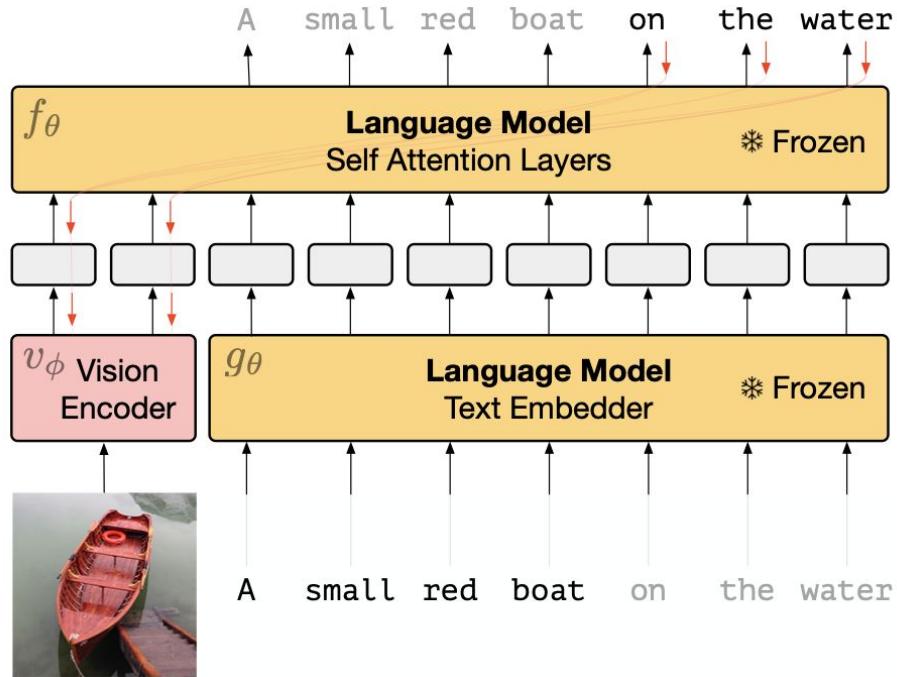


**Low-Rank Adaptation (LoRA):** Injects trainable rank decomposition matrices into each Transformer layer of a pretrained model.

# Multimodal LMs: Frozen (2021)

## Frozen:

- Prefix tuning for adapting LLMs to image captioning.
- ~95% of the model kept frozen.
- Capable of compelling few-shot multi-modal reasoning.



# Multimodal LMs: Frozen (2021)

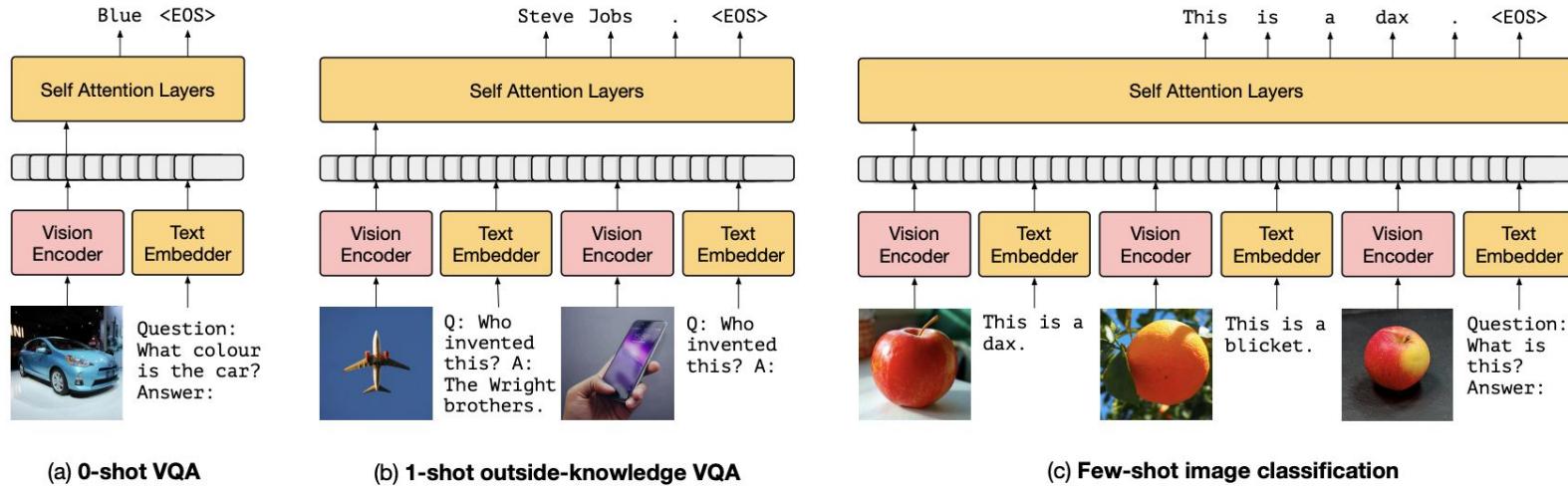


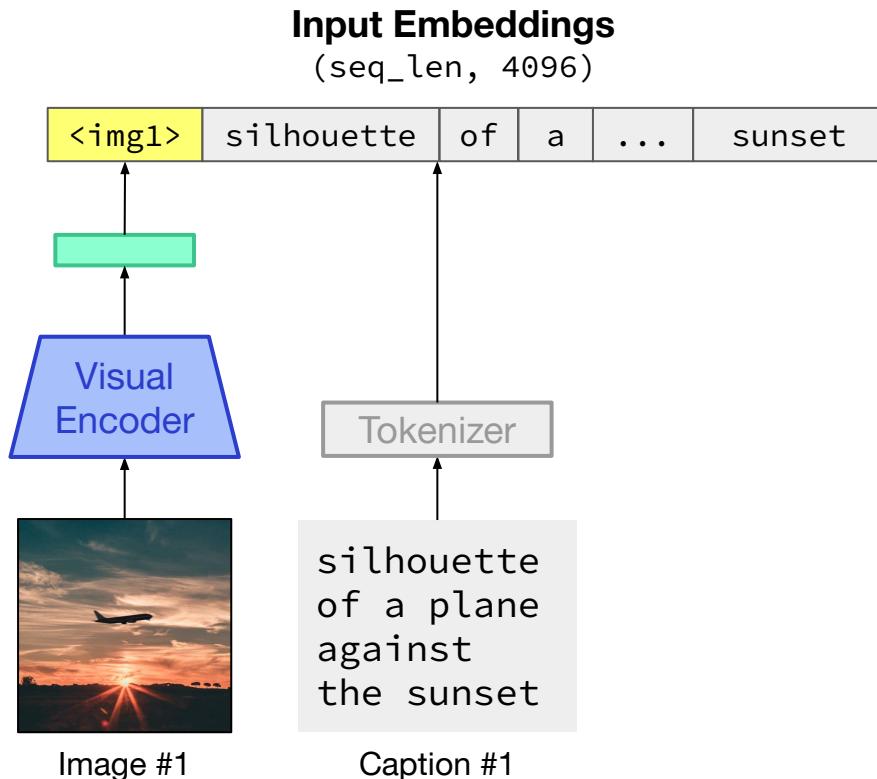
Figure 3: Inference-Time interface for *Frozen*. The figure demonstrates how we can support (a) visual question answering, (b) outside-knowledge question answering and (c) few-shot image classification via in-context learning.

