

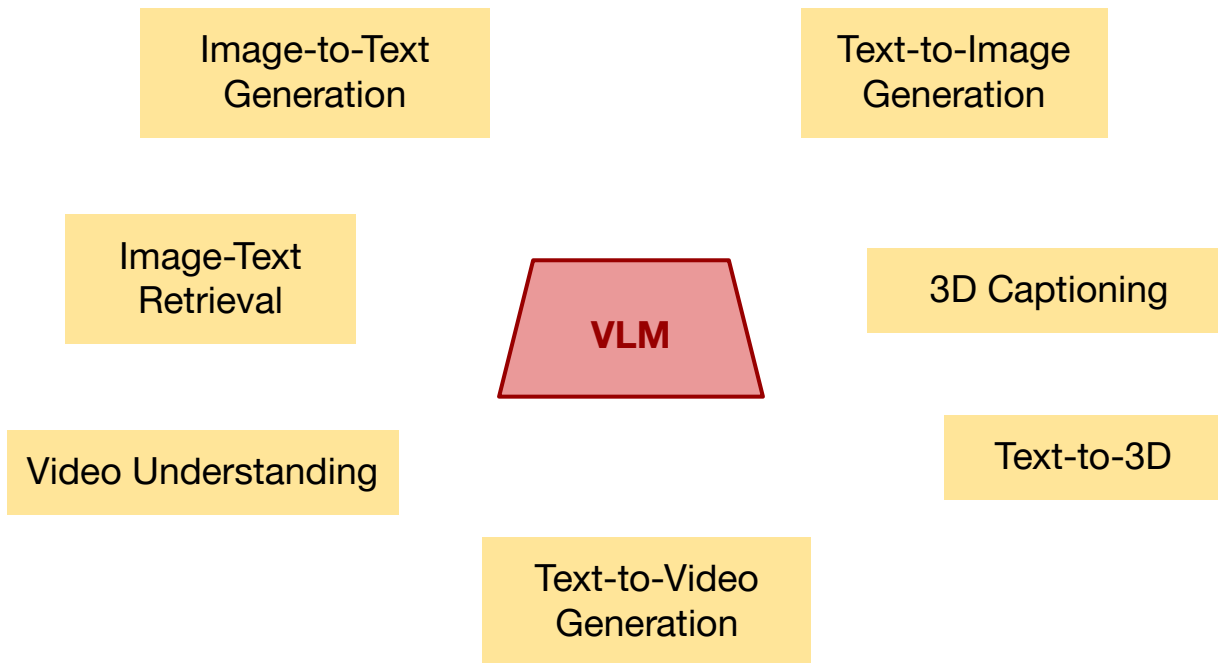
Multimodal Language Models

10-707

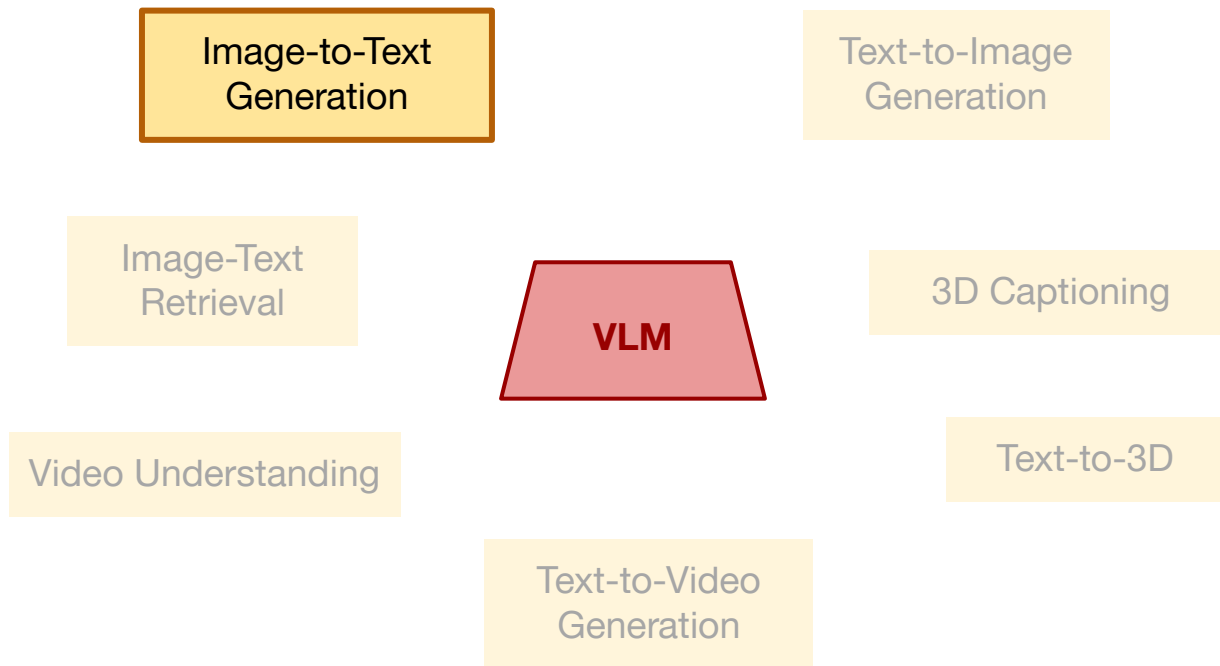
Jing Yu Koh

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

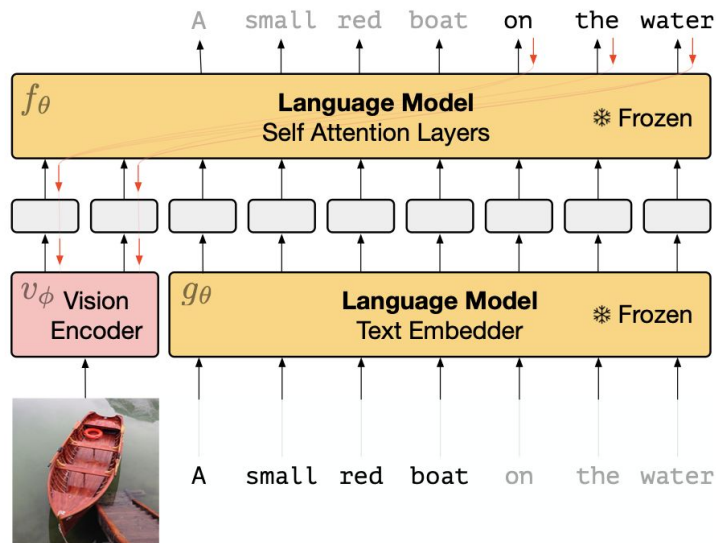


Image-to-Text Tasks

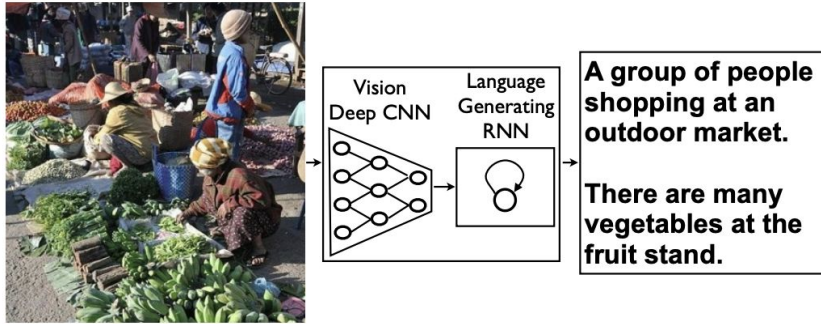


Image Captioning

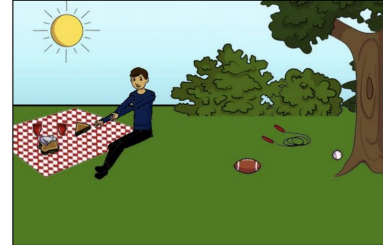
Show and Tell: A Neural Image Caption Generator ([Vinyals et al., 2014](#))
VQA: Visual Question Answering ([Agrawal et al., 2016](#))



