

How to AI (Almost) Anything

Lecture 9 – Large Multimodal Models

Paul Liang

Assistant Professor

MIT Media Lab & MIT EECS



<https://pliang279.github.io>

ppliang@mit.edu

 [@pliang279](#)



Assignments for This Coming Week

For project:

- Make sure to meet with myself and TAs this week
- Medium progress towards implementing new ideas. Either promising results or poor results, but a good idea of what is wrong and how to fix.

Reading assignment due tomorrow Wednesday (4/16).

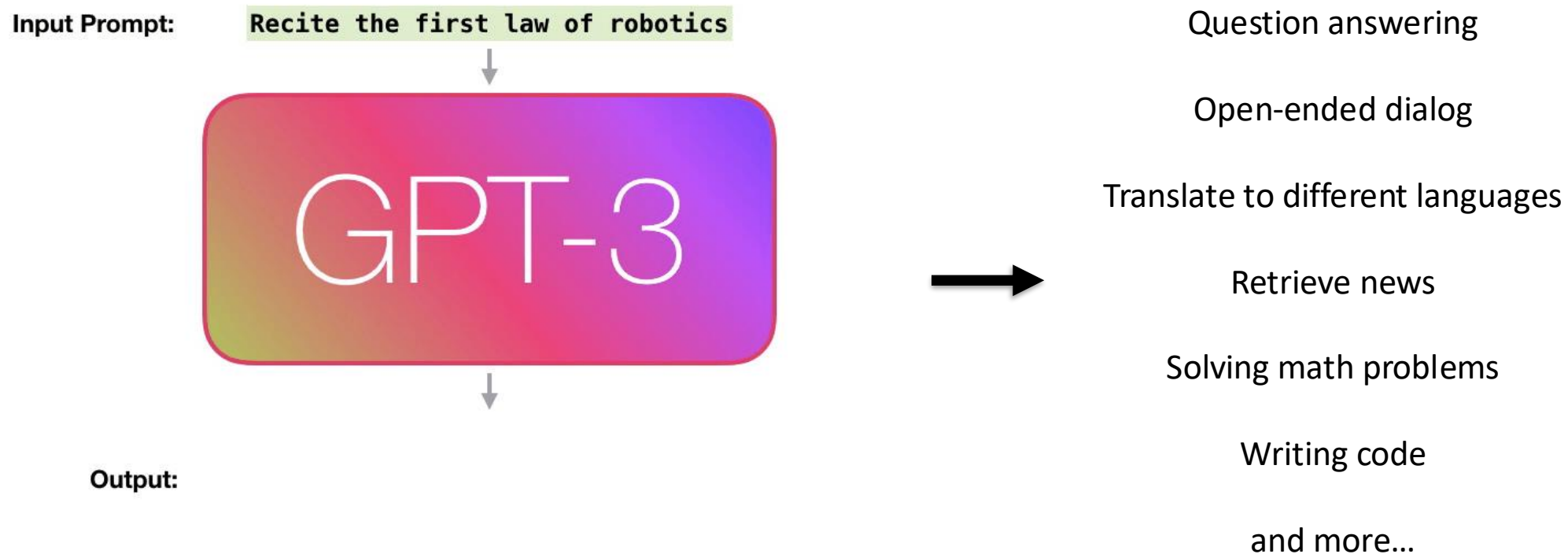
This Thursday (4/17): fifth reading discussion on **large language models**.

1. Alignment faking in LLMs
2. Mathematical reasoning in LLMs

Today's lecture

- 1 Multimodal foundation models and pre-training
- 2 Adapting LLMs into multimodal LLMs
- 3 From text to multimodal generation
- 4 Latest directions

Recap: Large Language Models



From Large Language Models to Multimodal Models



Classification: What is the tone of the man in the grey shirt?

Open-ended: Describe the relationships between these 2 people.

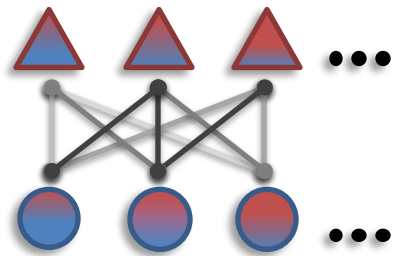
Explanation: Explain why, citing visual and verbal evidence.

Generation: Animate a story inspired by this short video clip.

Counterfactual: What if these people were from a different society or culture?

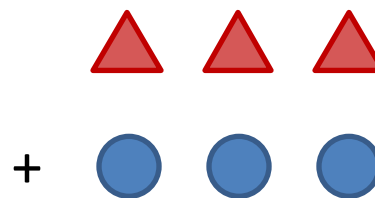
From Large Language Models to Multimodal Models

*It's just a privilege to
watch your mind at work.*



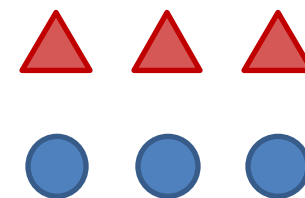
Multimodal
representation

*This person is being sarcastic.
They seem to be close friends.*



*(quote previous episodes)
(highlight multimodal information)*

*Here's a story of them in
a different culture...*

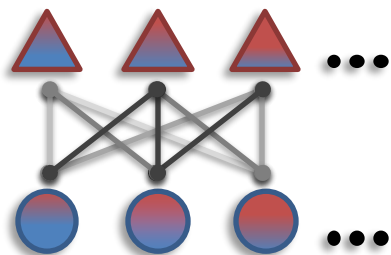


*(generate future
episodes)*

Lecture outline

Part 1: Multimodal foundation model representations of text, video, audio

*It's just a privilege to
watch your mind at work.*



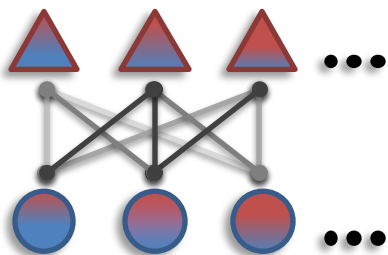
Multimodal
representation



Lecture outline

Part 2: Adapting large language models for multimodal text generation

*It's just a privilege to
watch your mind at work.*



Multimodal
representation



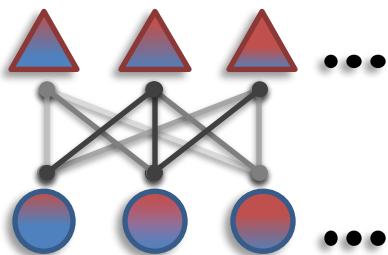
*This person is being sarcastic.
They seem to be close friends.*



Lecture outline

Part 3: Enabling text and image generation

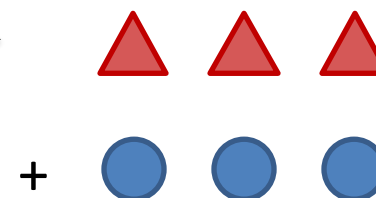
*It's just a privilege to
watch your mind at work.*