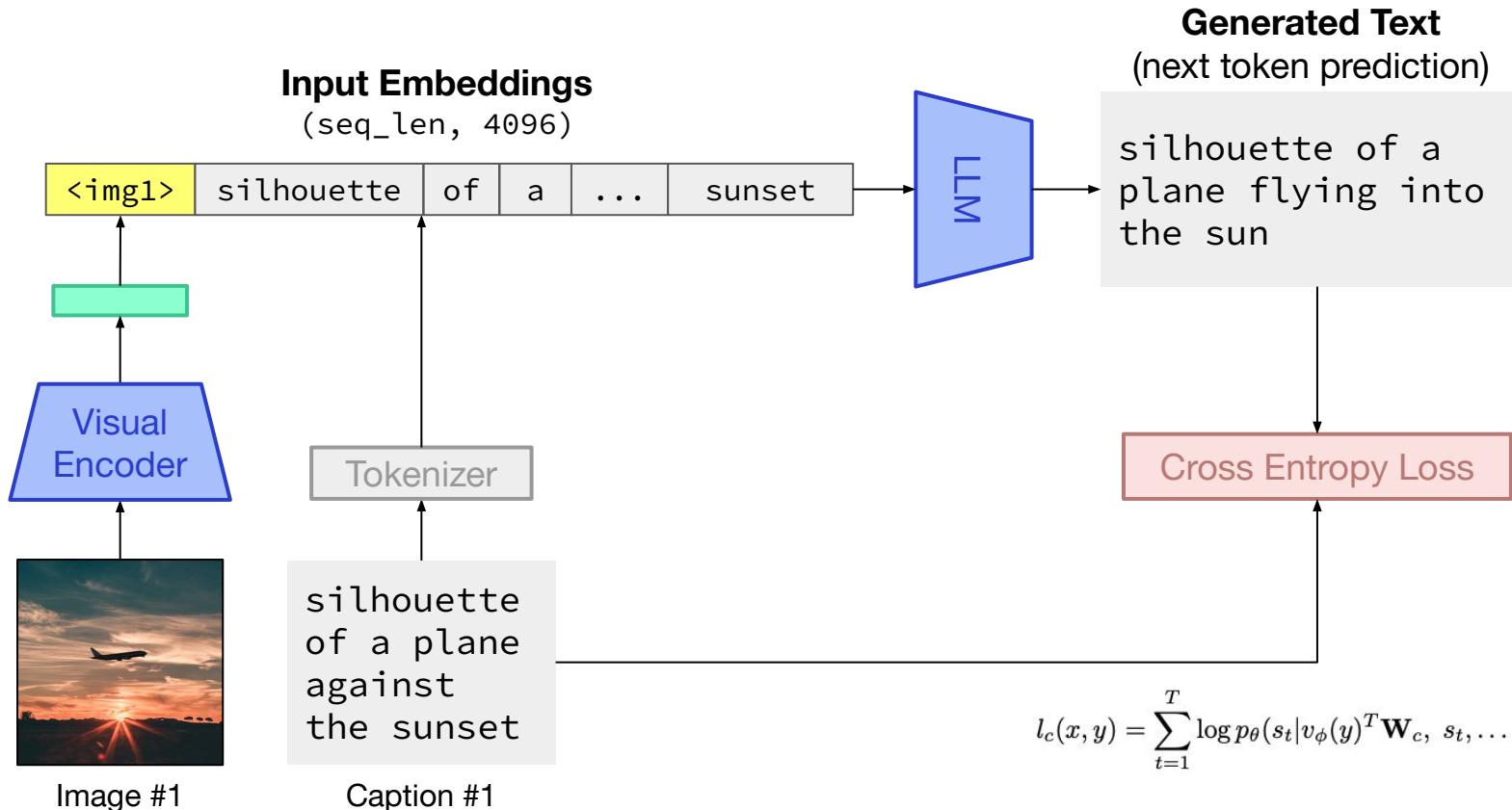


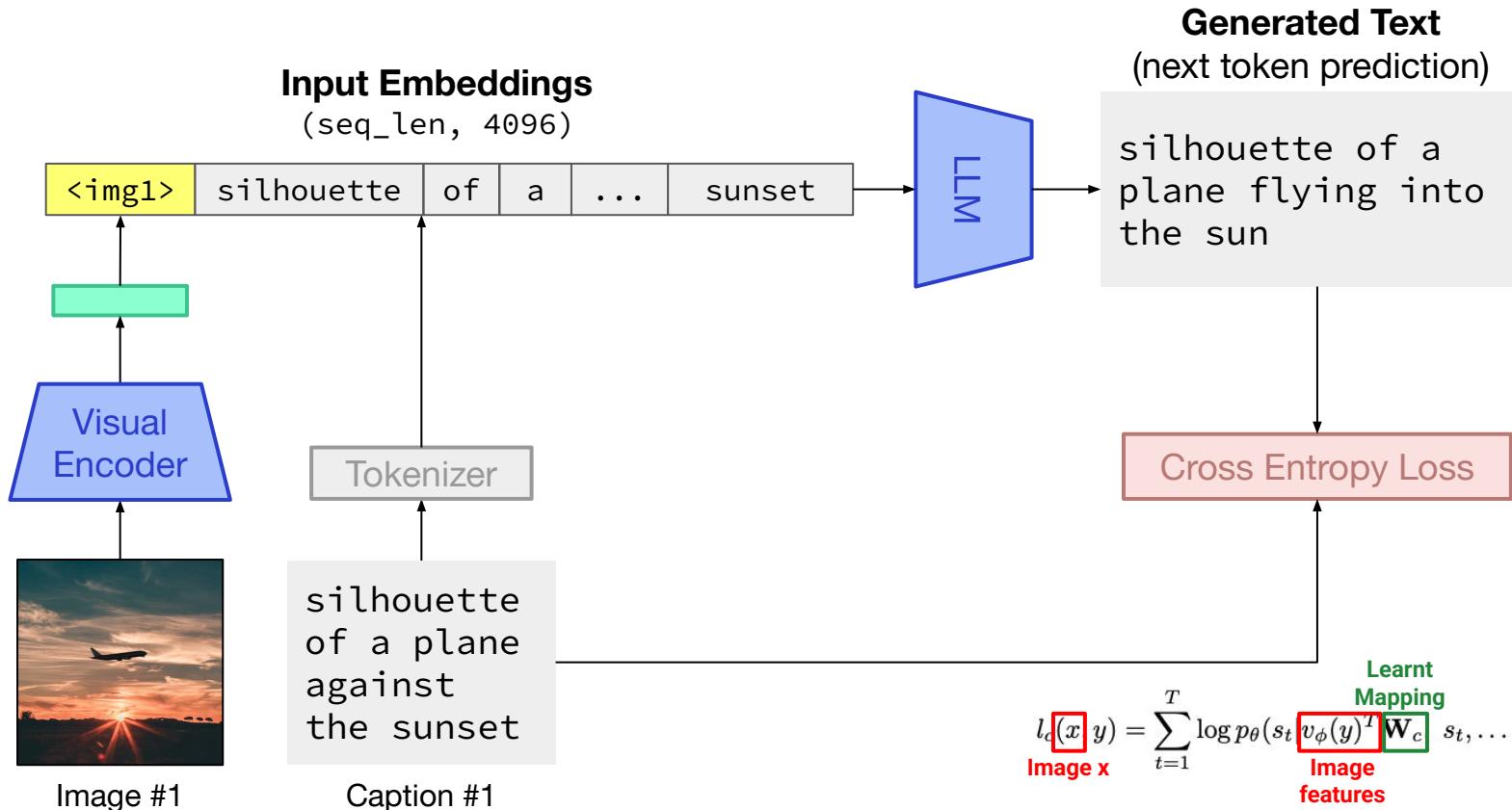
Image #1

silhouette  
of a plane  
against  
the sunset

Caption #1

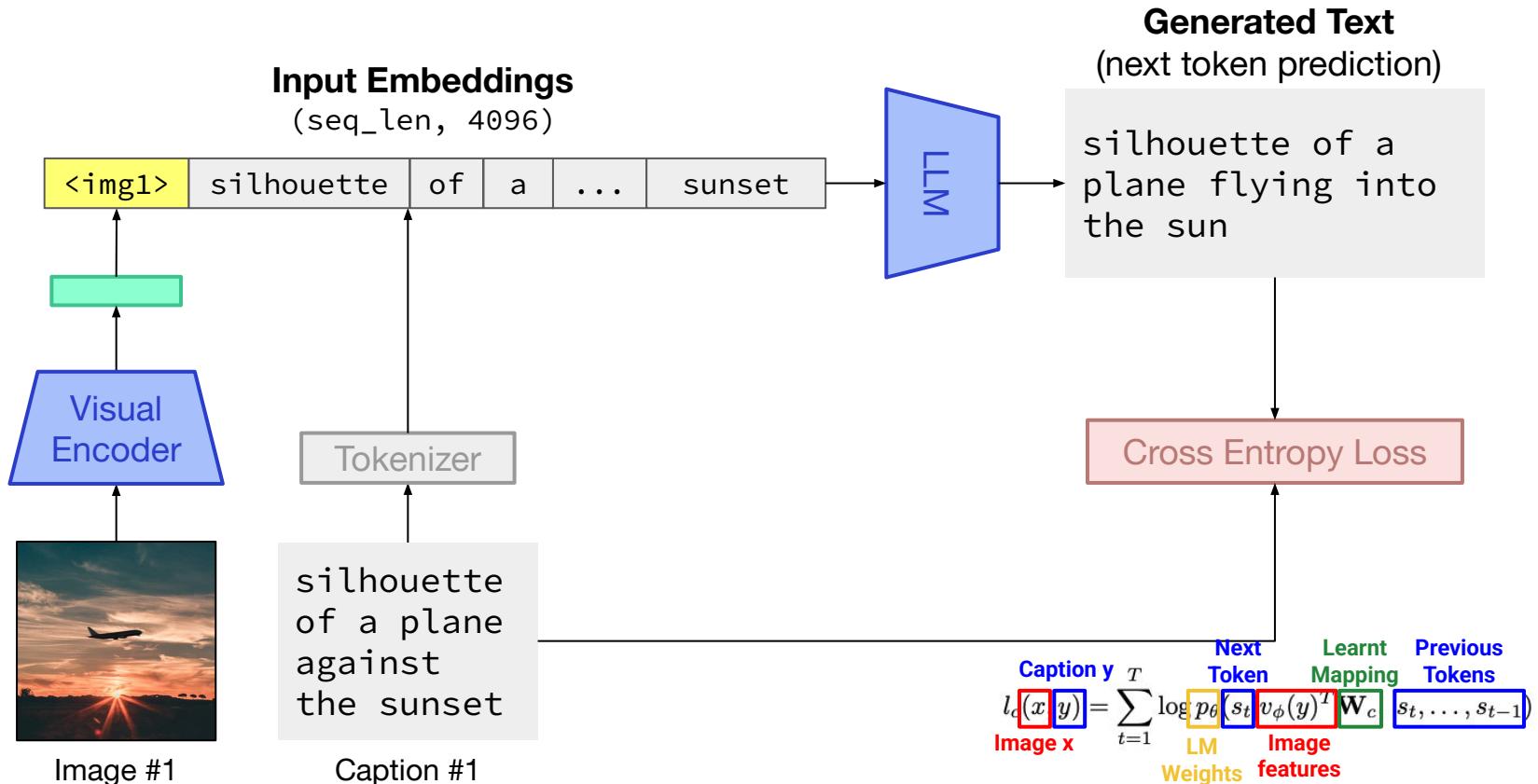






$$l_d(x, y) = \sum_{t=1}^T \log p_\theta(s_t | v_\phi(y)^T \mathbf{W}_c, s_t, \dots, s_{t-1})$$