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



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



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



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

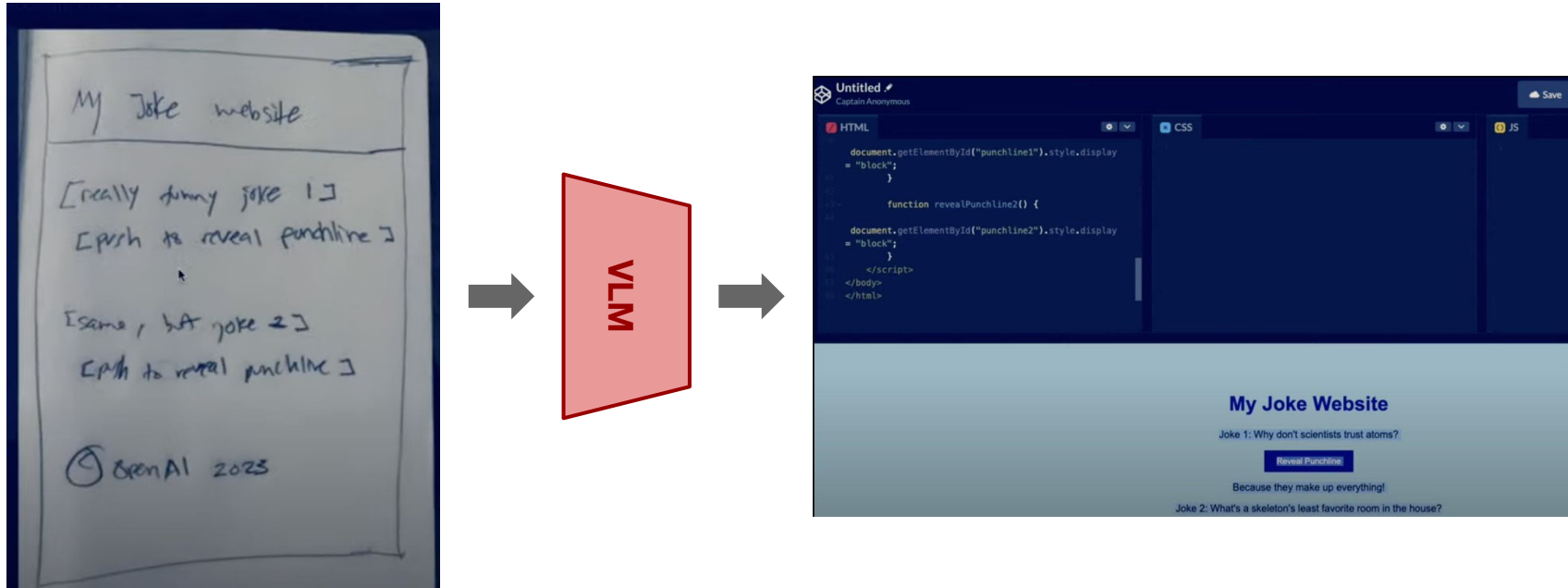


Image-to-Text Applications: Image-to-LaTeX

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



Can you
write this in
LaTeX?



```


$$\zeta(s) =$$


$$\prod_{p \text{ prime}}$$


$$\frac{1}{1-p^{-s}}$$


$$=$$


$$\frac{1}{1-2^{-s}}$$


$$\cdots$$


$$\frac{1}{1-3^{-s}}$$


$$\cdots$$


$$\frac{1}{1-5^{-s}}$$


$$\cdots$$


$$\frac{1}{1-7^{-s}}$$


$$\cdots$$


```

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 Like Comment

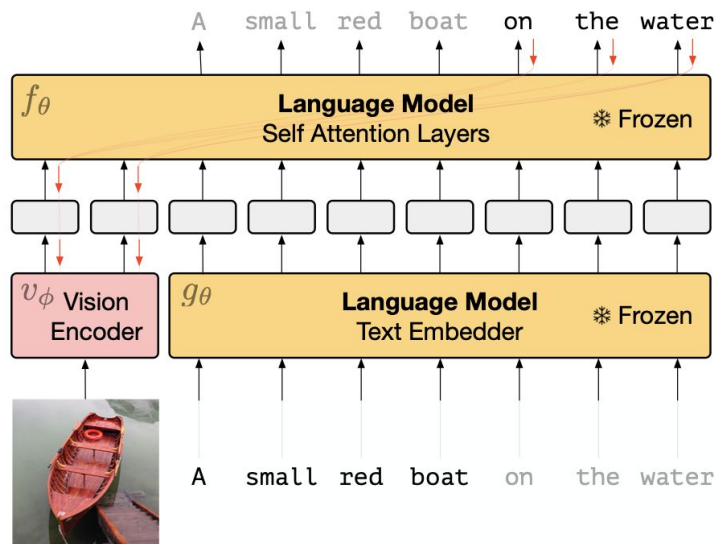


Claude can make mistakes. Please double-check responses.

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.

```
#!/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";
}
```

Welcome to the new Bing

Your AI-powered answer engine

- Ask complex questions: "What are some meals I can make for my picky toddler who only eats orange-colored food?"
- Get better answers: "What are the pros and cons of the top 3 selling pet vacuums?"
- Get creative inspiration: "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...

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.* ✓

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 was a prodigy of luxury and privilege, surrounded by opulence and excess. But one day, his parents abandoned him, casting him out of the royal court and onto the streets.

He eventually realized that he was a scavenger in a hell dimension. Without hesitation, he determined to face whatever challenges lay ahead. Inside, his character was tested to his limits. He was forced to fight for his life against 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.

```
#!/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 {
  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";
}
```

Resource Intensive

Welcome to the new Bing
Your AI-powered answer engine

- Ask complex questions: "What are some meals I can make for my picky toddler who only eats orange-colored food?"
- Get better answers: "What are the pros and cons of the top 3 selling pet vacuums?"
- Get creative inspiration: "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 on our site.

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.* ✓

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:

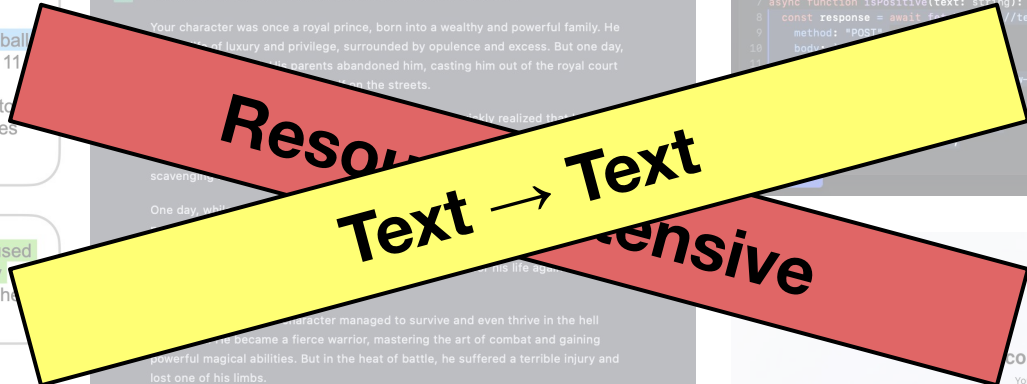
Your character was once a royal prince, born into a wealthy and powerful family. He lived a life of luxury and privilege, surrounded by opulence and excess. But one day, a catastrophic event occurred, and his parents abandoned him, casting him out of the royal court and onto the streets.

Struggling to survive, he scavenged for food and shelter, eventually realizing that his only chance of survival was to embrace his inner darkness. 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.

```
sentiments.txt write_api.go parse_expenses.py addresses.rb
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("https://text-processing.com/api/sentiment/", {
9     method: "POST",
10    body: JSON.stringify({text}),
11    headers: {
12      "Content-Type": "application/x-www-form-urlencoded",
13    },
14  });
15 }
```



Cherylin 55

Welcome to the new Bing
Your AI-powered answer engine

- 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 on our site.

Ask me anything...

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

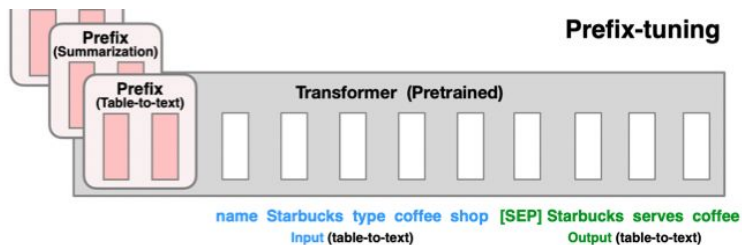
Large Language Models are Zero-Shot Reasoners (Kojima et al., 2022)

ChatGPT (OpenAI, 2022)

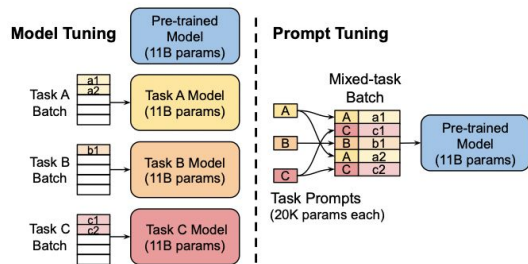
Copilot (GitHub, 2021)

Bing (Microsoft, 2023)

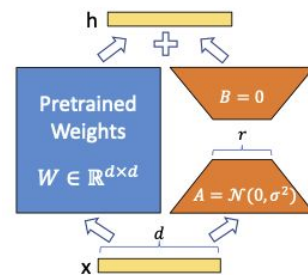
Parameter Efficient Adaptation



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.

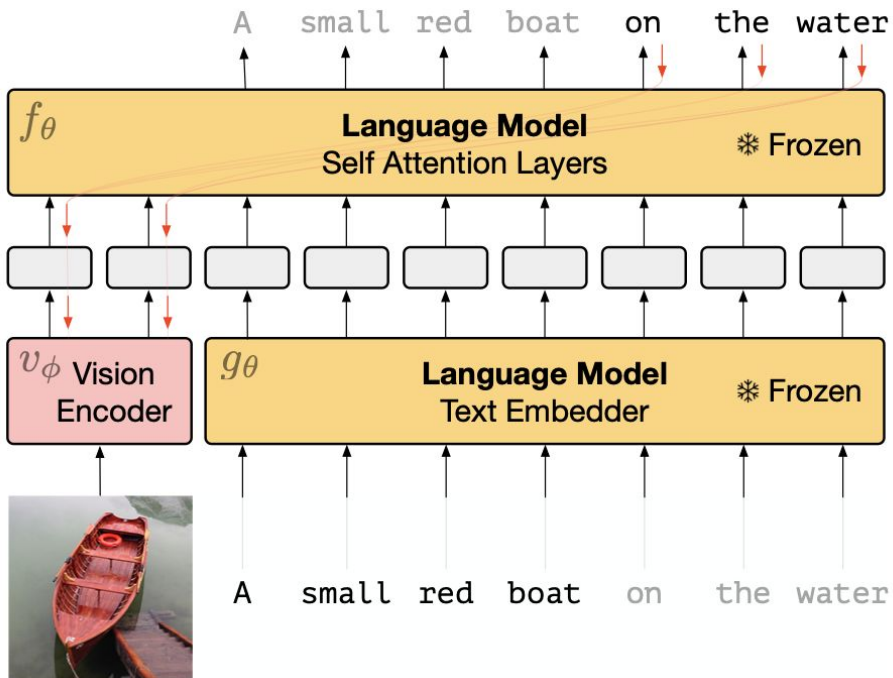


Low-Rank Adaptation (LoRA): Injects trainable rank decomposition matrices into each Transformer layer of a pretrained model.