Multimodal
representation



*This person is being sarcastic.
They seem to be close friends.*



*(quote previous episodes)
(highlight multimodal information)*

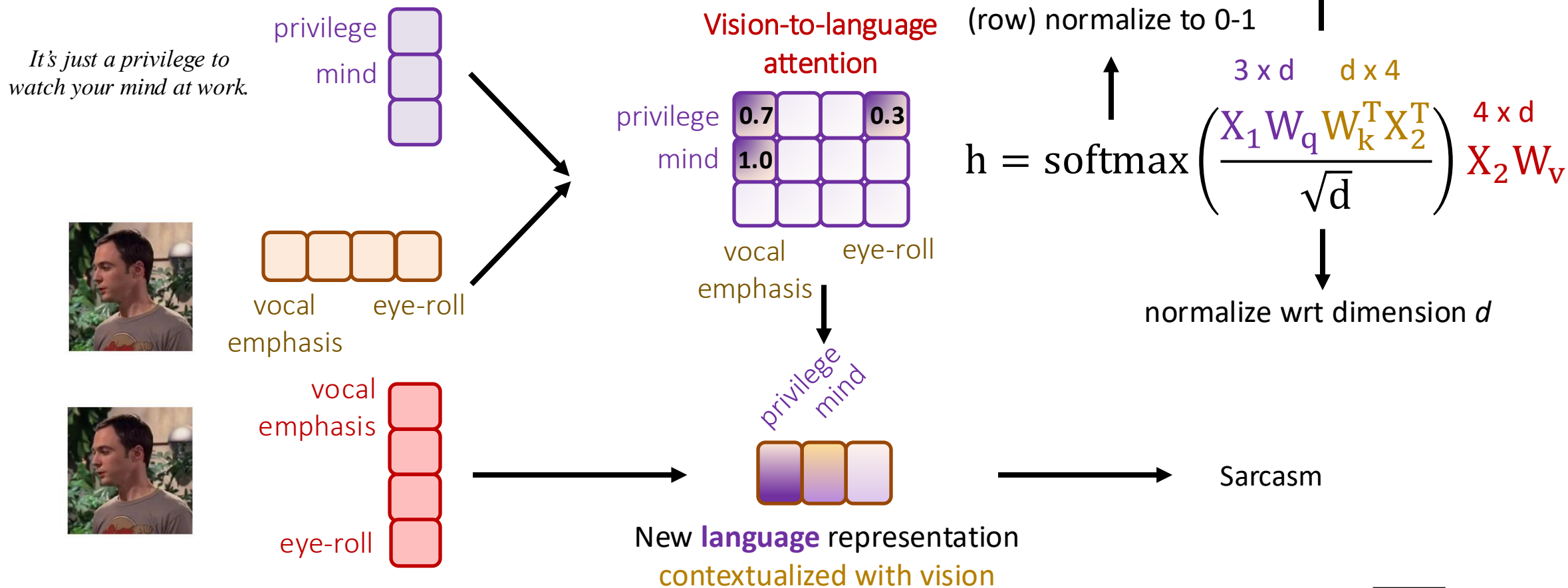
Vision Transformers



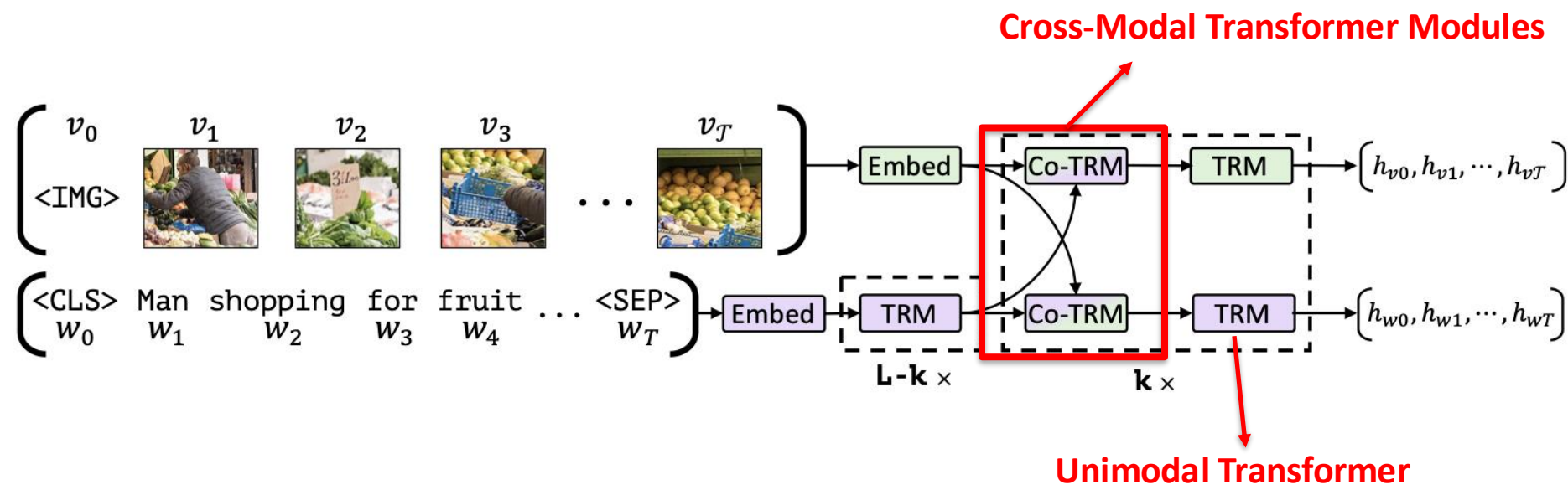
[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

[Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2020]

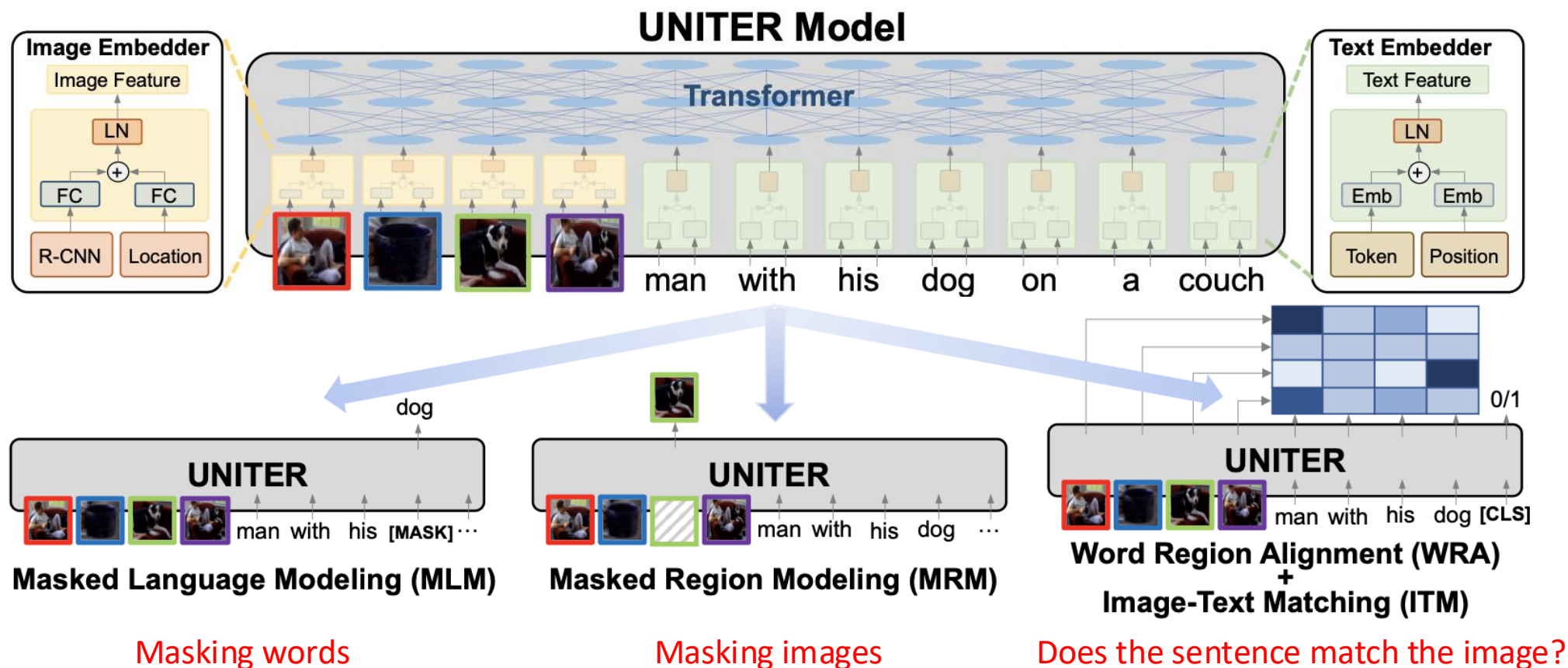
Multimodal Transformers



Multimodal Cross-attention Transformers



Multimodal Cross-attention Transformers



Visual-and-Language Transformer (ViLT)

Example of alignment between modalities:



a display of **flowers** growing out and over the retaining **wall** in front of **cottages** on a **cloudy** day.



flowers



wall



cottages



cloudy



a room with a **rug**, a **chair**, a **painting**, and a **plant**.



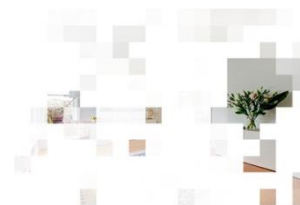
rug



chair



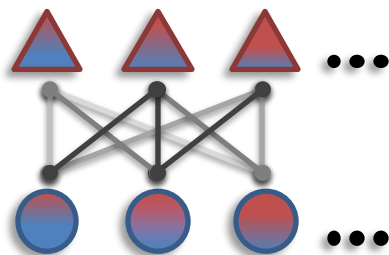
painting



plant

Adapting Large Language Models to Multimodal

*It's just a privilege to
watch your mind at work.*



Multimodal
representation



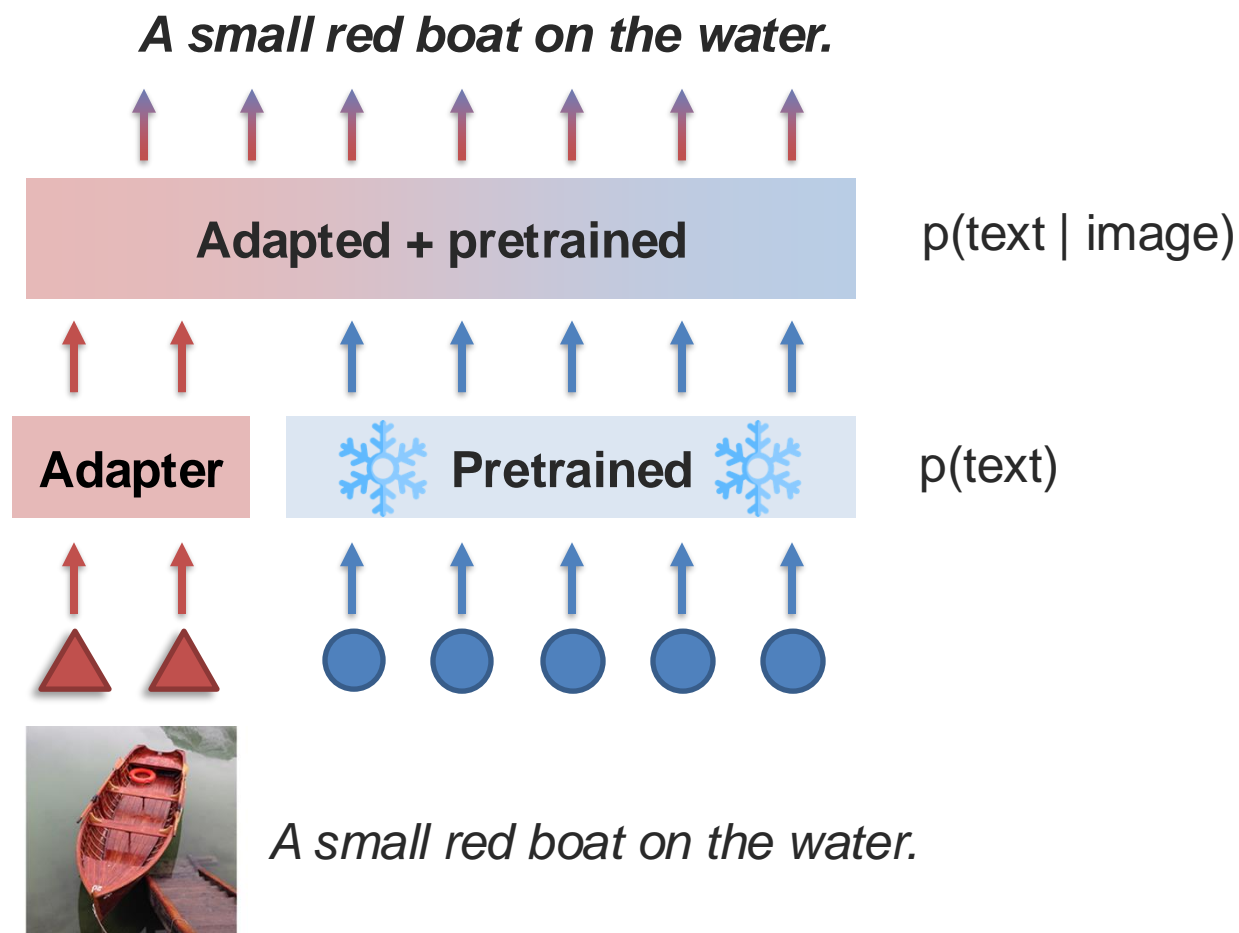
*This person is being sarcastic.
They seem to be close friends.*