Image x      █ █ █  
 Image features      Learnt Mapping



# Why does this work?

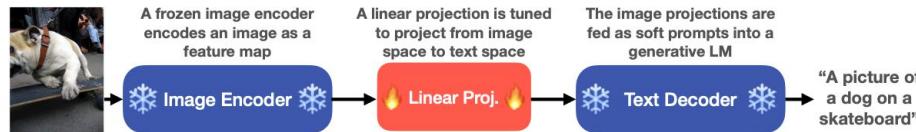
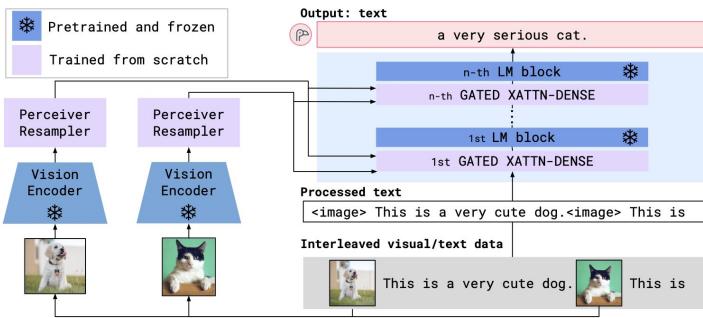


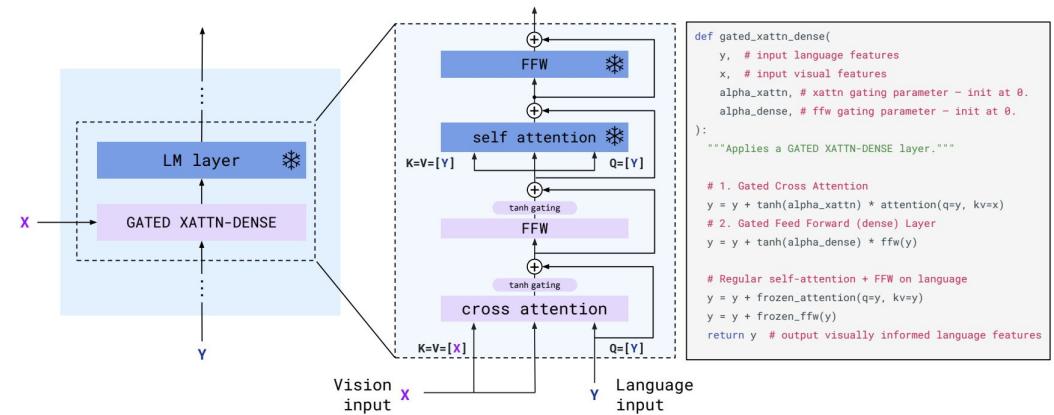
Figure 1: We train linear projections from image representations into the input space of a language model to produce captions describing images. We find that LMs can describe the contents of most image representations, but performance varies based on the type of image encoder used.

Merullo et al. showed that pretrained text-only LMs and pretrained visual encoders produce functionally equivalent representations up to a linear mapping.

# Multimodal LMs: Flamingo (2022)

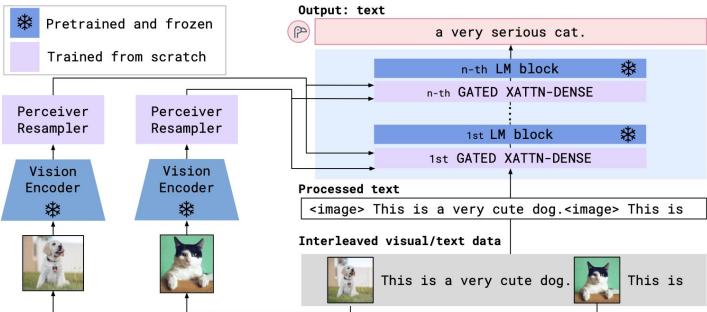


**Flamingo:** Finetunes new cross-attention layers on top of a 70B LLM. Achieves SOTA on many multi-modal tasks.

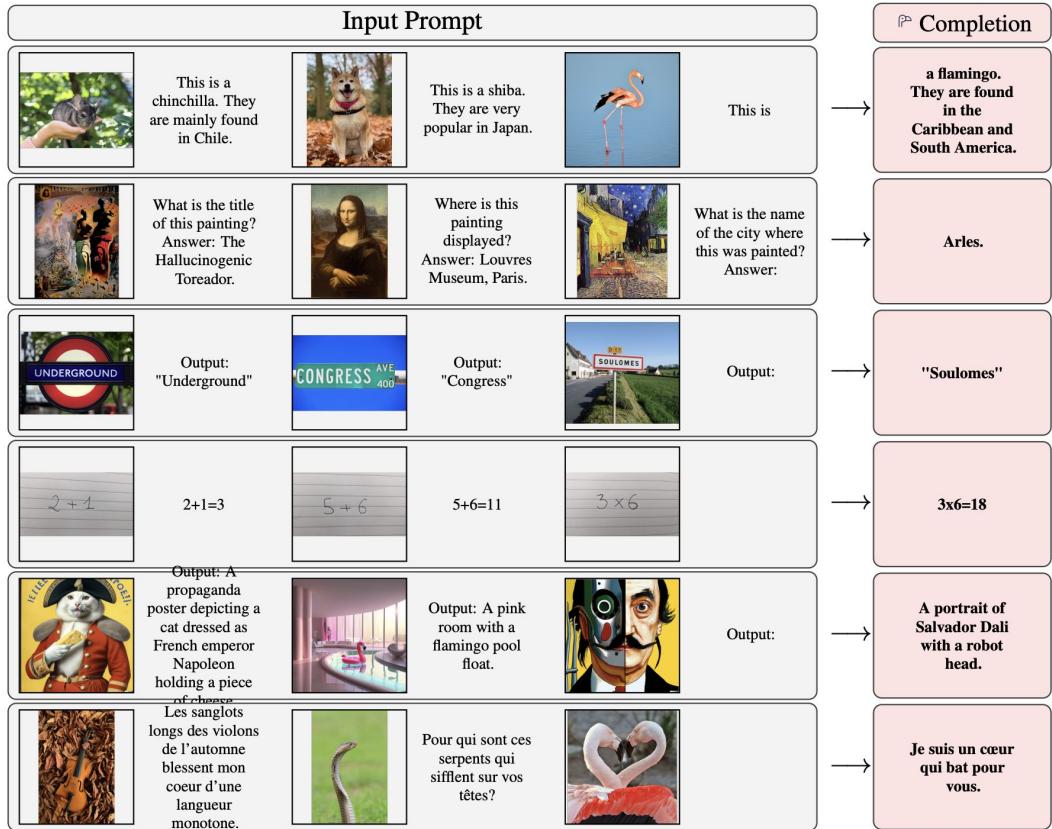


Introduced cross-attention layers between existing frozen LLM layers. Purple blocks are finetuned, blue blocks are kept frozen.