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)

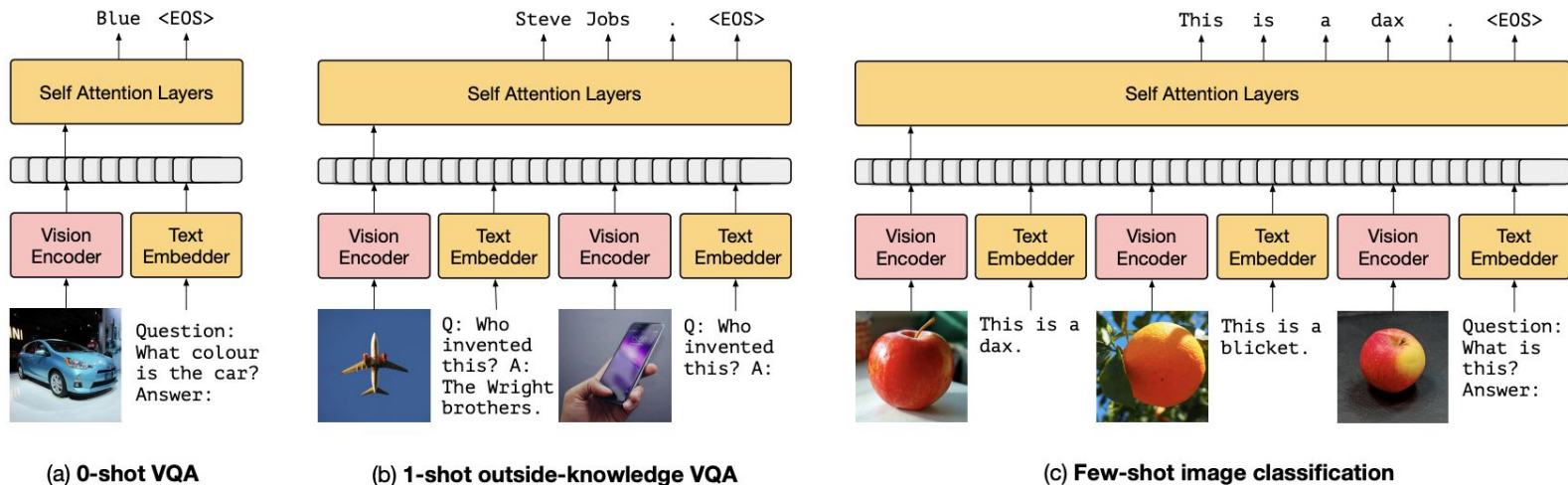


Figure 3: Inference-Time interface for *Frozen*. The figure demonstrates how we can support (a) visual question answering, (b) outside-knowledge question answering and (c) few-shot image classification via in-context learning.

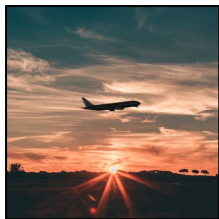


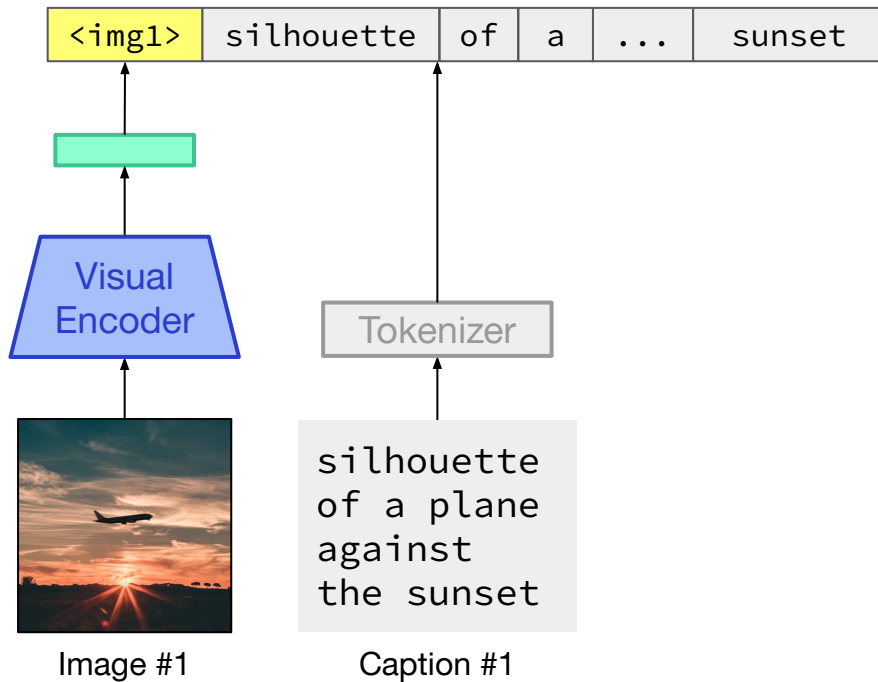
Image #1

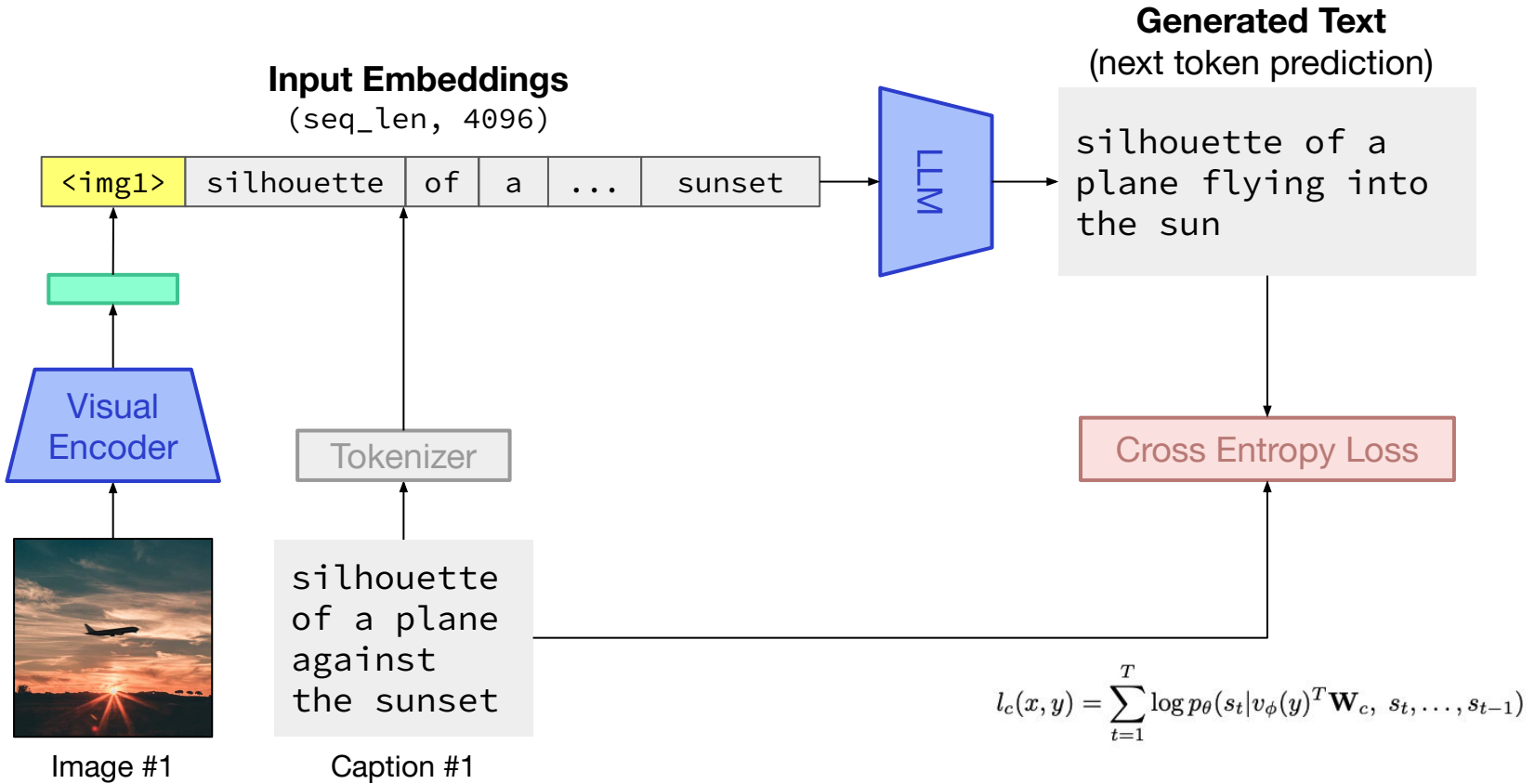
silhouette
of a plane
against
the sunset