Adapting Large Language Models

Conditioning via prefix tuning

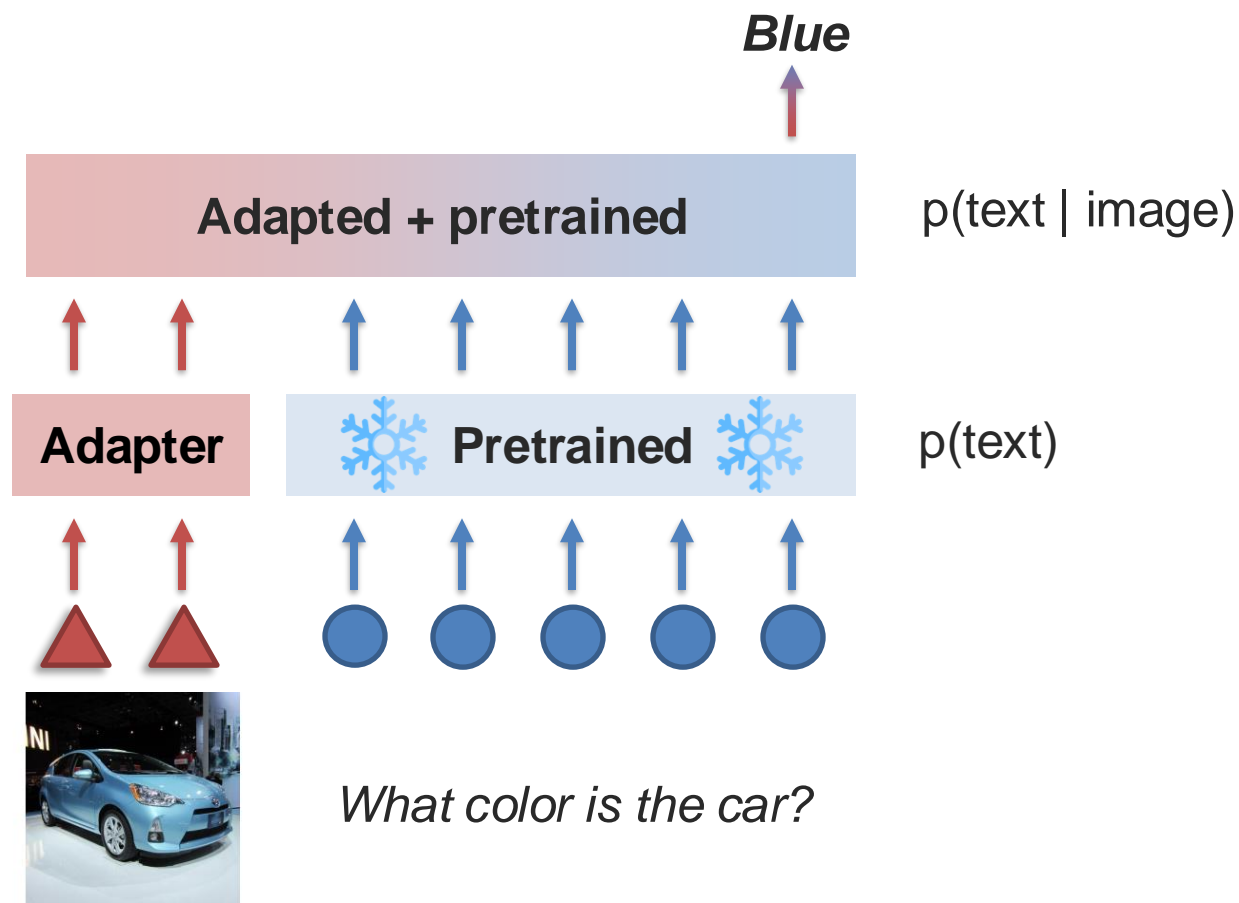
Modeling $p(\text{text} \mid \text{image})$:



Adapting Large Language Models

Conditioning via prefix tuning

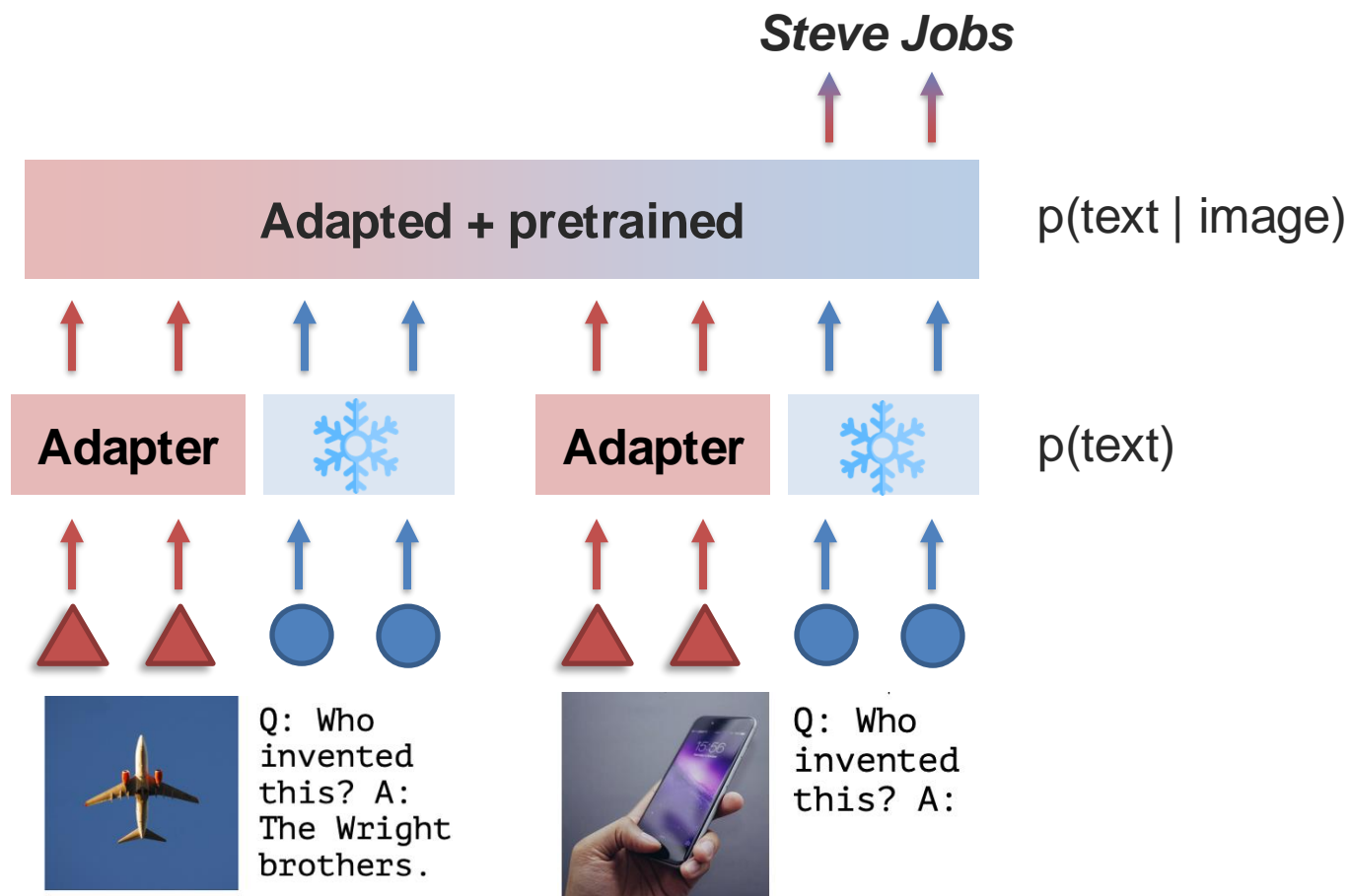
Modeling $p(\text{text} \mid \text{image})$:



Adapting Large Language Models

Conditioning via prefix tuning

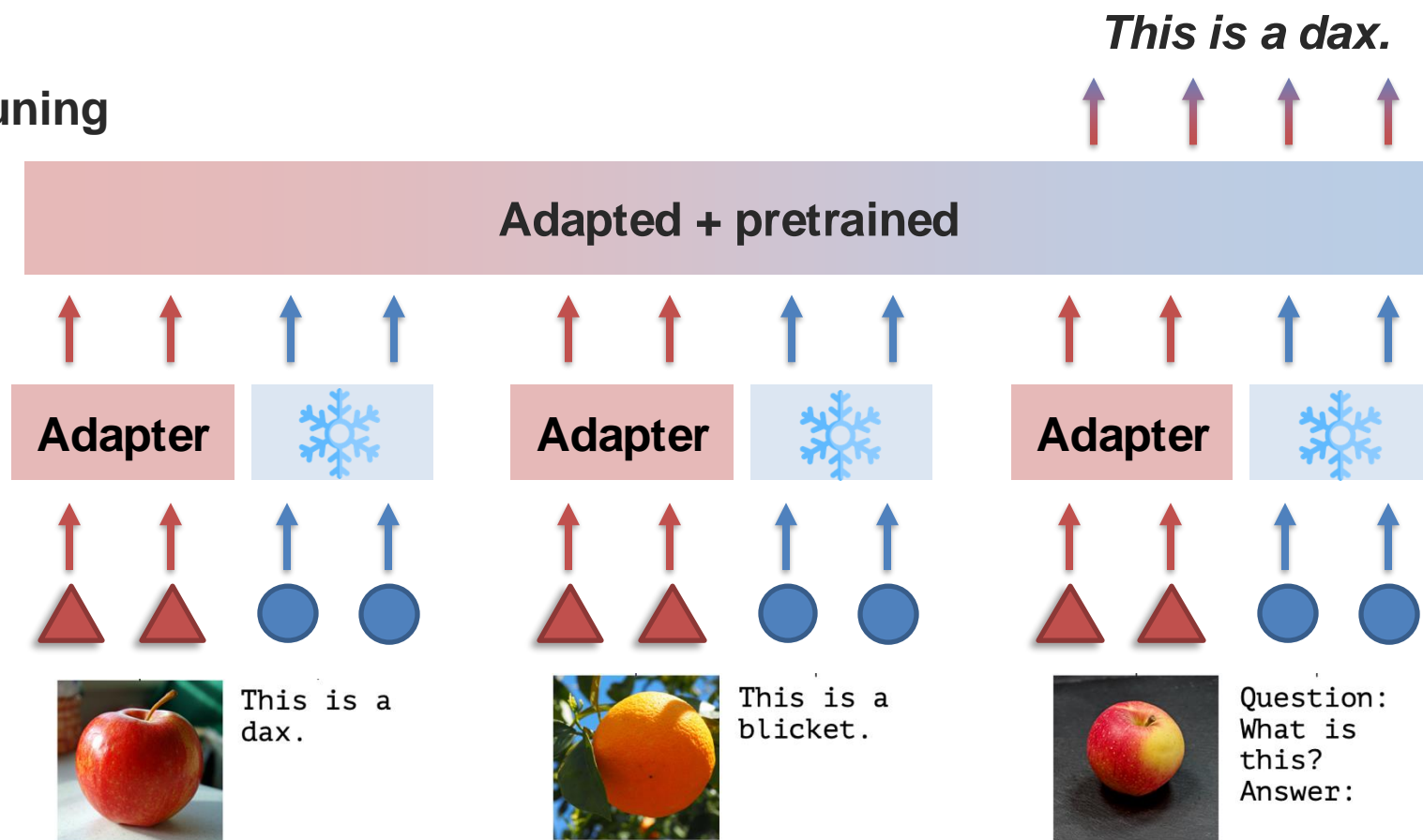
Modeling $p(\text{text} \mid \text{image})$:



Adapting Large Language Models

Conditioning via prefix tuning

Modeling $p(\text{text} \mid \text{image})$:



Adapting Large Language Models

Flamingo



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?