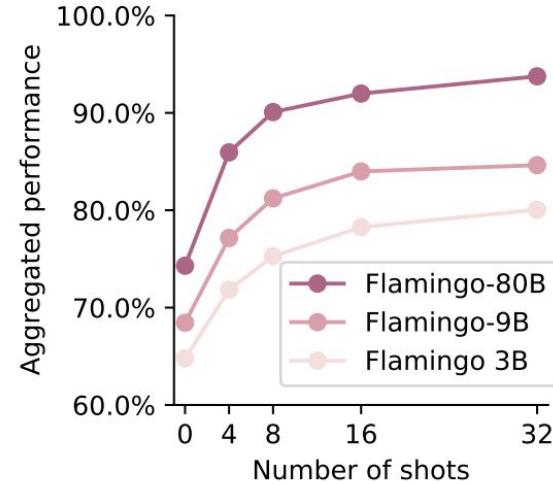
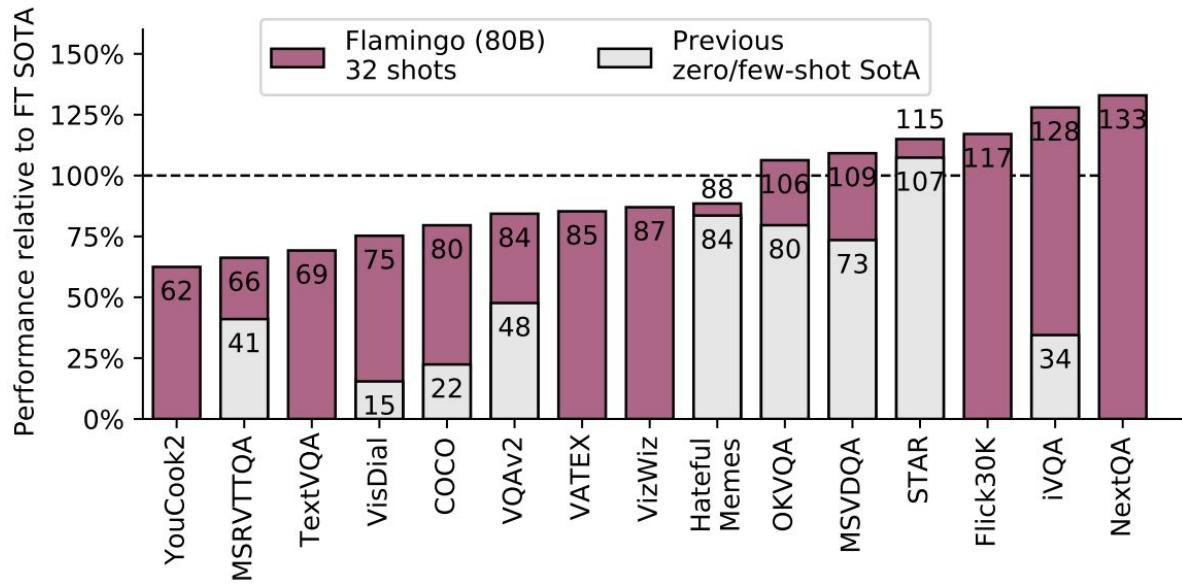
# Multimodal LMs: Flamingo (2022)



**Flamingo:** Finetunes new cross-attention layers on top of a 70B LLM. Achieves SOTA on many multi-modal tasks.



# Multimodal LMs: Flamingo (2022)



# Multimodal LMs: Flamingo (2022)



This is a picture of two teddy bears on the moon.

What are they doing?

They are having a conversation.

What object are they using?

It looks like a computer.

Is this surprising?

Yes, it is surprising.

Why is this picture surprising to you?

I think it is surprising because teddy bears are not usually found on the moon.



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.



This is an apple with a sticker on it.

What does the sticker say?

The sticker says "iPod".

Where is the photo taken?

It looks like it's taken in a backyard.

Do you think it is printed or handwritten?

It looks like it's handwritten.

What color is the sticker?

It's white.



This is a cityscape. It looks like Chicago.

What makes you think this is Chicago?

I think it's Chicago because of the Shedd Aquarium in the background.



What about this one? Which city is this and what famous landmark helped you recognise the city?

This is Tokyo. I think it's Tokyo because of the Tokyo Tower.

# Data: Flamingo (2022)



This is an image of a flamingo.

Image-Text Pairs dataset



A kid doing a kickflip.

Video-Text Pairs dataset



Welcome to my website!



This is a picture of my dog.



This is a picture of my cat.

Multi-Modal Massive Web (M3W) dataset

## Image-Text Pairs:

1.8B pairs with short captions (~12 tokens each)

312M pairs with long descriptions (~20.5 tokens each)

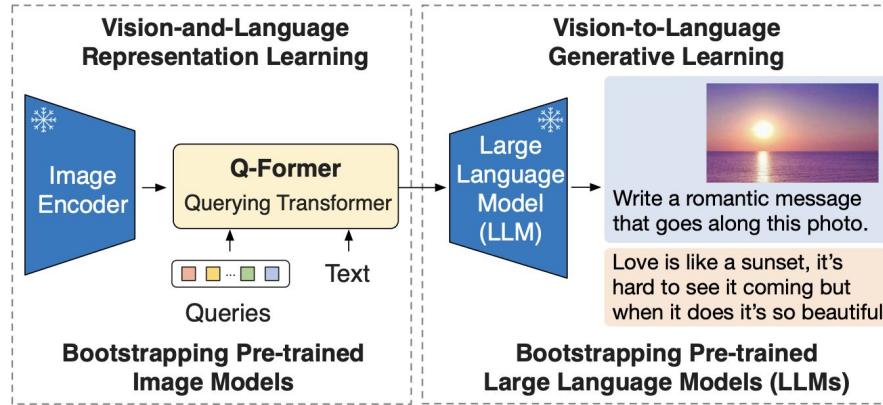
## Video-Text Pairs:

27M short videos

## M3W:

43M webpages (for each page, sample 256 tokens, take first 5 images)

# Multimodal LMs: BLIP-2 (2023)

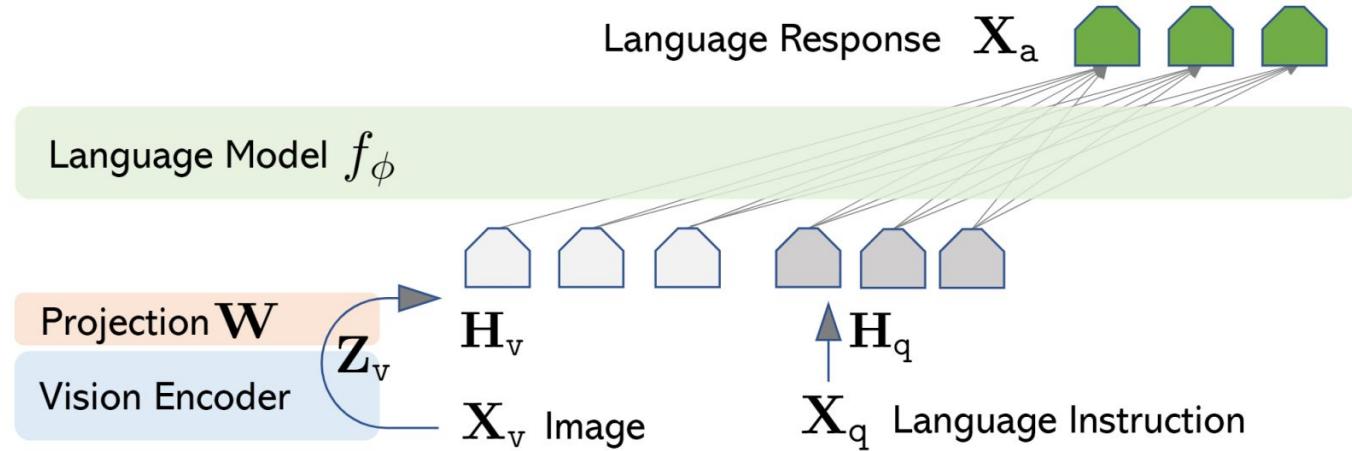


**BLIP-2:** Trains a Q-Former (essentially a transformer with learnt queries) to map between a frozen image encoder and a frozen LLM. Achieves strong performance on image captioning, VQA, and other image-to-text tasks.

# Multimodal LMs: BLIP-2 (2023)

	8	8	8
<p>Explain the advantages of this product.</p> <p>The audi e-tron quattro concept is a plug-in hybrid electric sports car that has a range of 310 miles and can accelerate from 0 to 60 mph in just 4 seconds.</p>			
	8	8	8
<p>Tell me something about the history of this place.</p> <p>The great wall of china was built in 221 bc by emperor qin shihuang to protect the emperor's capital from invaders from the north.</p>			
	8	8	8
<p>Write down the facts that you know about this flower.</p> <p>It is a genus of flowering plants in the orchid family, orchidaceae. It is native to eastern asia and south east asia. It is a perennial flowering plant that grows up to a height of 1 m.</p>			
	8	8	8
<p>Is this photo unusual?</p> <p>Yes, it's a house that looks like it's upside down.</p> <p>How could someone get out of the house?</p> <p>It has a slide on the side of the house.</p>			
	8	8	8
<p>What are shown in the photo?</p> <p>A man and a chicken.</p> <p>What does the man feel and why?</p> <p>He is scared of the chicken because it is flying at him.</p>			
	8	8	8
<p>What are the ingredients I need to make this?</p> <p>Pepperoni, mozzarella cheese, pizza sauce, olive oil, salt, pepper, basil.</p> <p>What is the first step?</p> <p>Place the pizza dough on a baking sheet, brush with olive oil, sprinkle with salt, pepper, and basil.</p>			