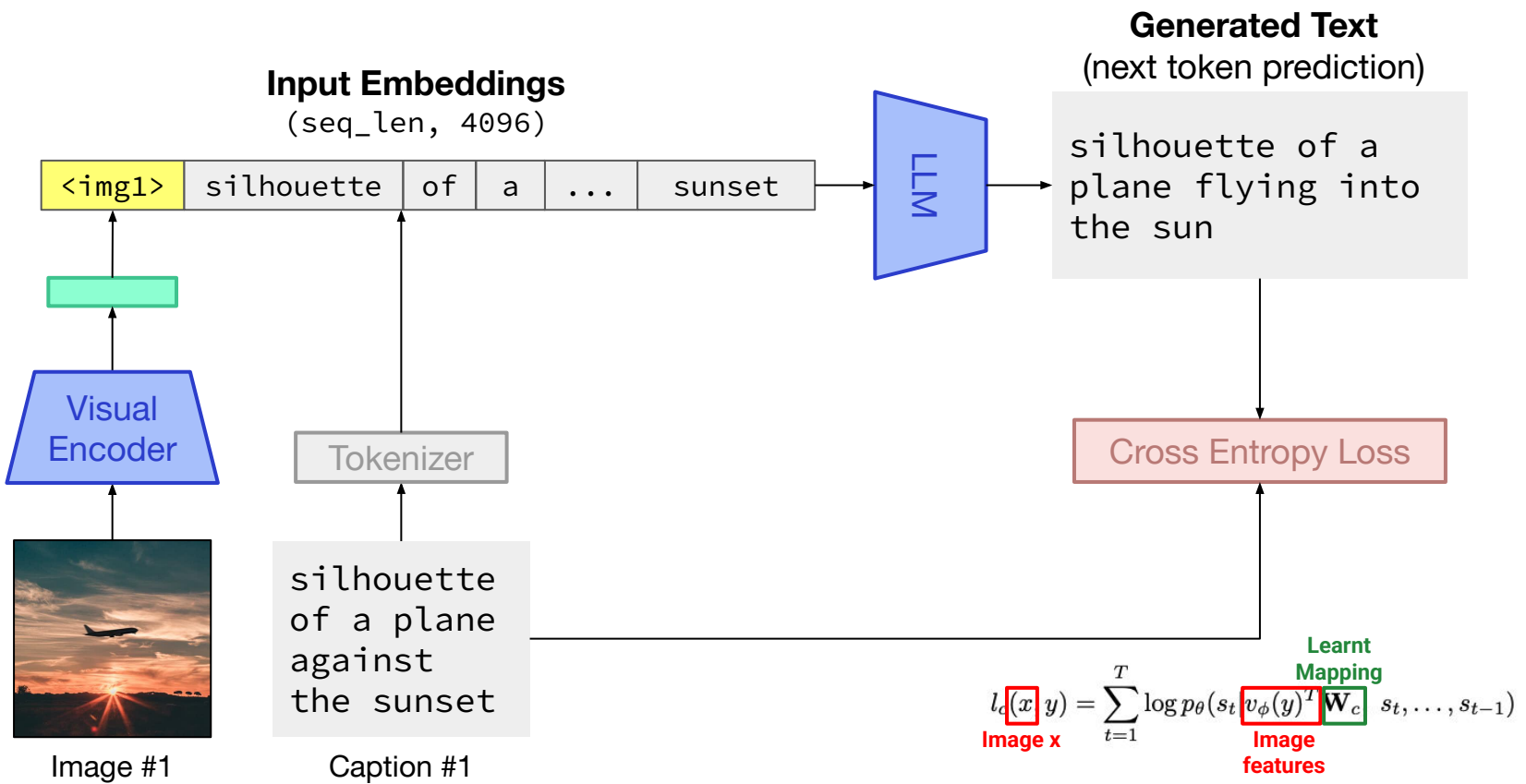
Caption #1

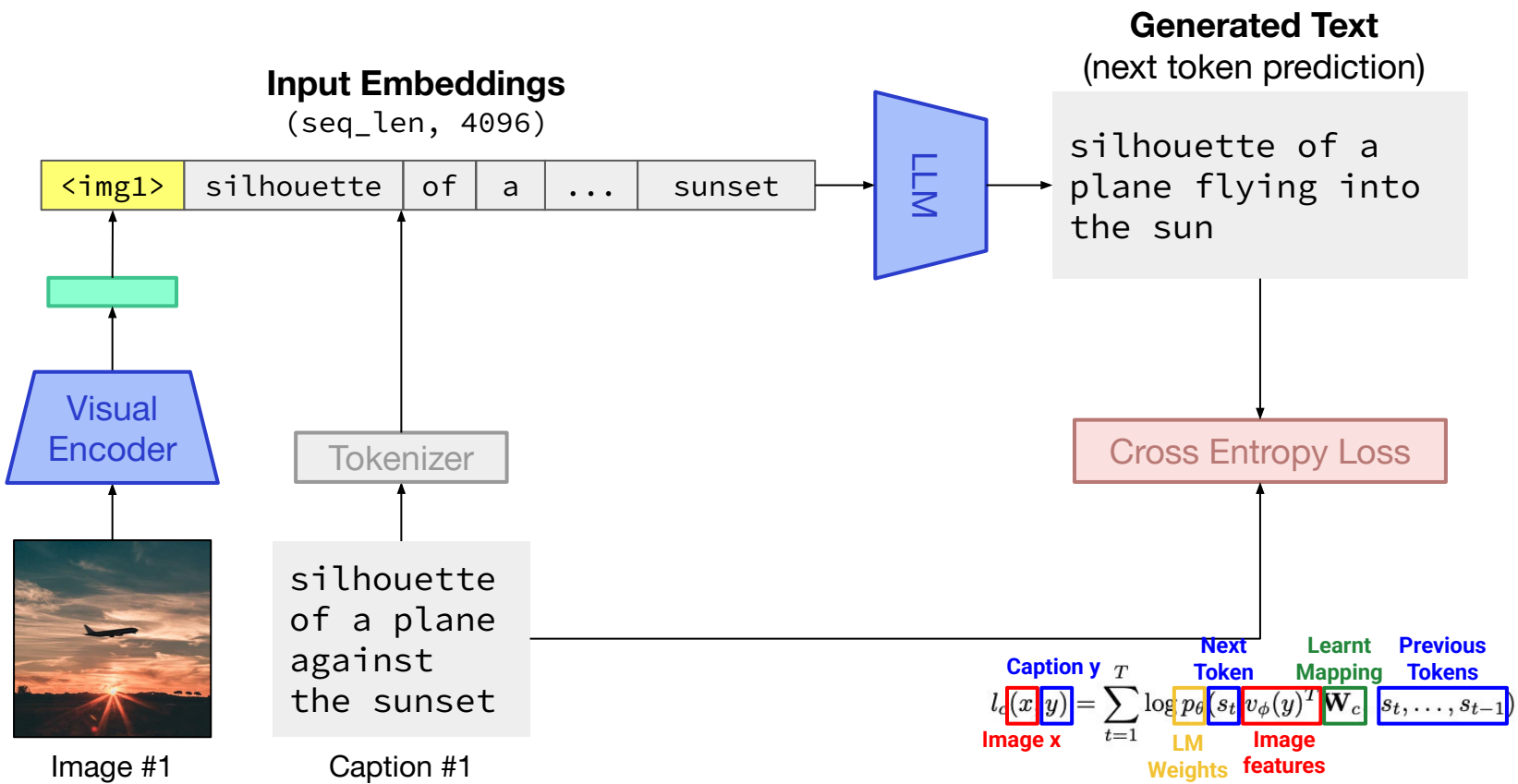
Input Embeddings

(seq_len, 4096)









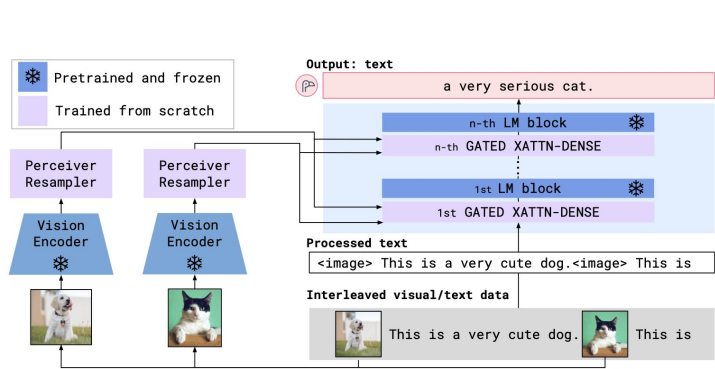
Why does this work?