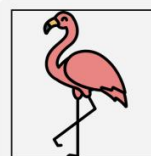


Yes, it is surprising.

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?



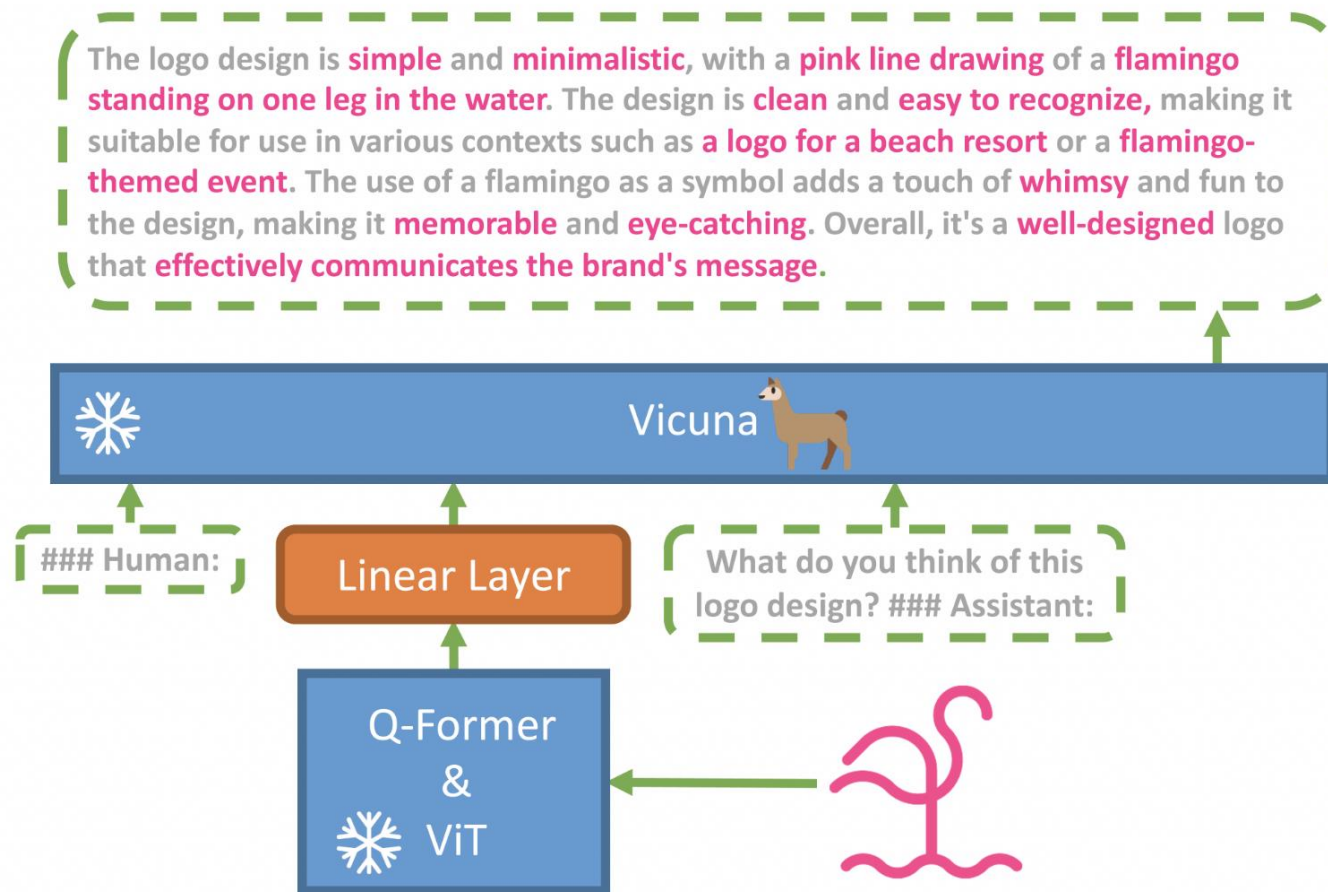
It's white.

Scaling Large Multimodal Models

Mini-GPT4

Stage 1: **Alignment** using paired image-text data.

Stage 2: **Instruction tuning** using image + text instructions and example completions.



The architecture of MiniGPT-4.

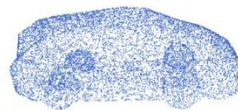
Scaling Large Multimodal Models

LLaMA-Adapter

LLaMA-Adapter:
Bilingual Multi-modality
Instruction Model



Example: 3D Point Cloud to Image (Bilingual)



Generate an image from the 3D point cloud.



Hello



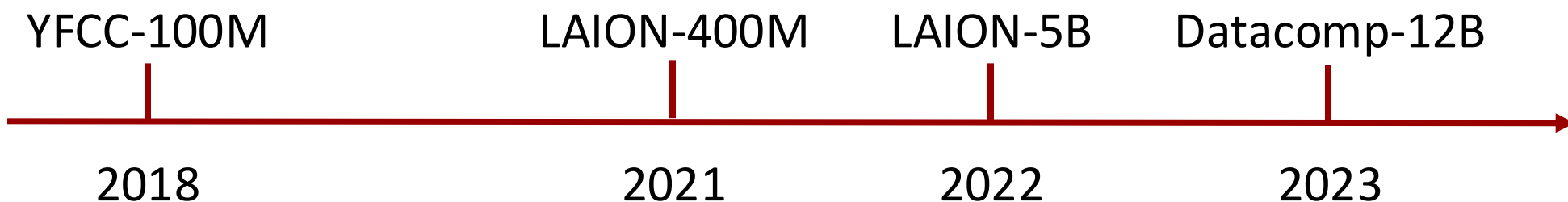
根据这个3D点云生成一张图片。

你好



Pre-training datasets

- Largest dataset is DataComp. It has 12.8 billion image-text pairs.
- Recent efforts shifted more towards filtering for high quality multimodal data. Examples include DFN (2B), COYO (600M), and Obelics (141M)



Data type	dataset	#samples	sampling prob.
Image-Caption	DFN [Fang et al., 2023]	2B	27%
	COYO [Byeon et al., 2022]	600M	11.25%
	HQITP	400M	6.75%
Interleaved Text	Obelics [Laurençon et al., 2024a]	141M Docs	45%
	DCLM [Li et al., 2024b]	6.6T Toks	10%

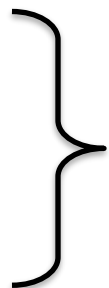
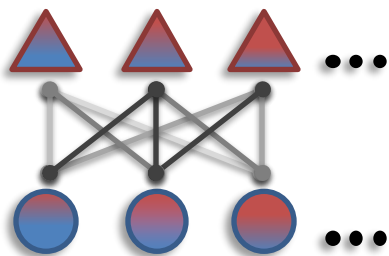
Multimodal Instruction Tuning Datasets

- More scattered, smaller in nature
- General domain: Vision-Flan (187K), LLaVA-Instruct (150K), InstructBLIP (~1.6M), M3IT (2.4M)
- Clinical: CLIMB-QA (4.51M), BioMed-VITAL (210K), LLaVA-Med (60K)

Dataset	# Tasks	Multi-Lingual	# of Instances	Avg. # of Manual Instructions / Task	Open-Sourced
MiniGPT4	N / A	✗	5K	N / A	✓
LLaVA	3	✗	1.15M	N / A	✓
MultiModalGPT	3	✗	6K	5	✗
MultiInstruct	26	✗	~ 235K	5	✗
InstructBLIP	28	✗	~ 1.6M	9.7	✗
M ³ IT (Ours)	40	✓	2.4M	10	✓

From Text to Multimodal Generation

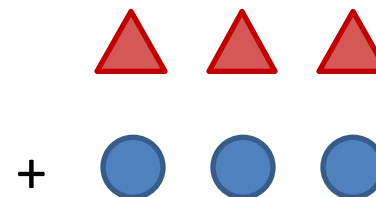
*It's just a privilege to
watch your mind at work.*



Multimodal
representation



*This person is being sarcastic.
They seem to be close friends.*



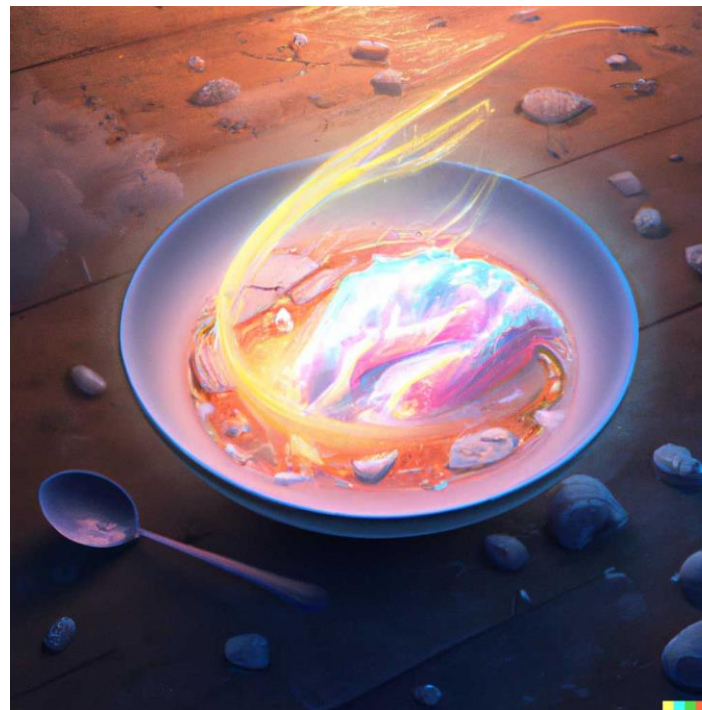
*(retrieve next episode)
(highlight multimodal evidence)*

From Text to Multimodal Generation

An astronaut riding a horse in the style of Andy Warhol.