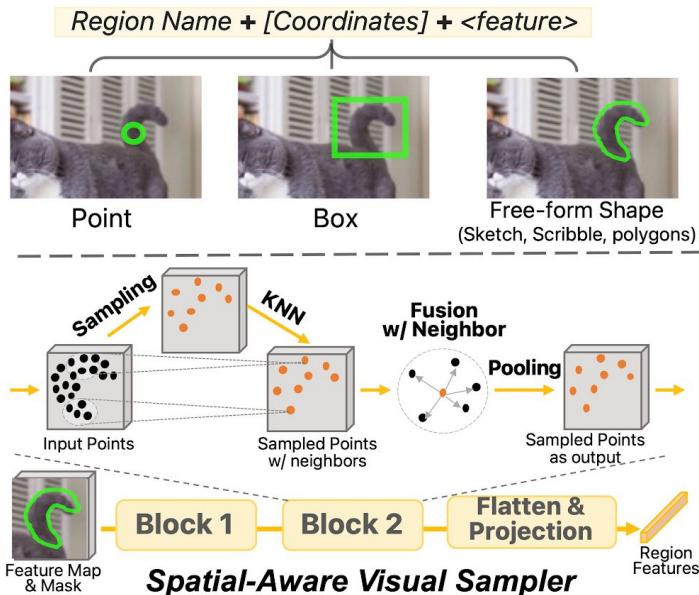
# Multimodal LMs: LLaVA (2023)



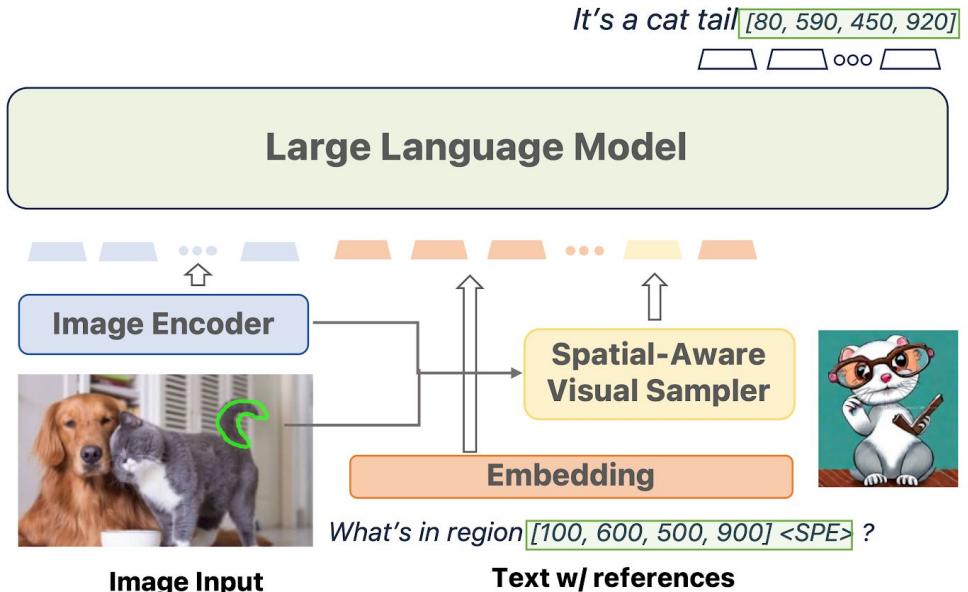
**LLaVA:** Finetunes a linear layer ( $W$ ) over a frozen vision encoder and a frozen LLM.  
Showcases strong performance by finetuning on paired data of images and text instructions  
(some GPT-4 generated).

# Multimodal LMs: Ferret (2023)

## Hybrid Region Representation

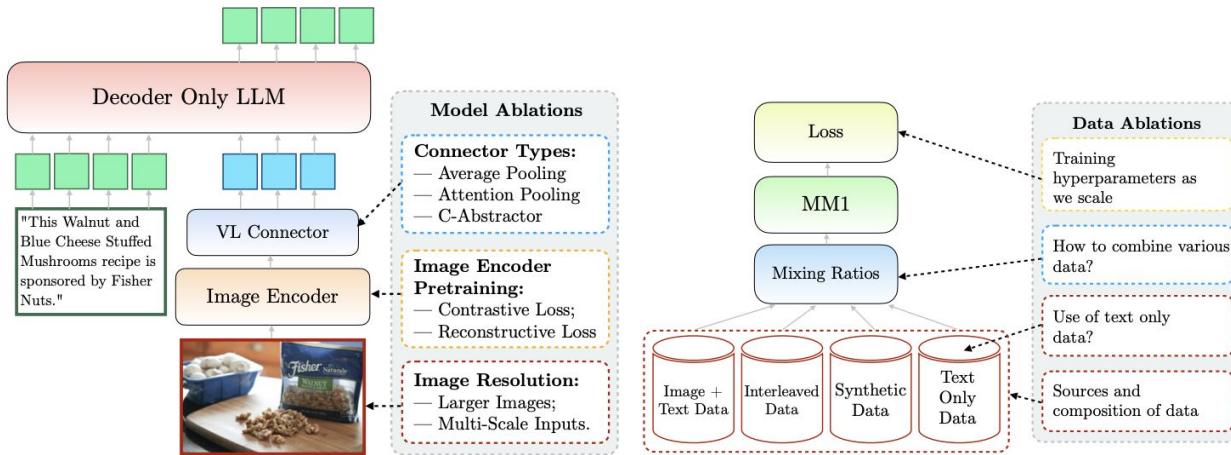


## Ferret Model



**Ferret:** Hybrid input representations, enabling fine-grained grounding and referencing.

# Multimodal LMs: MM1 (2024)

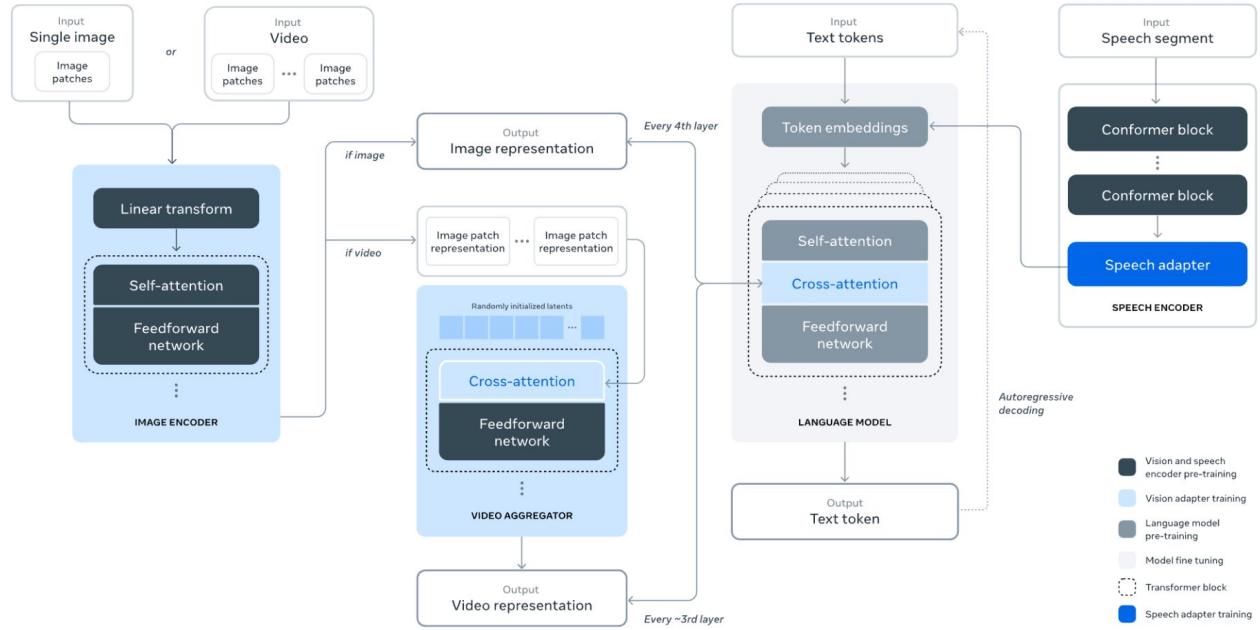


**MM1:** ViT-L image encoder + C-Abstractor (CNN) mapper + 1.2B decoder-only LM

# Multimodal LMs: MM1 (2024) Analysis

- **Biggest impact on results:**
  - Input image resolution has the highest impact to results (224px → 336px leads to +3% boost)
  - Model size (+1%) and training data composition (+1%) are also impactful
- **Vision-Language Connector**
  - Number of visual tokens and image resolution matters most
  - The particular architecture has little effect
- **Pretraining Data**
  - Interleaved data is instrumental for few-shot and text-only performance
  - Caption data lifts zero-shot performance
  - Text-only data helps with few-shot and text-only performance
  - Synthetic captions help with few-shot learning

# Multimodal LMs: Llama3V (2024)



**Figure 28 Illustration of the compositional approach to adding multimodal capabilities to Llama 3 that we study in this paper.** This approach leads to a multimodal model that is trained in five stages: **(1)** language model pre-training, **(2)** multi-modal encoder pre-training, **(3)** vision adapter training, **(4)** model finetuning, and **(5)** speech adapter training.