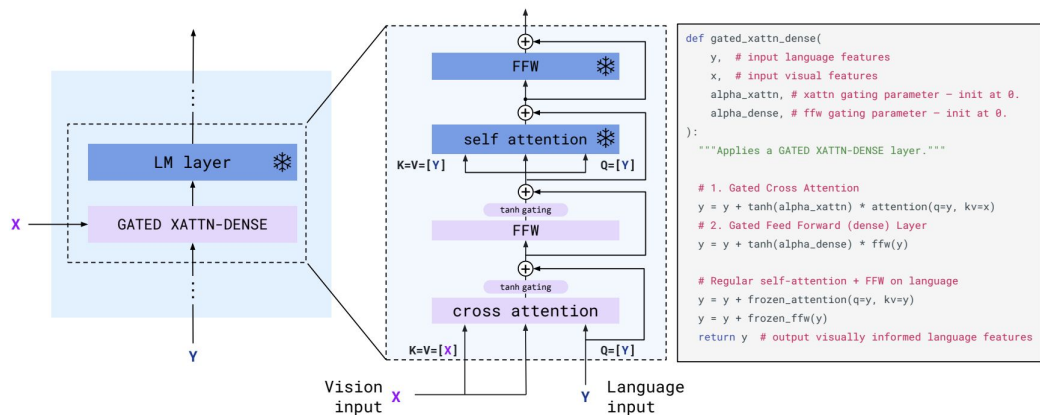
Figure 1: We train linear projections from image representations into the input space of a language model to produce captions describing images. We find that LMs can describe the contents of most image representations, but performance varies based on the type of image encoder used.

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)

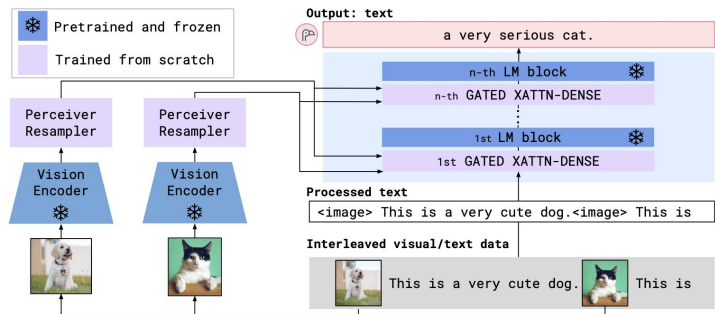


Flamingo: Finetunes new cross-attention layers on top of a 70B LLM. Achieves SOTA on many multi-modal tasks.



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

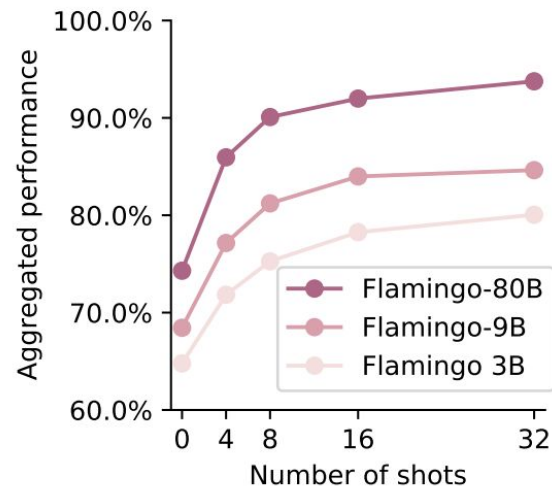
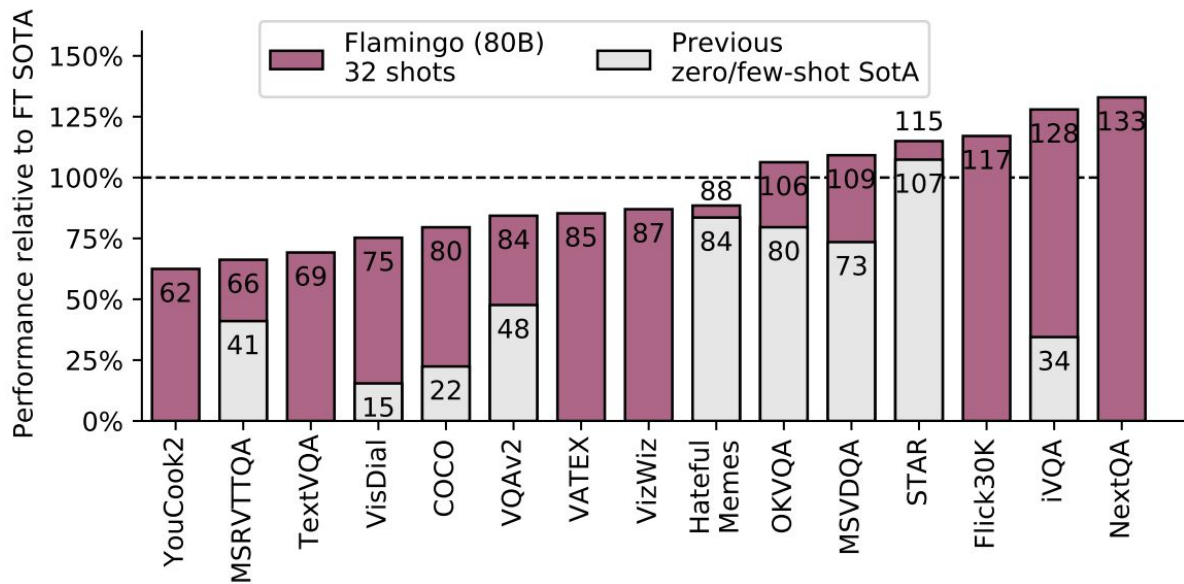
Multimodal LMs: Flamingo (2022)




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

Input Prompt				Completion
	This is a chinchilla. They are mainly found in Chile.		This is a shiba. They are very popular in Japan.	a flamingo. They are found in the Caribbean and South America.
	This is			
	What is the title of this painting? Answer: The Hallucinogenic Toreador.		Where is this painting displayed? Answer: Louvres Museum, Paris.	
	What is the name of the city where this was painted? Answer:			Arles.
	Output: "Underground"		Output: "Congress"	"Soulomes"
	2+1=3		5+6=11	3x6=18
	Output: A propaganda poster depicting a cat dressed as French emperor Napoleon holding a piece of cheese.		Output: A pink room with a flamingo pool float.	A portrait of Salvador Dali with a robot head.
	Les sanglots longs des violons de l'automne blessent mon coeur d'une langueur monotone.		Pour qui sont ces serpents qui sifflent sur vos têtes?	Je suis un cœur qui bat pour vous.

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.

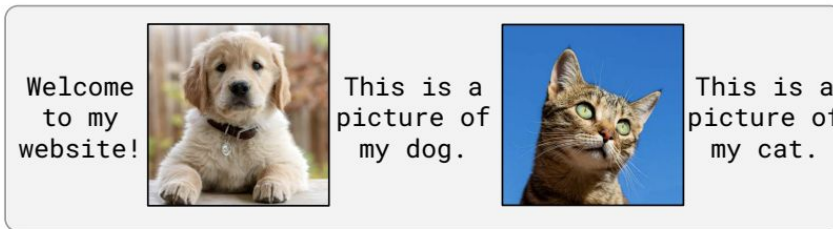
Data: Flamingo (2022)



Image-Text Pairs dataset



Video-Text Pairs dataset



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)

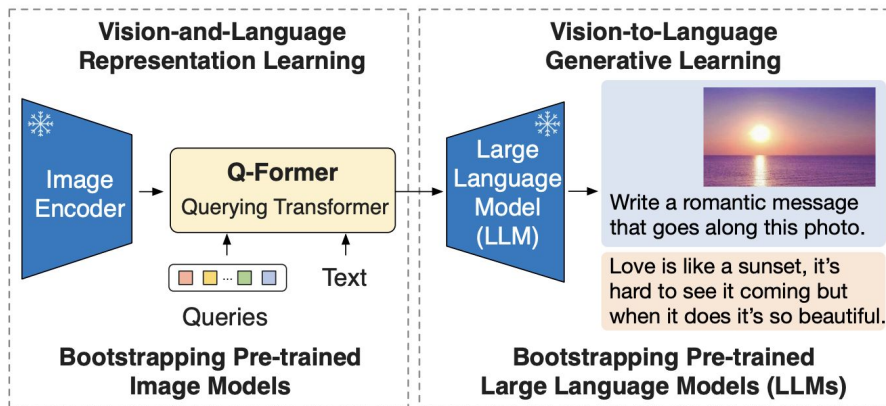
Video-Text Pairs:

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)




Explain the advantages of this product.

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.




Tell me something about the history of this place.

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.



Write down the facts that you know about this flower.

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.



Is this photo unusual?

Yes, it's a house that looks like it's upside down.

How could someone get out of the house?

It has a slide on the side of the house.




What are shown in the photo?

A man and a chicken.

What does the man feel and why?

He is scared of the chicken because it is flying at him.