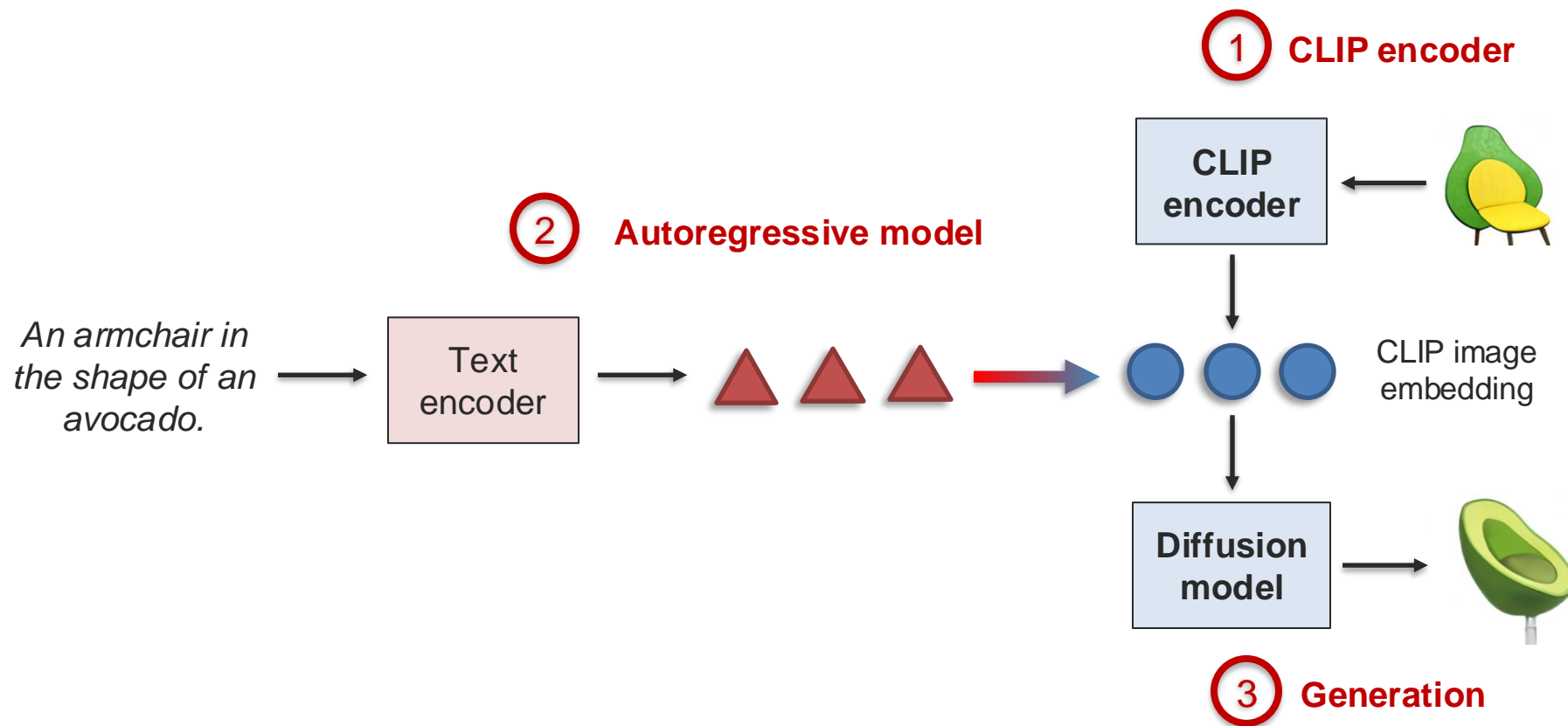
A bowl of soup that is a portal to another dimension as digital art



From Text to Multimodal Generation

Directly training diffusion models with conditional information

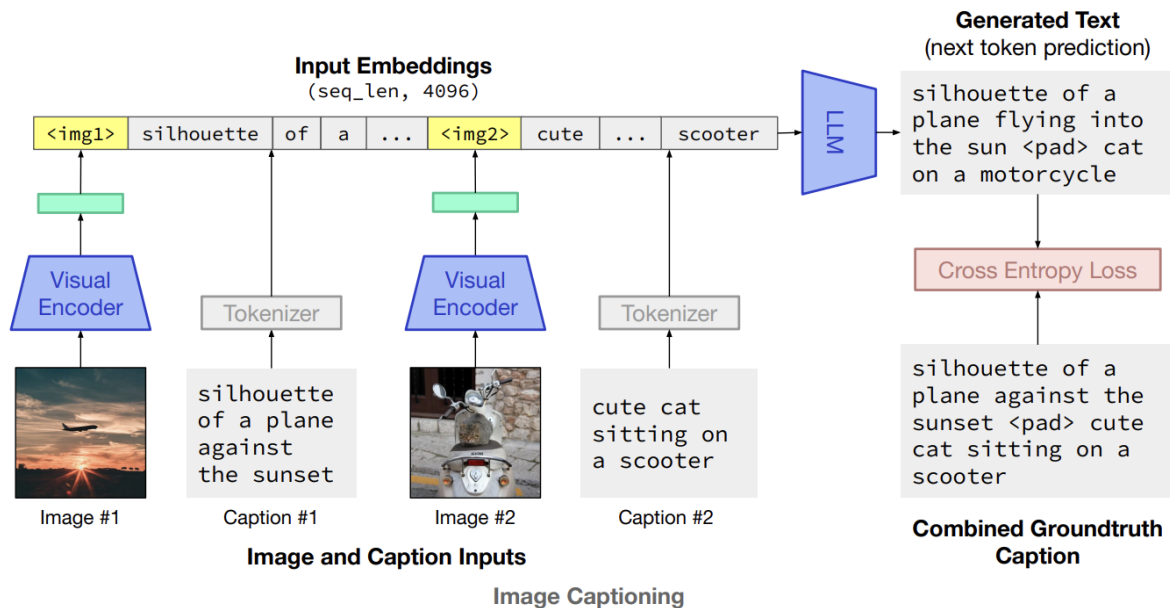
Conditional latent variables are pretrained CLIP embeddings, then diffusion model to generate image.



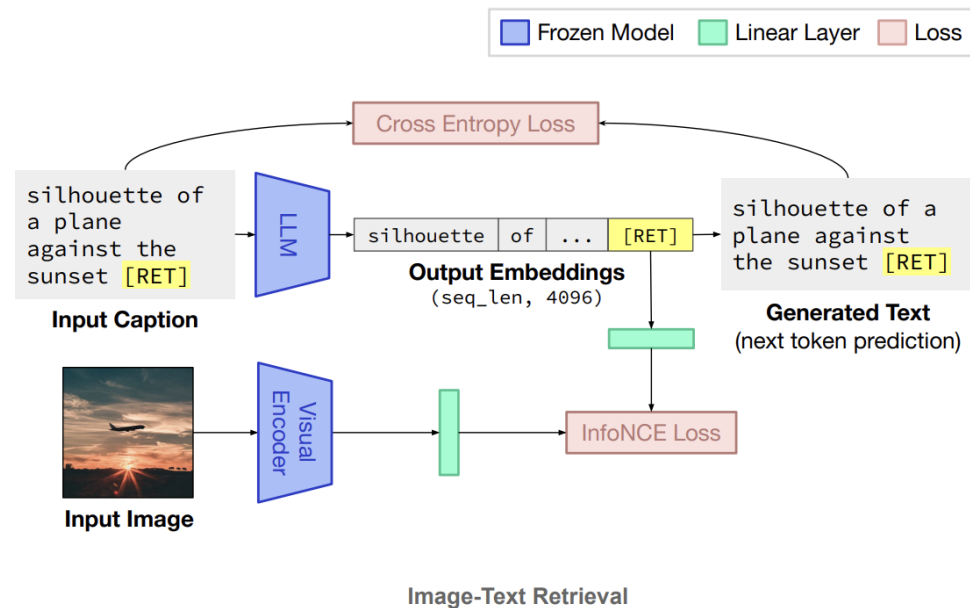
Grounding LMs for Image Retrieval

LIMBeR + CLIP. Trainable in 1 day on 1 GPU

Interleaved images and text



CLIP, with a frozen LLM



Grounding LMs for Image Retrieval

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?

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.



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



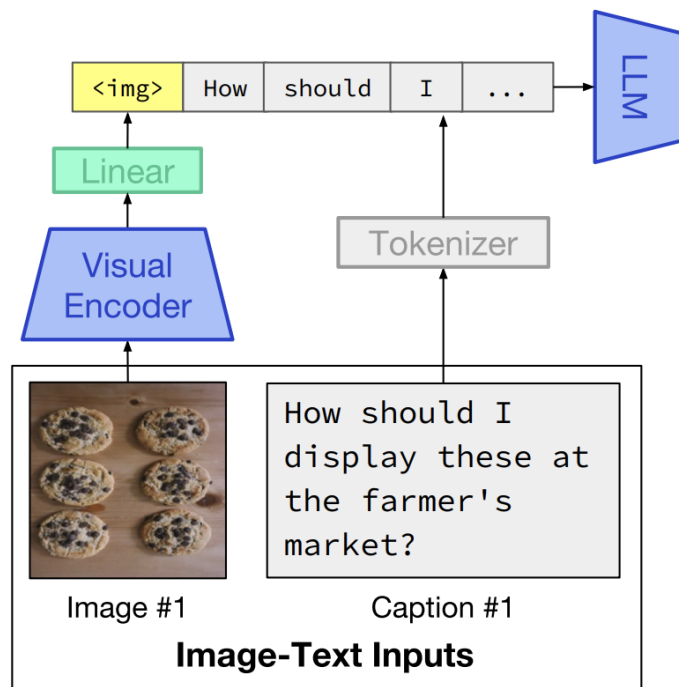
After this, what do I do?

The macaron is then baked in a hot oven.