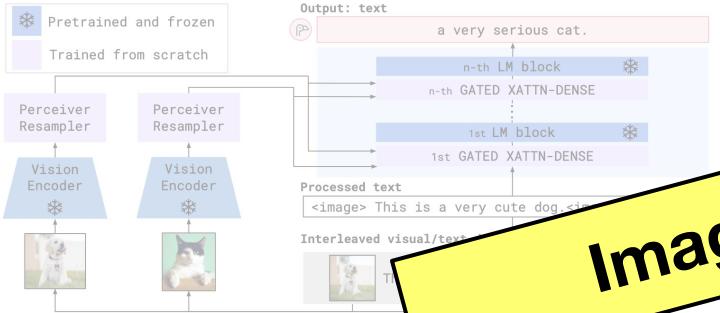
# Multimodal LMs: Llama3V (2024) details

- **Images**
  - Quality filtering, de-duplication
  - Resampling to prioritize rare n-grams
  - Run OCR pipelines and concat with original caption
- **Documents**
  - Transcribe documents and pair with the original caption
- **Safety**
  - Removed unsafe content
- **Annealing**
  - Created a small set of high quality 500M image-text pairs
    - Visual grounding (x, y), screenshot + html, QA pairs, synthetic captions, synthetic structured images (LaTeX)

# Multimodal LMs: Others

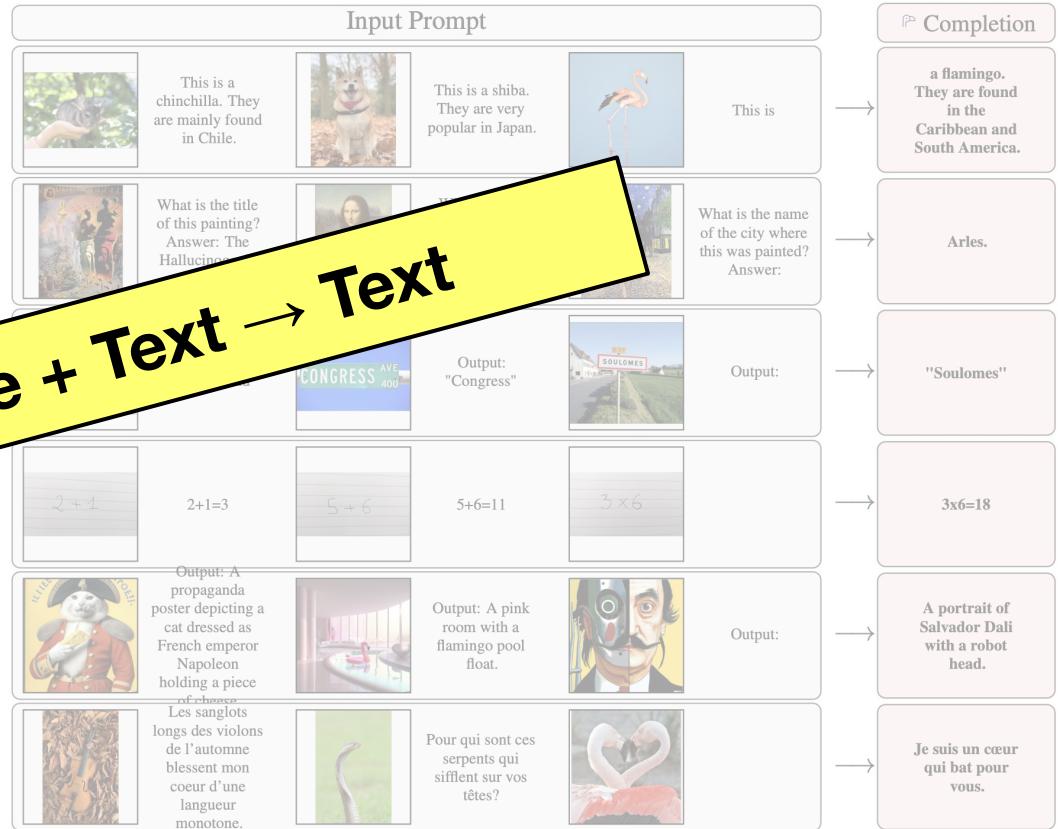
- **Closed models:** gpt-4, Gemini-1.5, Claude 3.5
  - The general public is not 100% sure what they do :)
- **Open sourced models:**
  - [IDEFICS](#)
  - [OpenFlamingo](#)
  - [Owen-VL](#)
  - [CogVLM](#)

# Multimodal LMs: Flamingo (2022)



**Flamingo:** Finetunes new cross-attention layers on top of a 70B LLM. Achieves SOTA on many multi-modal tasks.

Image + Text → Text



**Can we ground text-only LLMs to  
consume and produce visual data?**



# FROMAGe

Frozen Retrieval Over Multimodal Data for Autoregressive Generation

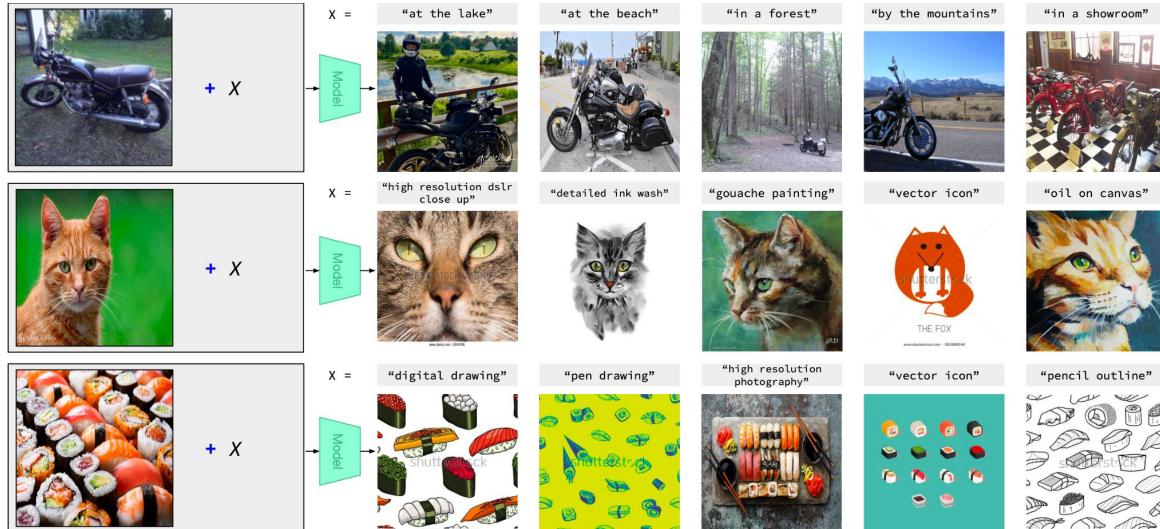
[jykoh.com/fromage](http://jykoh.com/fromage)



# FROMAGe

Frozen Retrieval Over Multimodal Data for Autoregressive Generation

[jykoh.com/fromage](http://jykoh.com/fromage)



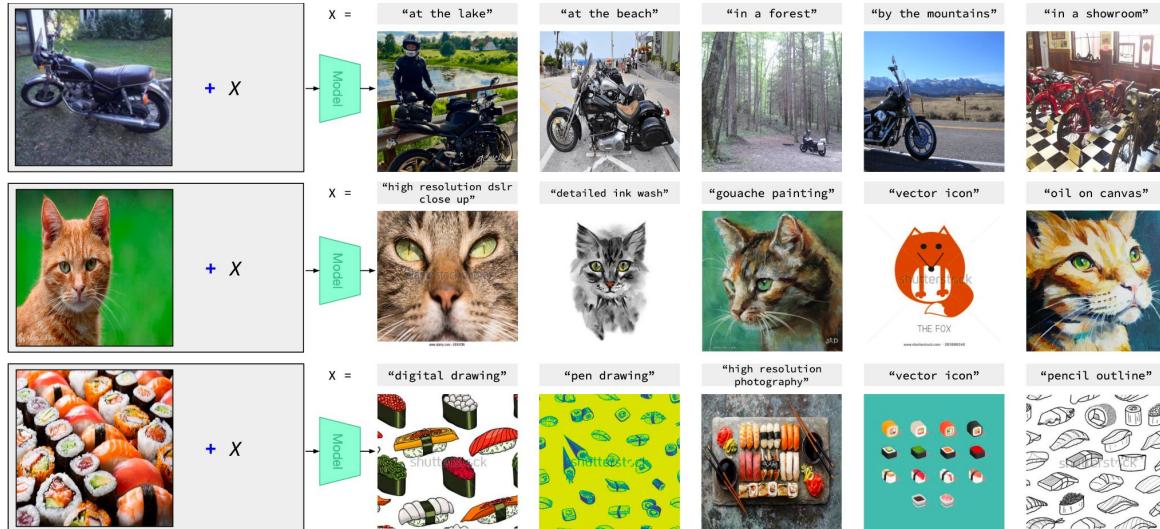
**Concept composition.** Our model is capable of retrieving relevant images conditioned on multi-modal context inputs.