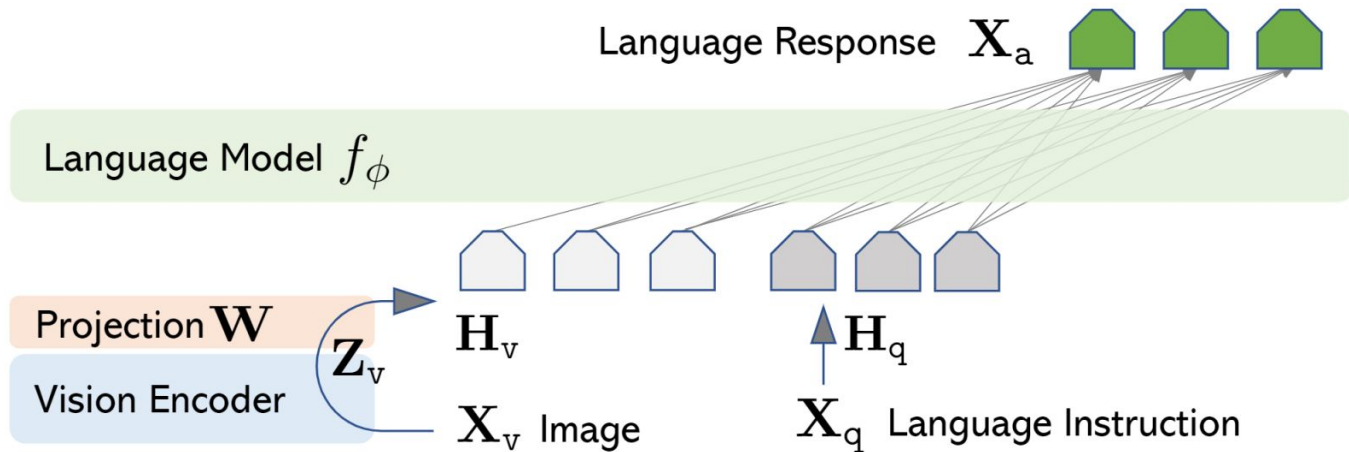
What are the ingredients I need to make this?

Pepperoni, mozzarella cheese, pizza sauce, olive oil, salt, pepper, basil.

What is the first step?

Place the pizza dough on a baking sheet, brush with olive oil, sprinkle with salt, pepper, and basil.

Multimodal LMs: LLaVA (2023)

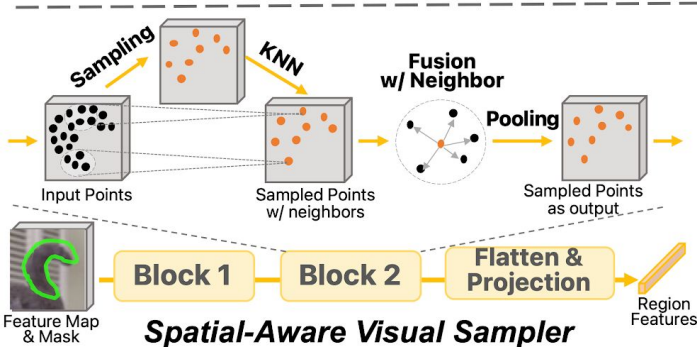
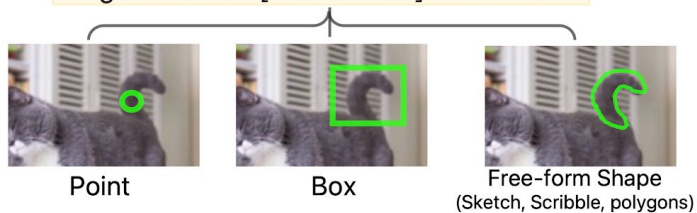


LLaVA: Finetunes a linear layer (W) over a frozen vision encoder and a frozen LLM. Showcases strong performance by finetuning on paired data of images and text instructions (some GPT-4 generated).

Multimodal LMs: Ferret (2023)

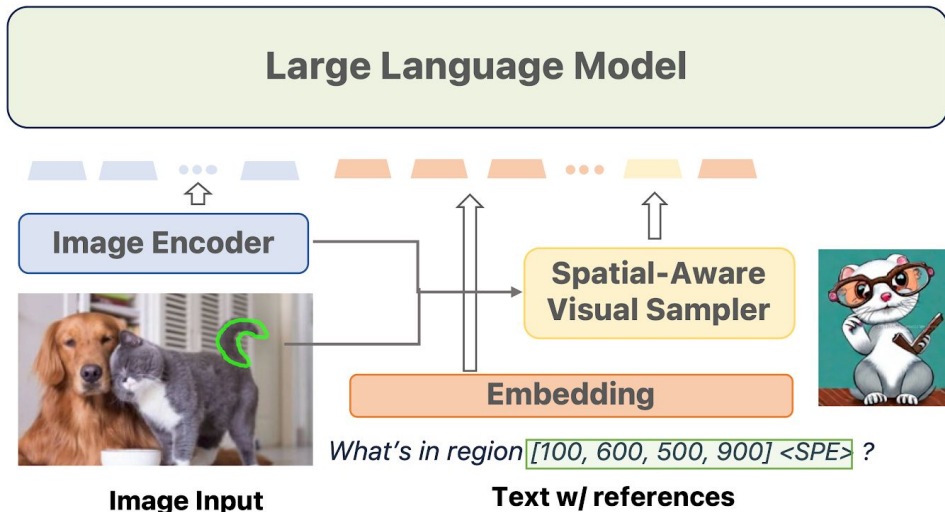
Hybrid Region Representation

Region Name + [Coordinates] + <feature>



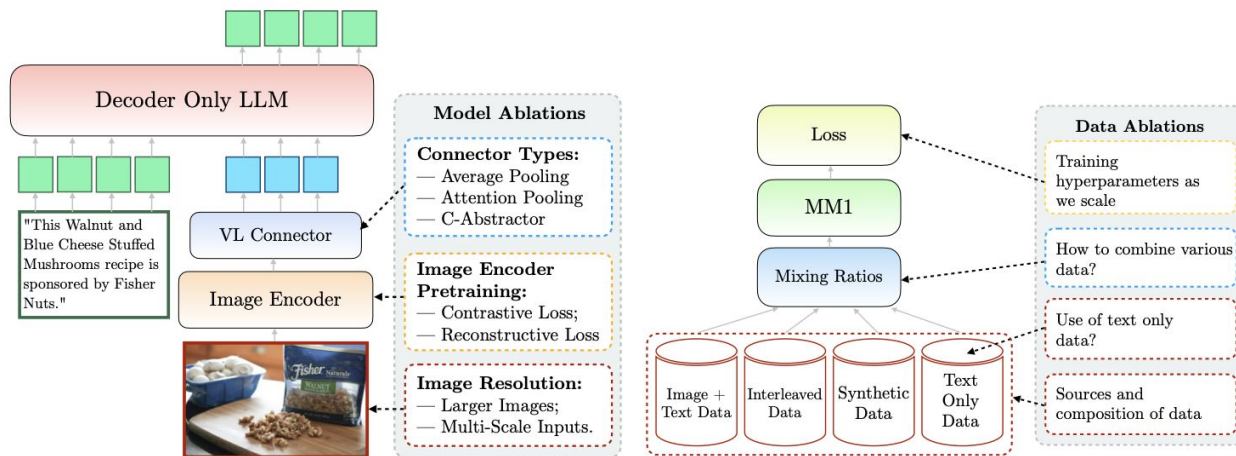
Ferret Model

It's a cat tail [80, 590, 450, 920]



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)

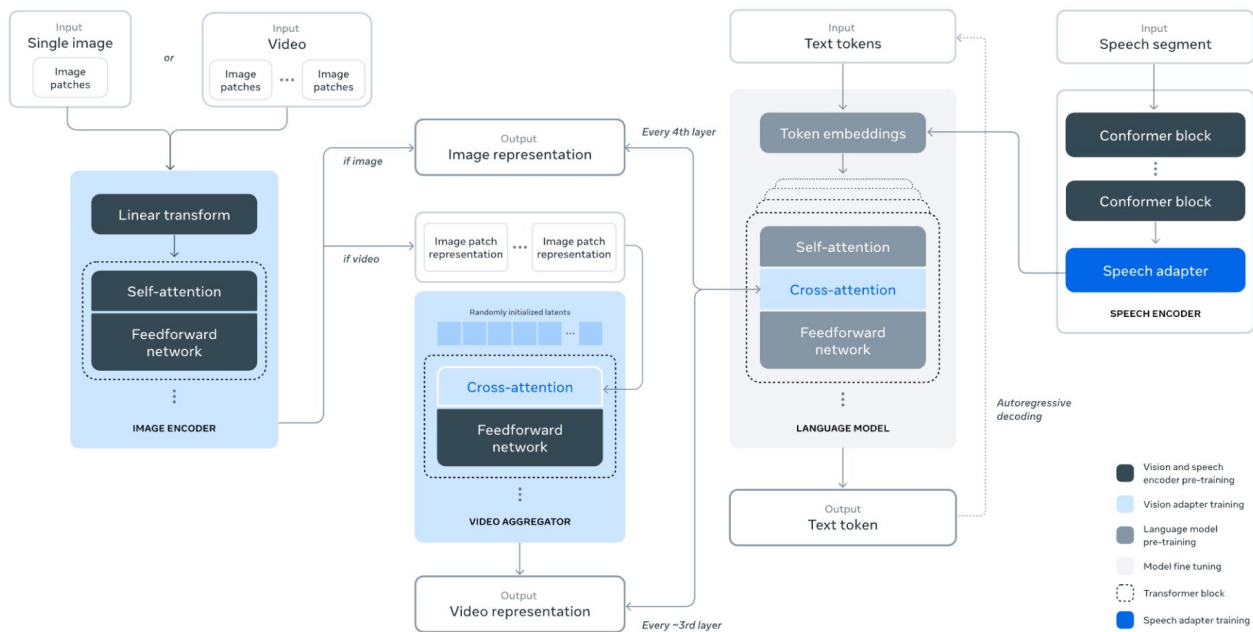


Figure 28 Illustration of the compositional approach to adding multimodal capabilities to Llama 3 that we study in this paper. This approach leads to a multimodal model that is trained in five stages: **(1)** language model pre-training, **(2)** multi-modal encoder pre-training, **(3)** vision adapter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