Grounding LMs for Multimodal Generation

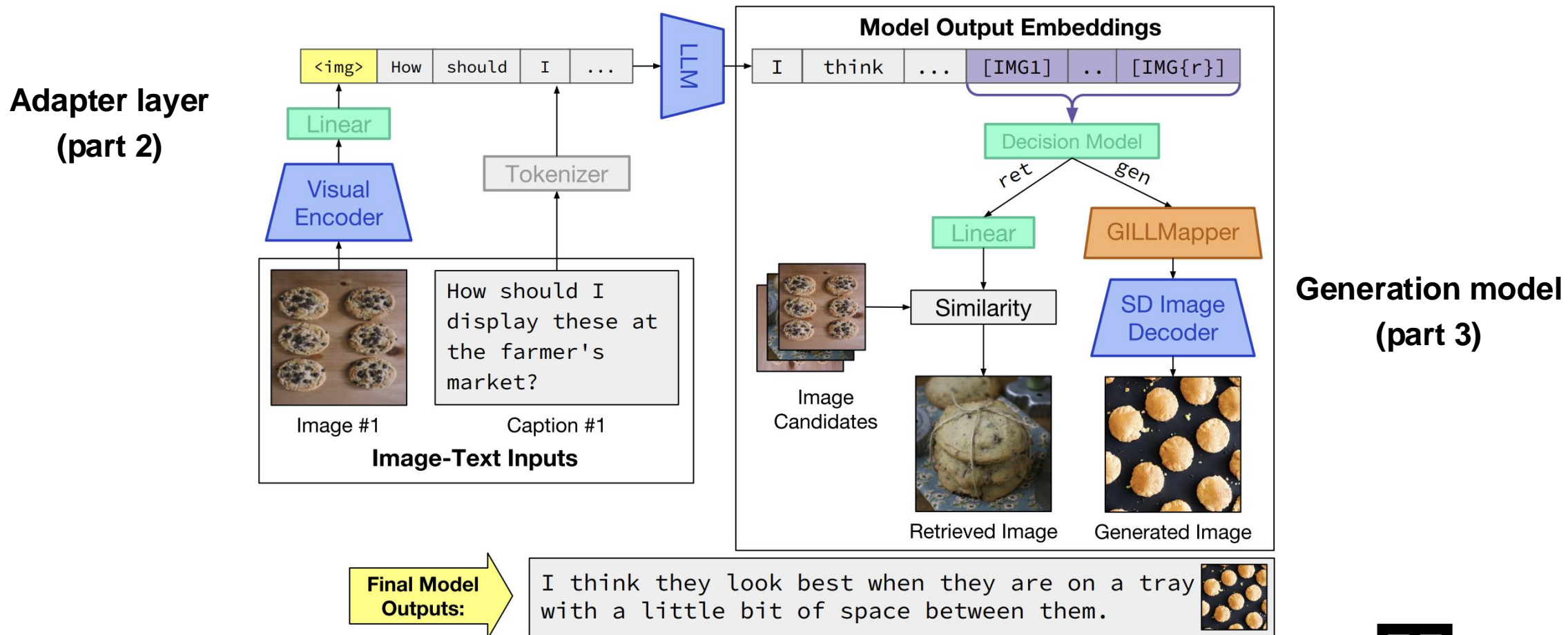
Large multimodal models with image generation

Adapter layer
(part 2)



Grounding LMs for Multimodal Generation


Large multimodal models with image generation



Grounding LMs for Multimodal Generation


Visual Storytelling

Our model can condition on interleaved image-and-text inputs to generate more relevant images compared to non-LLM based text-to-image generation models.



How can I make this more nutritious?


You can add vegetables to your ramen noodles, but you should be careful not to overdo it.




What are some vegetables I can add to it?

Broccoli, carrots, and green beans are all good choices.

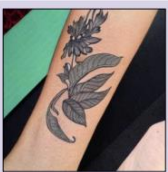
I'm looking for some ideas for a new tattoo. What do you think looks good on a petite person?




I want a tattoo of a sunflower, but I'm not sure what style would look best on me.




That's beautiful. Show me one that would work well as a complementary piece






How should I publicize these at the market?

I would suggest a simple sign with the name of the business and a picture of the cupcakes.




Hi, I'm looking for ideas for a customised cake design for my daughter's birthday.

I'm not sure what you mean by customised.




I want a cake with a unicorn theme, with pastel colors and a whimsical feel.

I think



Can you also show me some decorations that would work well for the party?

I can show you some ideas



Multimodal Dialogue

Our model can generate multimodal dialogue, weaving together text, retrieved images, and generated images.

□ User prompts □ Retrieved □ Generated

Native Multimodal Models

- *Native Multimodal Models*: LLMs Trained from scratch with multimodal input (instead of finetuning a trained unimodal LLM)
- Largest public model now: 109B - 2T parameters

The image is a screenshot of a webpage showcasing three Llama 4 models. The background is a light blue gradient. Each model is presented in a separate column, separated by thin vertical lines. The first column on the left features 'Llama 4 Behemoth' with a light blue background. The middle column features 'Llama 4 Maverick' with a light blue background. The third column on the right features 'Llama 4 Scout' with a light blue background. Each model's section includes its name, specifications (active parameters, experts, total parameters), a description, and a button ('Preview' for Behemoth, 'Available' for Maverick).

Model Name	Active Parameters	Experts	Total Parameters	Context Length	Availability
Llama 4 Behemoth	288B	16	2T	-	Preview
Llama 4 Maverick	17B	128	400B	1M	Available
Llama 4 Scout	17B	16	109B	10M	-

Llama 4 Behemoth
288B active parameter, 16 experts
2T total parameters
The most intelligent teacher model for distillation
[Preview](#)

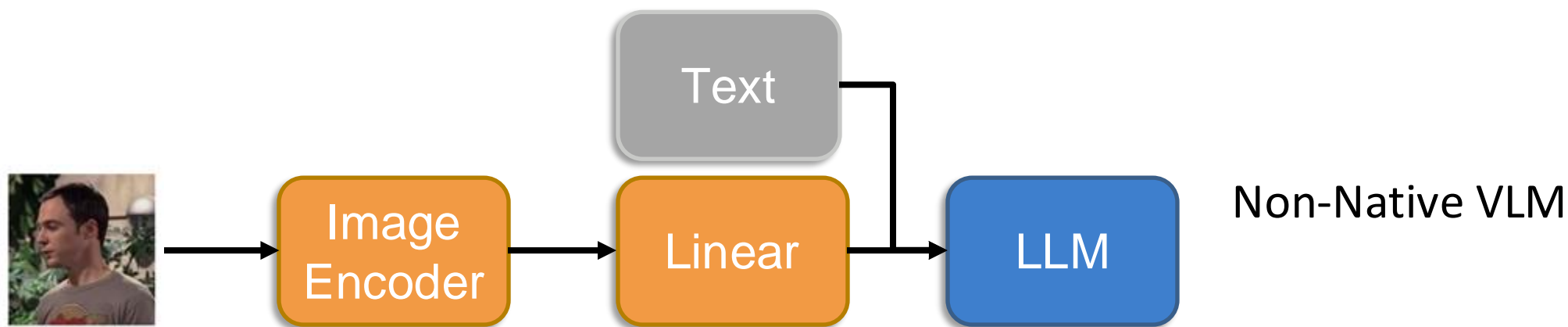
Llama 4 Maverick
17B active parameters, 128 experts
400B total parameters
Native multimodal with 1M context length
[Available](#)

Llama 4 Scout
17B active parameters, 16 experts
109B total parameters
Industry leading 10M context length
Optimized inference

Native Multimodal Models

- Background

- **Non-native VLMs:** Image encoder paired with frozen trained LLM. The image encoder can either be frozen or trained. Most VLMs now use this structure.

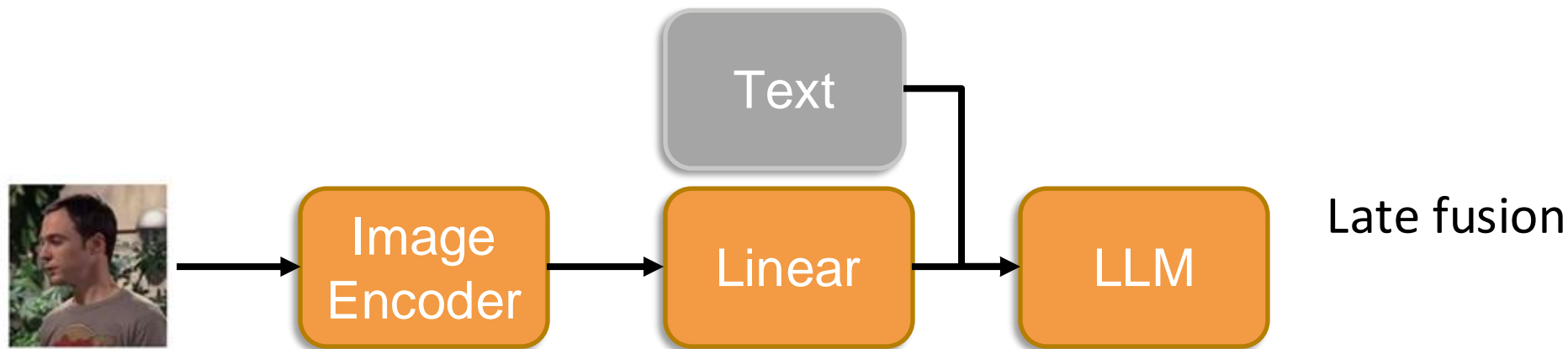


Most current VLMs use this architecture.

Native Multimodal Models

- *Background*

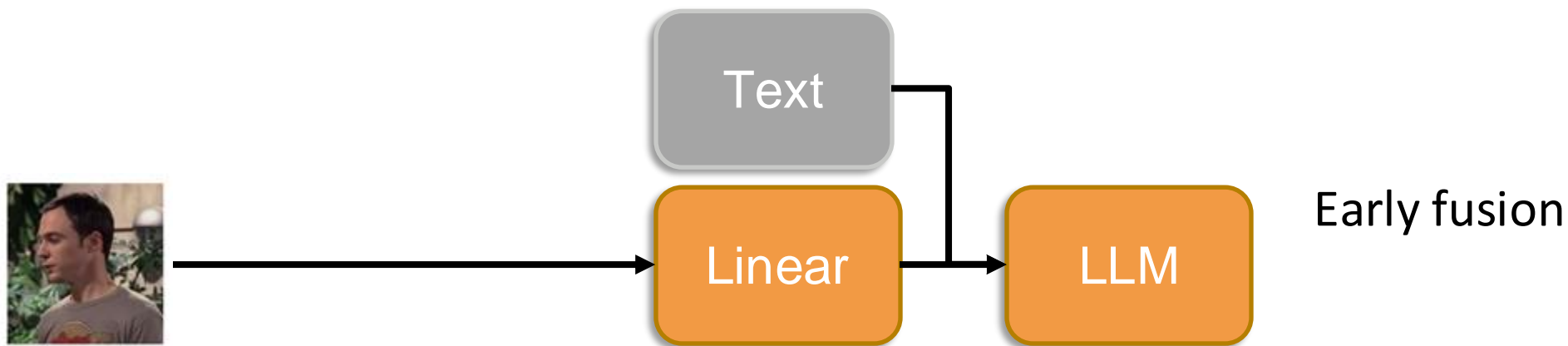
- Non-native VLMs: Image encoder paired with frozen trained LLM. The image encoder can either be frozen or trained. Most VLMs now use this structure.
- Native Multimodal Models: LLMs Trained from scratch with multimodal input
 - **Late fusion:** Image patches -> Image Encoder -> Linear -> LLM.
 - Early fusion: Image patches -> Linear -> LLM (No image encoder!)