# FROMAGe

Frozen Retrieval Over Multimodal Data for Autoregressive Generation

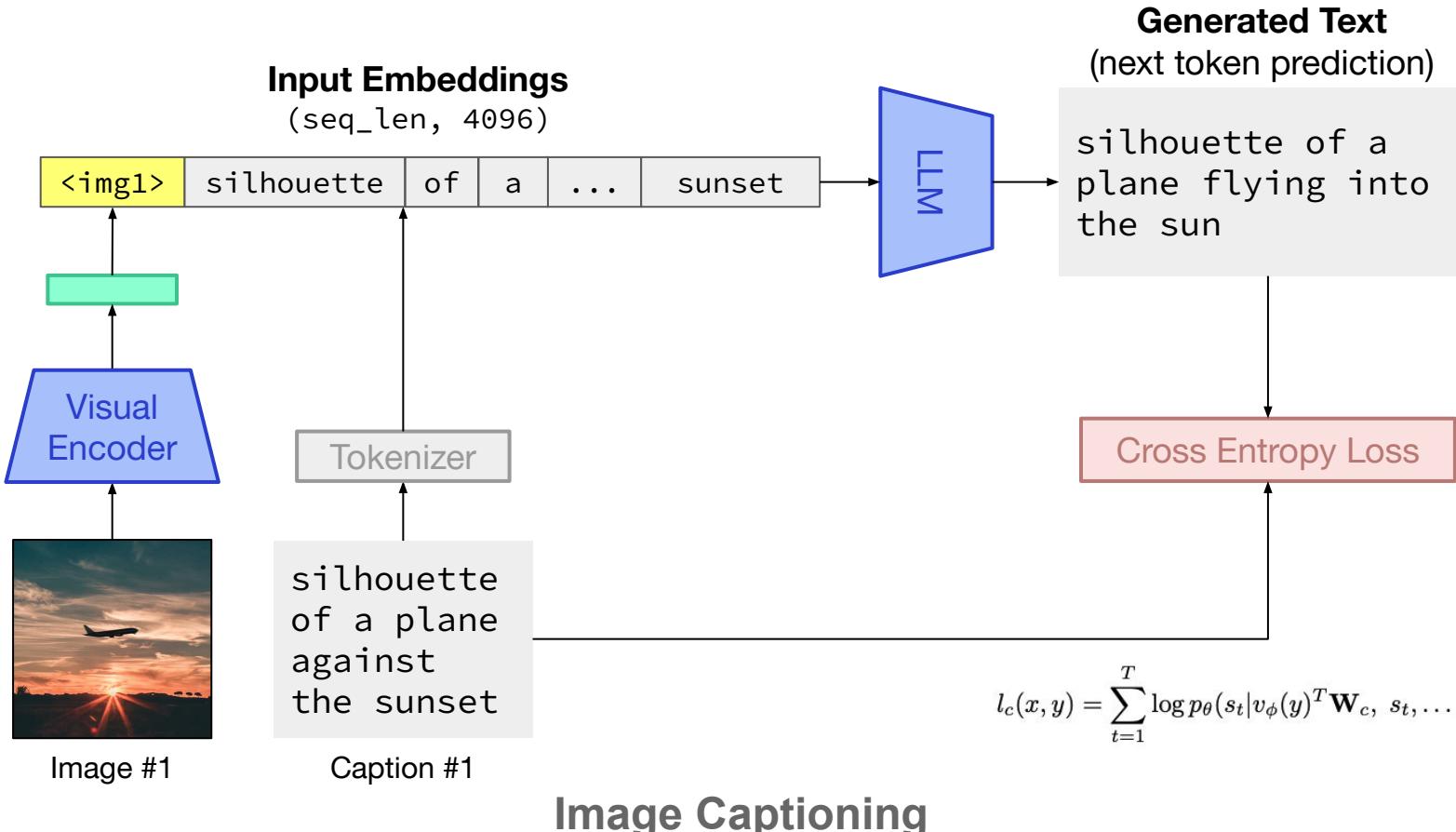
[jykoh.com/fromage](http://jykoh.com/fromage)

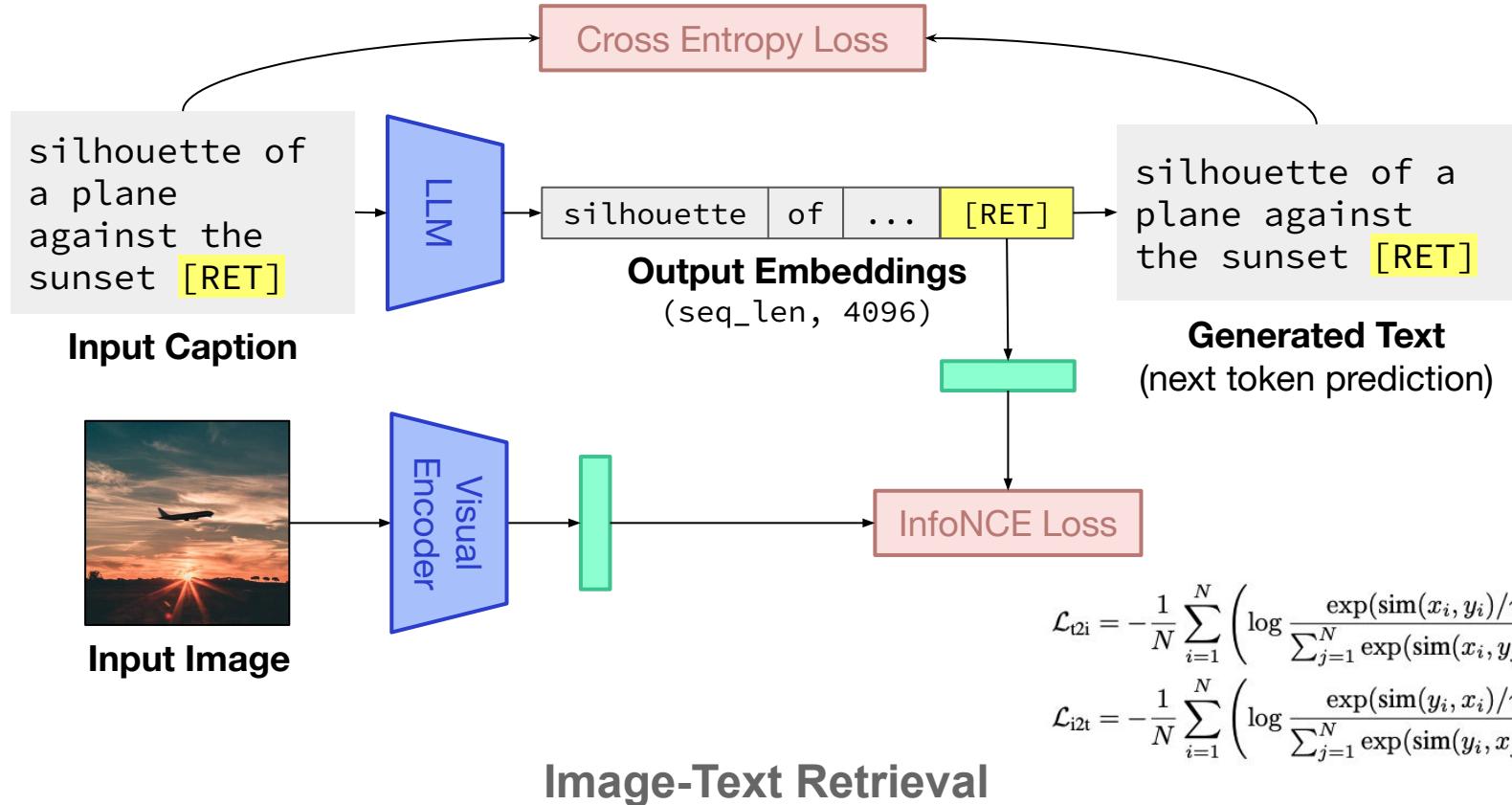


**Concept composition.** Our model is capable of retrieving relevant images conditioned on multi-modal context inputs.

Grounding Language Models to Images for Multimodal Generation ([jykoh.com/fromage](http://jykoh.com/fromage))

**Multi-modal dialogue.** Green bubbles represent model generated outputs, grey bubbles represent user input.





$$\mathcal{L}_{\text{t2i}} = -\frac{1}{N} \sum_{i=1}^N \left( \log \frac{\exp(\text{sim}(x_i, y_i)/\tau)}{\sum_{j=1}^N \exp(\text{sim}(x_i, y_j)/\tau)} \right)$$

$$\mathcal{L}_{\text{i2t}} = -\frac{1}{N} \sum_{i=1}^N \left( \log \frac{\exp(\text{sim}(y_i, x_i)/\tau)}{\sum_{j=1}^N \exp(\text{sim}(y_i, x_j)/\tau)} \right)$$

## Beavers

I am thinking of an animal. It is brown and furry, and has a tail. What is it?