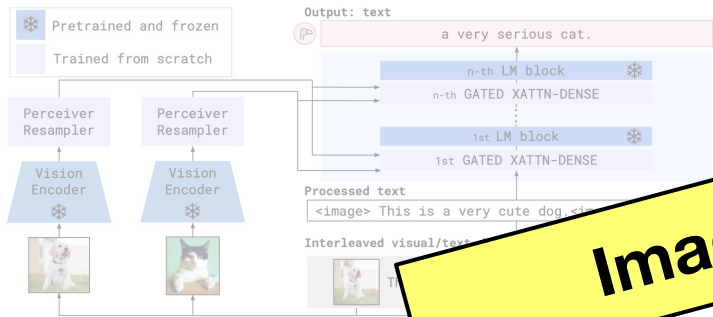
- **Annealing**

- 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



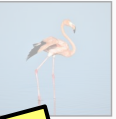

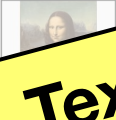



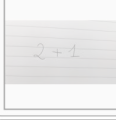
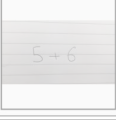
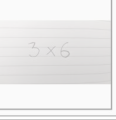

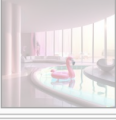
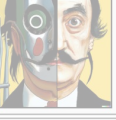
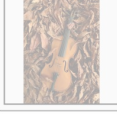
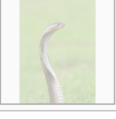
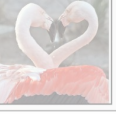
- **Closed models:** gpt-5, Gemini-3, Claude 4.6
 - The general public is not 100% sure what they do :)
- **Open sourced models:**
 - [Qwen3.5](#)
 - [Qwen3-VL](#)
 - [Molmo 2](#)
 - [Kimi-K2.5](#) (Huge model! Requires multiple 8 GPU nodes to run)

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

Input Prompt					Completion	
	This is a chinchilla. They are mainly found in Chile.		This is a shiba. They are very popular in Japan.		This is	a flamingo. They are found in the Caribbean and South America.
	What is the title of this painting? Answer: The Hallucinations		What is the name of the city where this was painted? Answer:			Arles.
	This is a very cute dog.		Output: "Congress"		Output:	"Soulomes"
	2+1=3		5+6=11			3x6=18
	Output: A propaganda poster depicting a cat dressed as French emperor Napoleon holding a piece of cheese.		Output: A pink room with a flamingo pool float.		Output:	A portrait of Salvador Dali with a robot head.
	Les sanglots longs de l'automne blessent mon coeur d'une langueur monotone.		Pour qui sont ces serpents qui sifflent sur vos têtes?			Je suis un cœur qui bat pour vous.

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



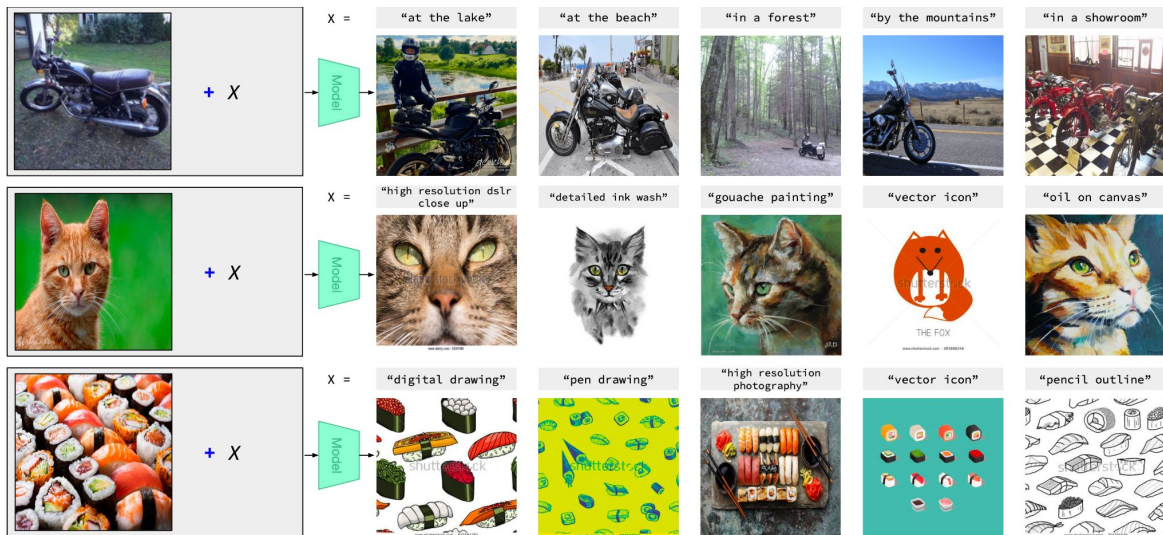
Frozen Retrieval Over Multimodal Data for Autoregressive Generation

jykoh.com/fromage

FROMAGe

Frozen Retrieval Over Multimodal Data for Autoregressive Generation

jykoh.com/fromage

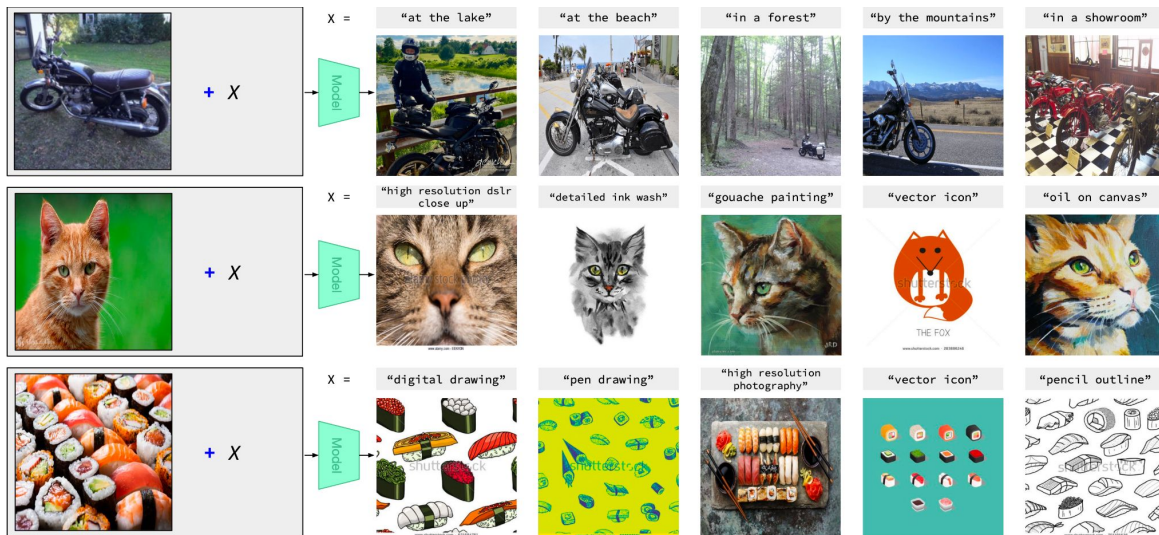


Concept composition. Our model is capable of retrieving relevant images conditioned on multi-modal context inputs.

FROMAGe

Frozen Retrieval Over Multimodal Data for Autoregressive Generation

jykoh.com/fromage



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)