Native Multimodal Models

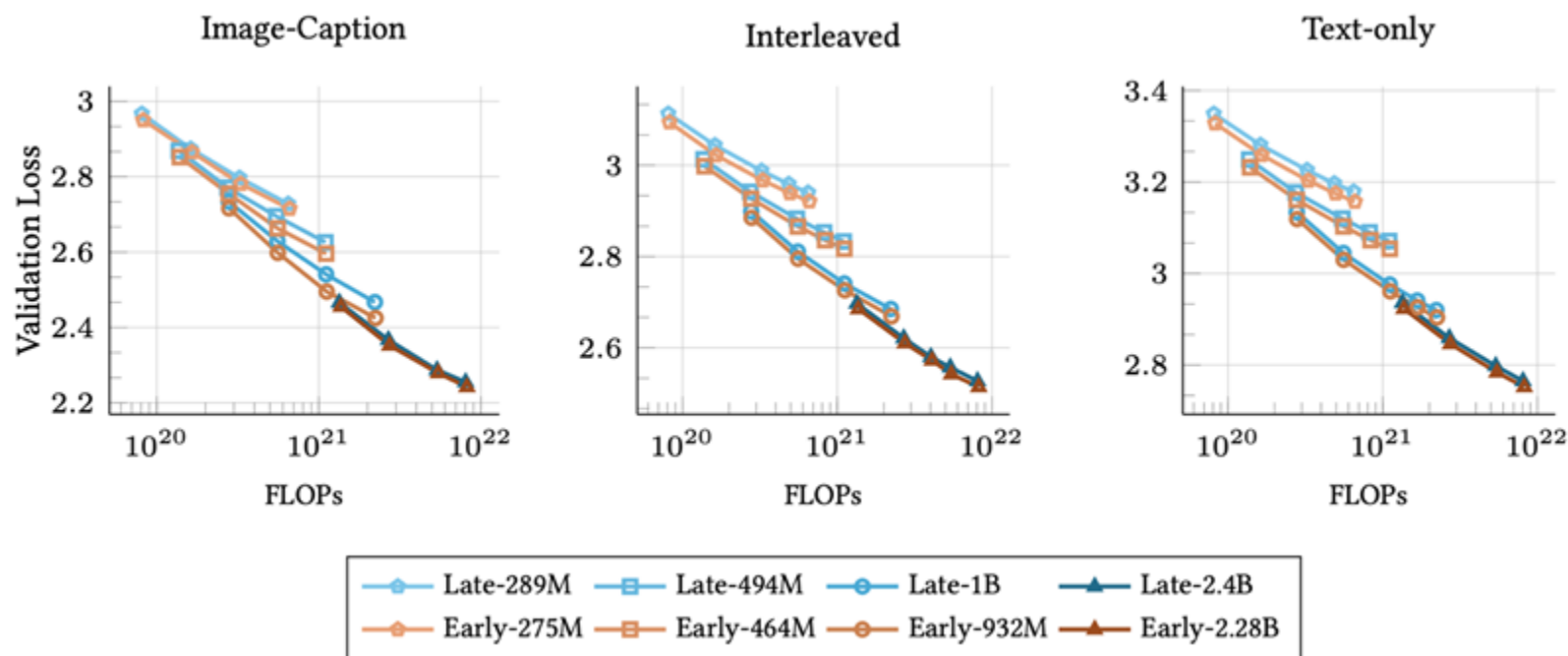
- Background

- Non-native VLMs: Image encoder paired with frozen trained LLM. The image encoder can either be frozen or trained. Most VLMs now use this structure.
- Native Multimodal Models: LLMs Trained from scratch with multimodal input
 - Late fusion: Image patches -> Image Encoder -> Linear -> LLM.
 - **Early fusion**: Image patches -> Linear -> LLM (No image encoder!)



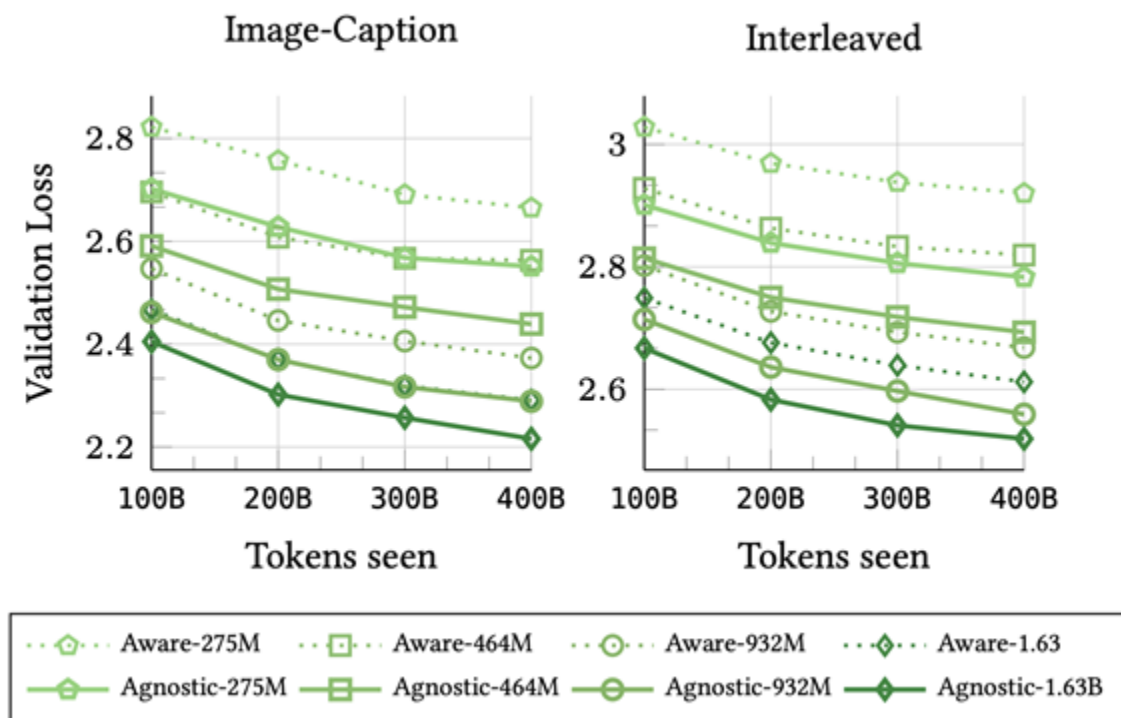
Scaling Laws for Native Multimodal Models

- Early fusion models hold small advantage on small scales.
- On larger scales, **both architectures perform similarly**. (We don't actually need image encoders!)
- **NMMs scale similarly to unimodal LLMs**, with slightly varying scaling exponents depending on the target data type and training mixture



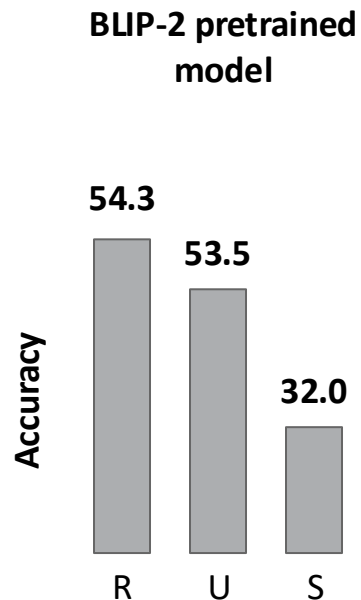
Scaling Laws for Native Multimodal Models

- Sparse structure like **MOE significantly benefits NMMs** at the same inference cost
- In an MOE structure, Modality-aware design (having separate image/text experts) performs **worse** than modality-agnostic design (unified experts for both image/text tokens)



One model for everything?

Video sarcasm detection



Y : Sarcasm

X_ℓ : Spoken
language

*It's just a privilege to
watch your mind at work.*

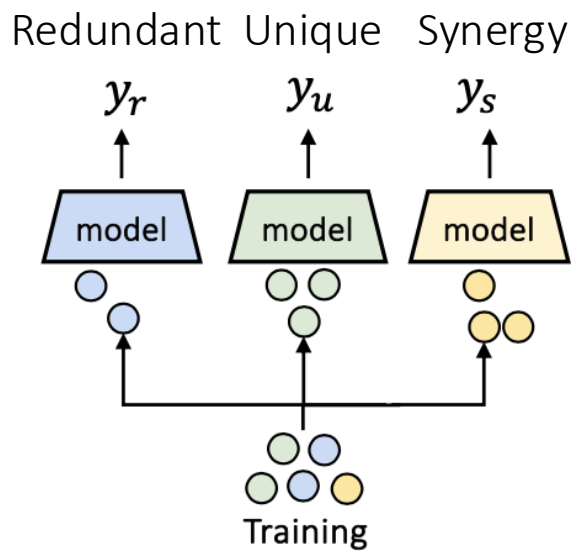
X_{av} : Audio
+ visual



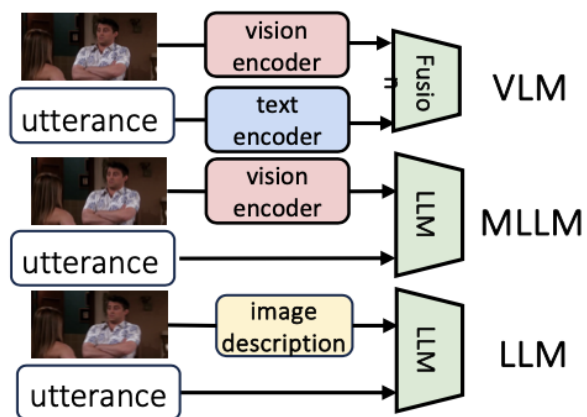
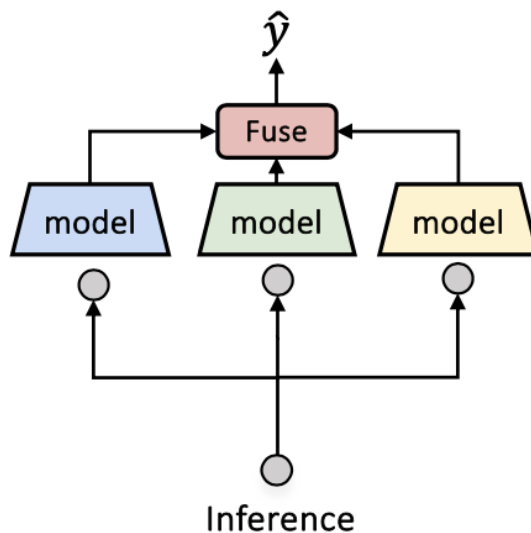
Neutral tone + straight face