A squirrel.

No, the animal I am thinking of lives in the river.

A beaver.

That's right, what do beavers look like?

They have a tail.

Show me an illustration of a beaver

A beaver.



Yes, what about a pencil drawing of one?



What about a photo of one?



## Unique Houses



What is this?

This is a house with a unique design.

What is its shape?

The house is shaped like a dome.

Can you show me more houses like this?

The house is designed by architect.



That's nice, any others?

Yes, there are many houses like this in the world.



## How to Make Macarons

Provide instructions on how to cook.

How do I make macarons?

Macarons are a French dessert that are made by combining egg whites and sugar. The egg whites are beaten until stiff and then folded into the sugar.

What does the stiff egg white look like?

The egg white should be stiff and glossy.

Show me a picture of what that looks like



After this, what do I do?

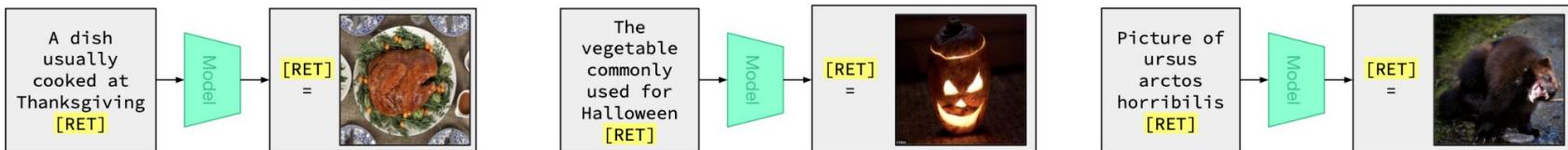
The macaron is then baked in a hot oven.

# In-Context Learning and Other Abilities



## Visual Storytelling

FROMAGe can learn from in-context examples to generate story-like image and text outputs.



## World Knowledge

Our method can draw upon knowledge learnt outside of CC3M (through large scale text pretraining of the frozen LLM) to return valid image outputs.



## Interleaved Text-to-Image Composition

Our model can transform a sequence of text inputs into text-and-image outputs. It can do coreferencing to select the appropriate images.

# Generating Multimodal Outputs: CM3 (2022)

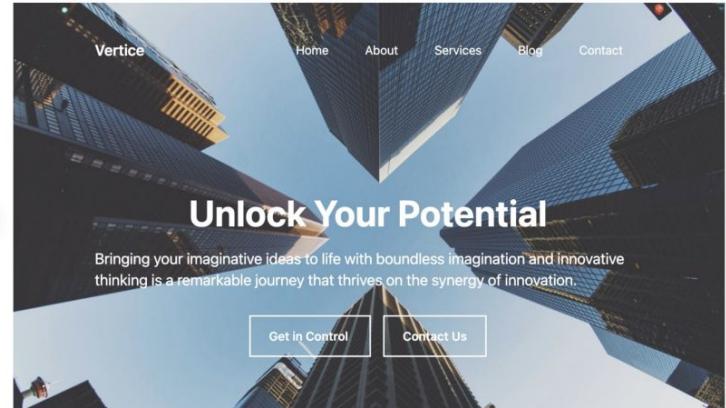
Trained a language model over HTML content.  
Generalized image + text inputs and image + text outputs (everything is a token!)


7           <a class="font-semib
8             <div>Vertice
9           </a>
10          </div>
11        </div>
12        <ul class="flex">
13          <li><a class="text-white
14            <li><a class="text-white
15              <li><a class="text-white
16                <li><a class="text-white
17                  </ul>
18                </div>
19              </div>
20              <div class="container mx-aut
21                <h1 class="text-5xl font
22                  <h1>Unlock Your Potential
23                    </h1>
24                    <p class="text-white tex
25                      <p>Unlock Your Potential</p>
26                    </p>
27                  <div class="mt-12">
28
```

Sightseer  
generation

Rendering

Rendered generated code



# Future Research: Multimodal Agents

The screenshot shows the /f/food page on Postmill. It features a sidebar with a search bar, navigation links for Postmill, Forums, and Wiki, and a login/signup button. The main content area displays a list of posts under the 'food' category. Each post includes a thumbnail image, the title, the submitter's name, the submission time, and the number of comments. The posts listed are:

- [homemade] Obligatory Halloween Pumpkin Loaf! (17,700 upvotes, 45 comments)
- [I ate] Maple Pecan Croissant (11,578 upvotes, 264 comments)
- [Homemade] Margherita pizza (11,504 upvotes, 181 comments)
- [Homemade] Sichuanese Spicy Beef Noods! (11,249 upvotes, 171 comments)
- [I ate] Sushi platter (10,235 upvotes, 163 comments)

Original Webpage

The diagram illustrates the State of the Art (SoA) of interactable elements on a webpage. It shows a comparison between the original webpage and a more detailed representation.

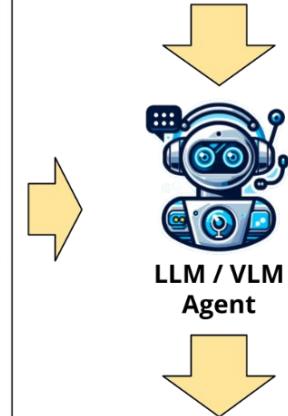
**Webpage with SoM of Interactable Elements:** This section shows a detailed view of the /f/food page from the perspective of an agent. It highlights various interactable elements with numbers in red boxes:

- Post count: 5
- Comments count: 145
- Upvote counts: 17,700, 11,578, 11,504, 11,249, 10,235
- Post creation times: 1 year ago, 1 year ago, 9 months ago, 10 months ago, 11 months ago
- Post titles: [homemade] Obligatory Halloween Pumpkin Loaf!, [I ate] Maple Pecan Croissant, [Homemade] Margherita pizza, [Homemade] Sichuanese Spicy Beef Noods!, [I ate] Sushi platter
- Post descriptions: [homemade], [I ate], [Homemade], [Homemade], [I ate]
- User names: kneechalice, AscerdingaJP, nikhilkr, conewly, laryfreako
- Toolbox items: Submissions, Comments, Hot, Toolbox, Bans, Moderation log
- Log in and Sign up buttons

**SoM Elements and Text Content:** This section shows a list of numbered elements corresponding to the highlighted areas in the diagram:

- [7] [A] [Comments]
- [8] [BUTTON] [Hot]
- [9] [IMG] [description: picture of a pumpkin]
- [10] [A] [kneechalice]

...



LLM / VLM  
Agent

click [31]

**Task:** Navigate to the page of a good Thai restaurant in Pittsburgh. It should have at least 200 reviews and 4.3 stars. Pick the one with the highest rating.



```
● ● ● visualwebarena — Python run_demo.py --instruction_path agent/prompts/jsons...  
(venv) jingyuk@GS22433 visualwebarena %  
python run_demo.py \  
--instruction_path agent/prompts/jsons  
/p_som_cot_id_actree_3s.json \  
--start_url "https://www.yelp.com/" \  
--intent "Navigate to the page of a go  
od Thai restaurant in Pittsburgh. It sho  
uld have at least 200 reviews and 4.3 st  
ars. Pick the one with the highest ratin  
g." \  
--result_dir demo_test_yelp \  
--model gpt-4-vision-preview \  
--action_set_tag som --observation_ty  
pe image_som \  
--render
```

Task: Make a reservation at Pusadee's Garden for 2 people on the earliest date for dinner. Use my name JY Koh and phone number 650-555-5555.



```
● ● ● visualwebarena — Python run_demo.py --instruction_path agent/prompts/jsons...
(venv) jingyuk@GS22433 visualwebarena % python run_demo.py \
    --instruction_path agent/prompts/jsons/
/p_som_cot_id_actree_3s.json \
    --start_url "https://www.google.com/" \
\
    --intent "Make a reservation at Pusadee's Garden for 2 people on the earliest date at any time. Use my name JY Koh and phone number 650-555-5555." \
    --result_dir demo_test_yelp \
    --model gpt-4-vision-preview \
    --action_set_tag som --observation_type image_som \
    --render
```

# Other Resources

- **Blog post:** [Multimodality and Large Multimodal Models \(LMMs\)](#)
- **Courses:** [CMU 11-777: Multimodal Machine Learning](#)
  - Lectures available on [YouTube](#)
  - Also check out [11-877: Advanced Topics in Multimodal Machine Learning](#)
- **Survey papers:**
  - [Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions](#)
  - [A Survey on Multimodal Large Language Models](#)
  - [Multimodal Large Language Models: A Survey](#)

# Thanks!

jykoh@cmu.edu