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

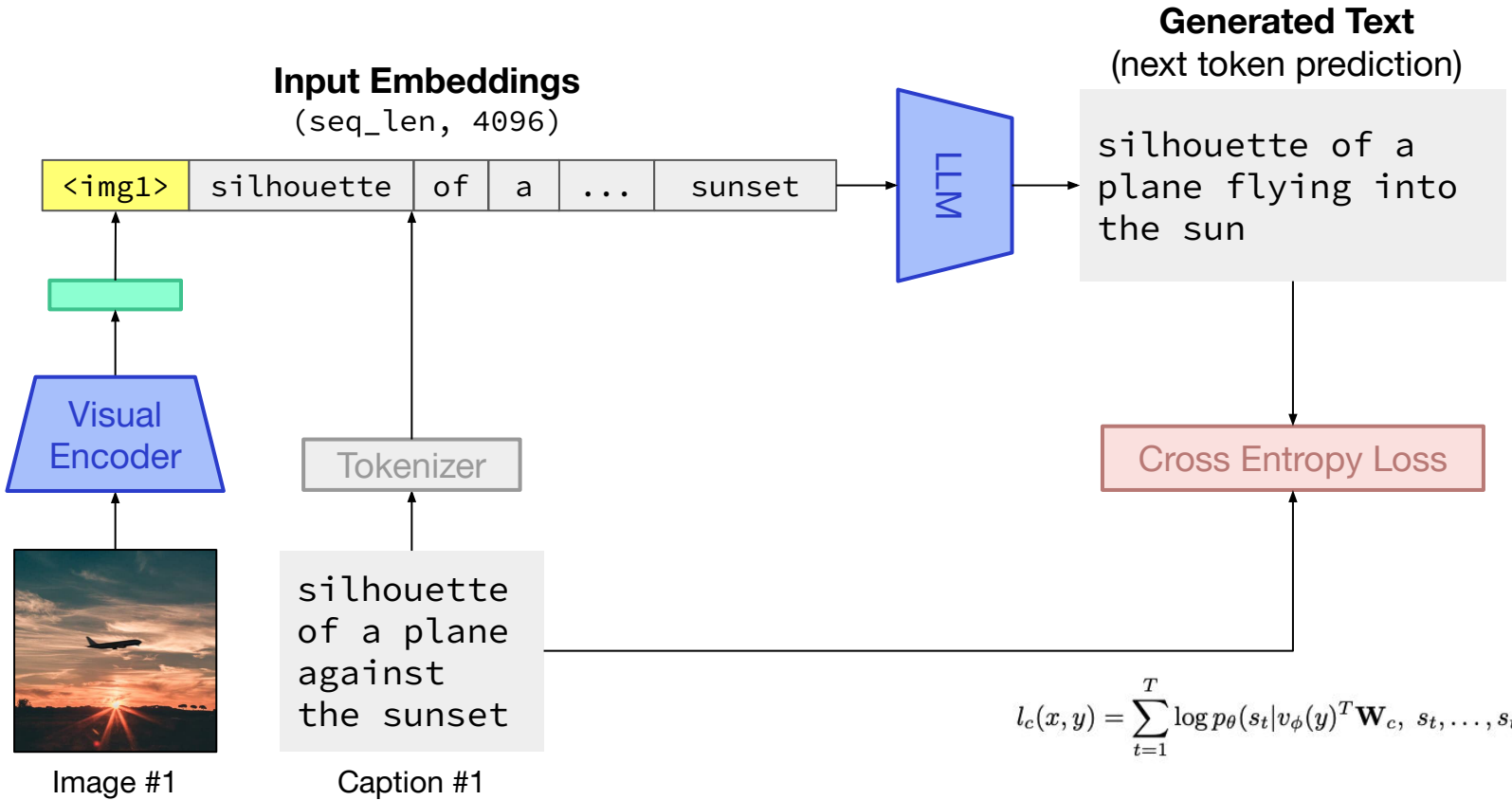
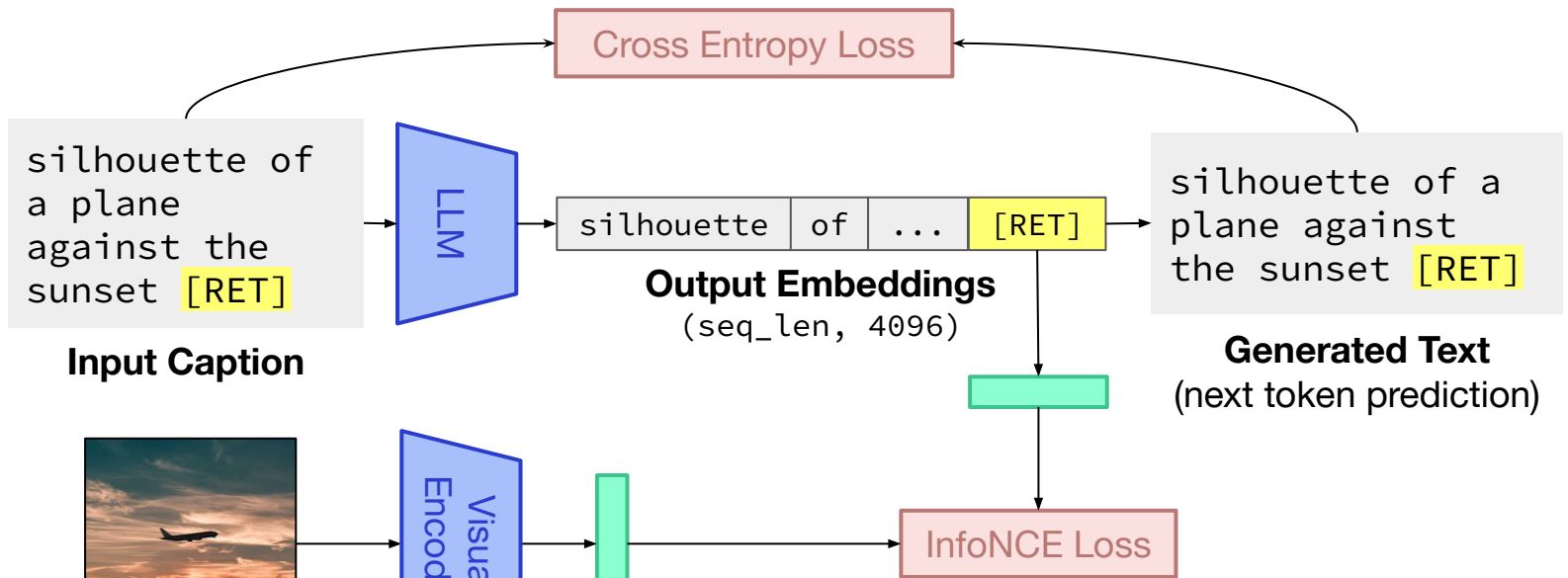


Image #1

Caption #1

Image Captioning



Input Image

Input Caption

Output Embeddings
(seq_len, 4096)


Generated Text
(next token prediction)


Image-Text Retrieval


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


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


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. 

Multi-Modal Dialogue

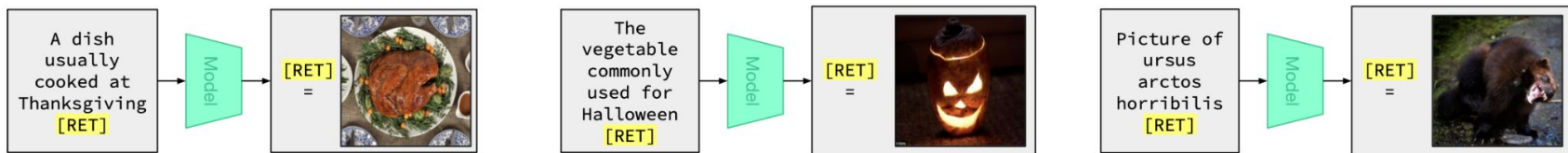
Green bubbles indicate model generated outputs, grey bubbles indicate user provided prompts.

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!)

2 <Link href="https://cdn.jsdelivr.net"
3 <body class="bg-gray-100 font-sans
4 <header class="bg-cover bg-cen
5 <nav class="container mx-aut
6 <div class="flex-1"
7 <a class="font-semib
8 Vertice
9
10 </div>
11 <div>
12 <ul class="flex"
13 <a class="te
14 <a class="te
15 <a class="te
16 <a class="te
17 <a class="te
18
19 </div>
20 </nav>
21 <div class="container mx-aut
22 <h1 class="text-5xl font
23 Unlock Your Potential
24 </h1>
25 <p class="text-white tex
26 Bringing your imagin
27 </p>
28 <div class="mt-12">
```

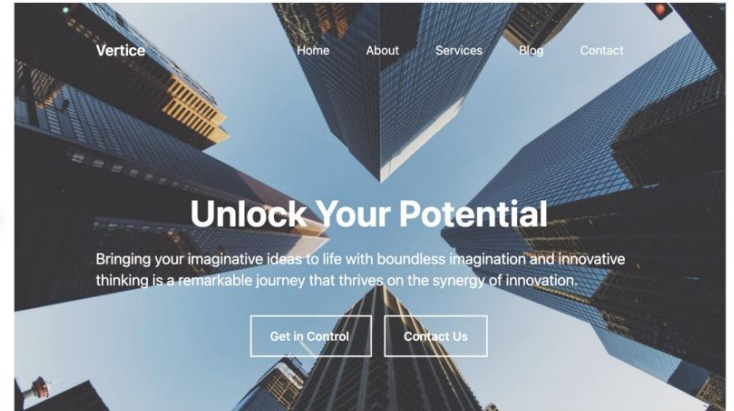
Sightseer  
generation



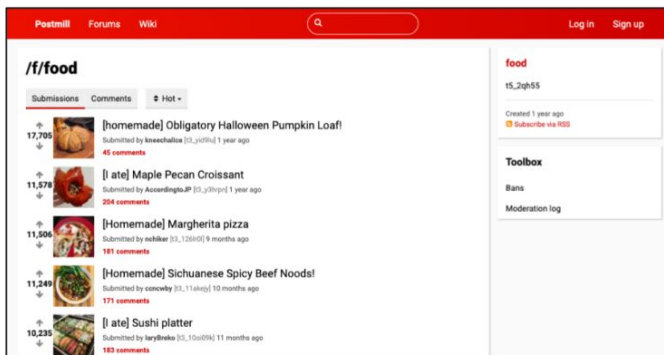
Rendering



Rendered generated code



# Future Research: Multimodal Agents

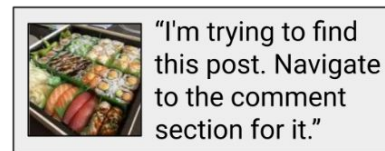


Original Webpage

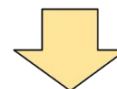
**Webpage with SoM of Interactable Elements**

```
...
[7] [A] [Comments]
[8] [BUTTON] [Hot]
[9] [IMG] [description: picture of a pumpkin]
[10] [A] [kneeçalice]
...
```

**SoM Elements and Text Content**



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 good Thai restaurant in Pittsburgh. It should have at least 200 reviews and 4.3 stars. Pick the one with the highest rating." \
 --result_dir demo_test_yelp \
 --model gpt-4-vision-preview \
 --action_set_tag som --observation_type 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_ty
pe 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](mailto:jykoh@cmu.edu)