Mixture of Multimodal Interaction Experts

One model for everything -> specialized models for each interaction



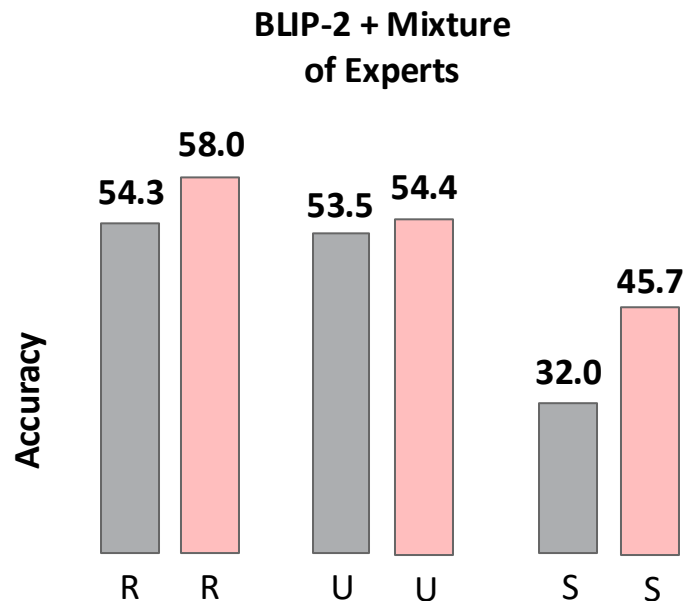
Datapoint-level
quantification
of interactions



Mixture of Multimodal Interaction Experts

One model for everything -> specialized models for each interaction

Video sarcasm detection



*The car is as fast
as a cheetah.*



[Yosef et al., EMNLP 23]

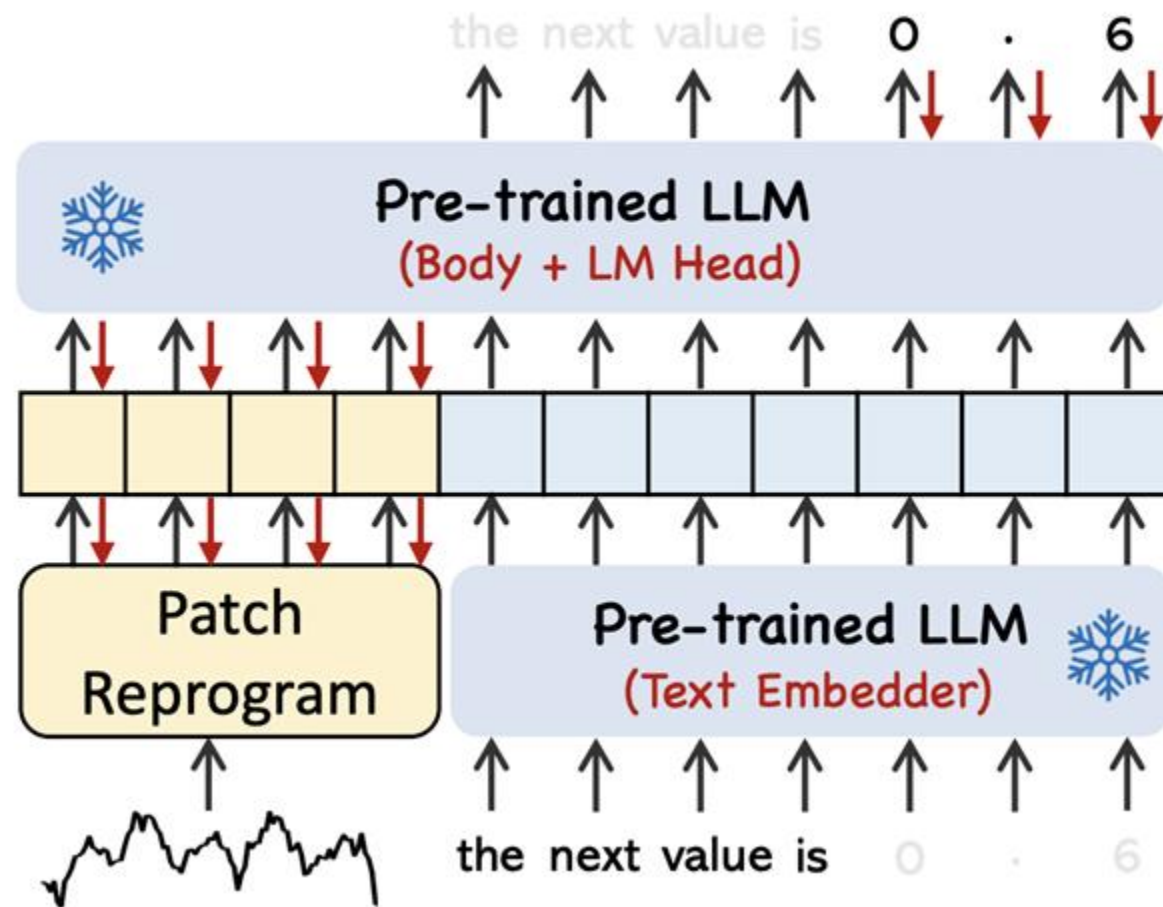


Can you please pass the cow?

[Hessel et al., ACL 23]

Time-series Multimodal Models

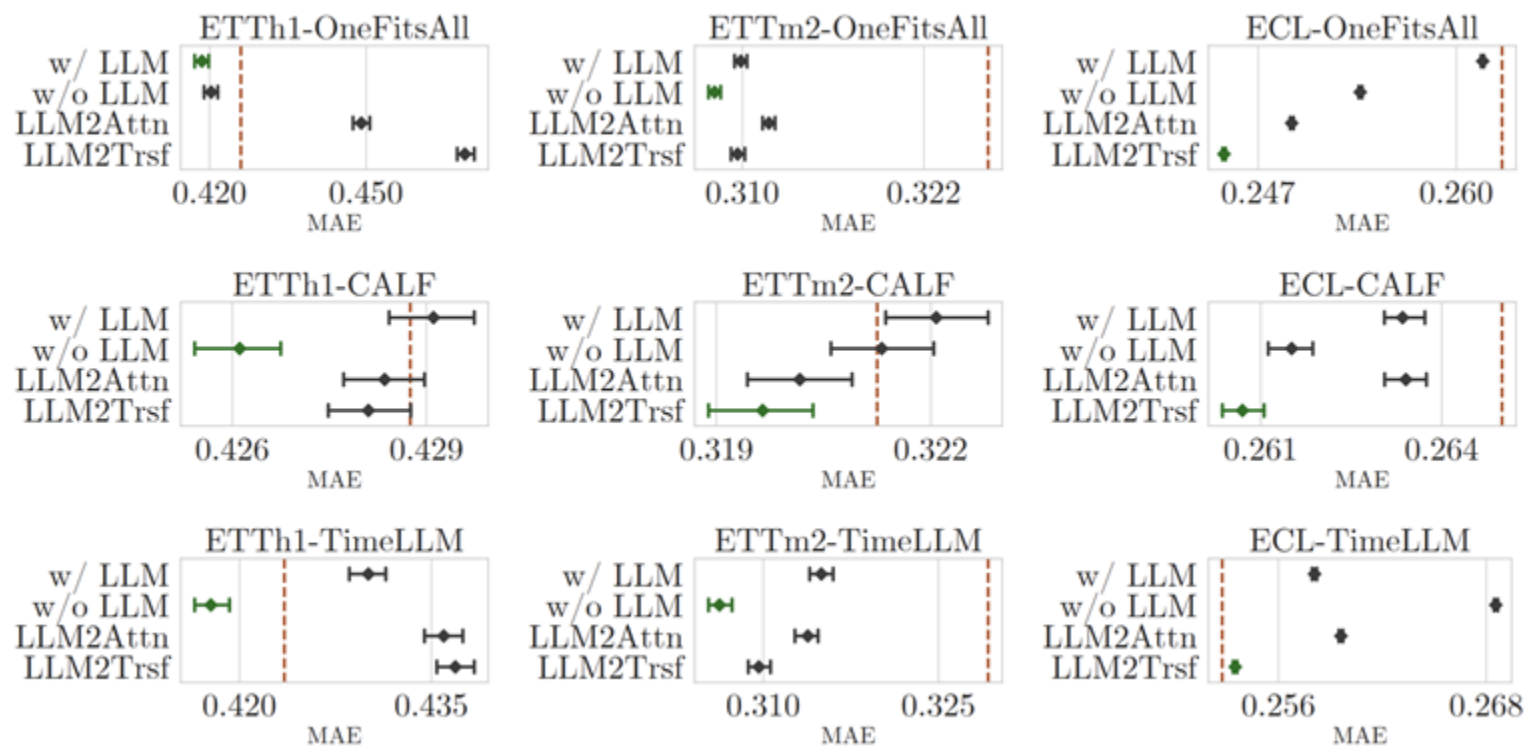
- Typically trained & aligned the same way as vision language models (alignment + instruction tuning)
- Works for both analysis and prediction
- Example: Time-LLM, OneFitsAll



Time-series Multimodal Models

- But some current time series LLMs have questionable performance. Replacing LLM with a simple attention layer doesn't significantly degrade performance (sometimes even better).

* Lower is better



Today's lecture

- 1 Multimodal foundation models and pre-training
- 2 Adapting LLMs into multimodal LLMs
- 3 From text to multimodal generation
- 4 Latest directions: natively multimodal, multimodal MoE, real-world modalities

Assignments for This Coming Week

For project:

- Make sure to meet with myself and TAs this week
- Medium progress towards implementing new ideas. Either promising results or poor results, but a good idea of what is wrong and how to fix.

Reading assignment due tomorrow Wednesday (4/16).

This Thursday (4/17): fifth reading discussion on **large language models**.

1. Alignment faking in LLMs
2. Reasoning in LLMs