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**.

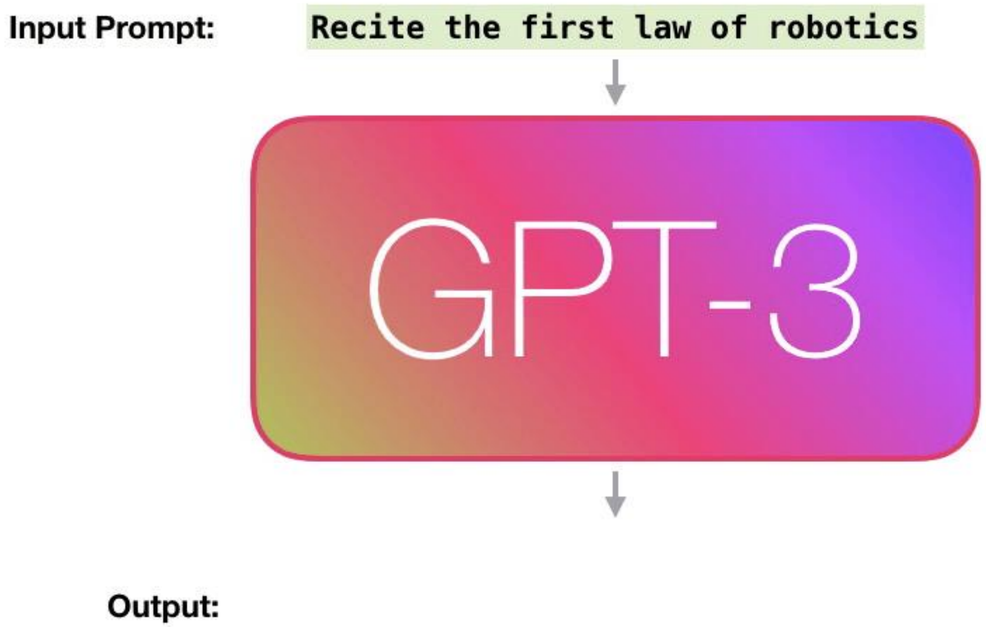
1. Alignment faking in LLMs
2. Mathematical reasoning in LLMs

multisensory intelligence

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



Question answering

Open-ended dialog

Translate to different languages

Retrieve news

Solving math problems

Writing code

and more…

[Brown et al., Language Models are Few-shot Learners. NeurIPS 2020]

multisensory intelligence

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.

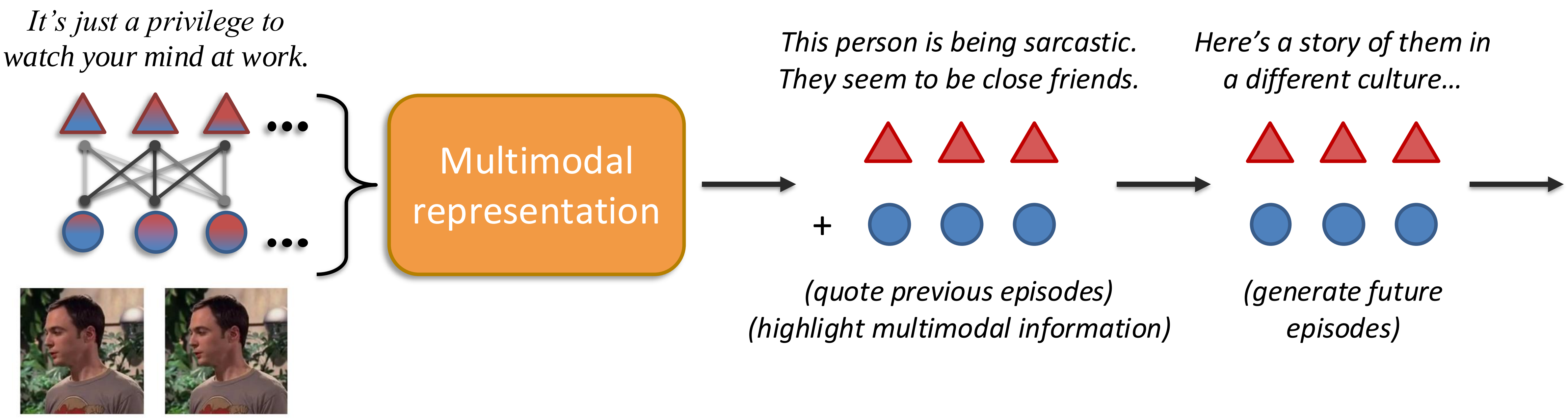
Generation: Animate a story inspired by this short video clip.

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

[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. ACM Computing Surveys 2024]

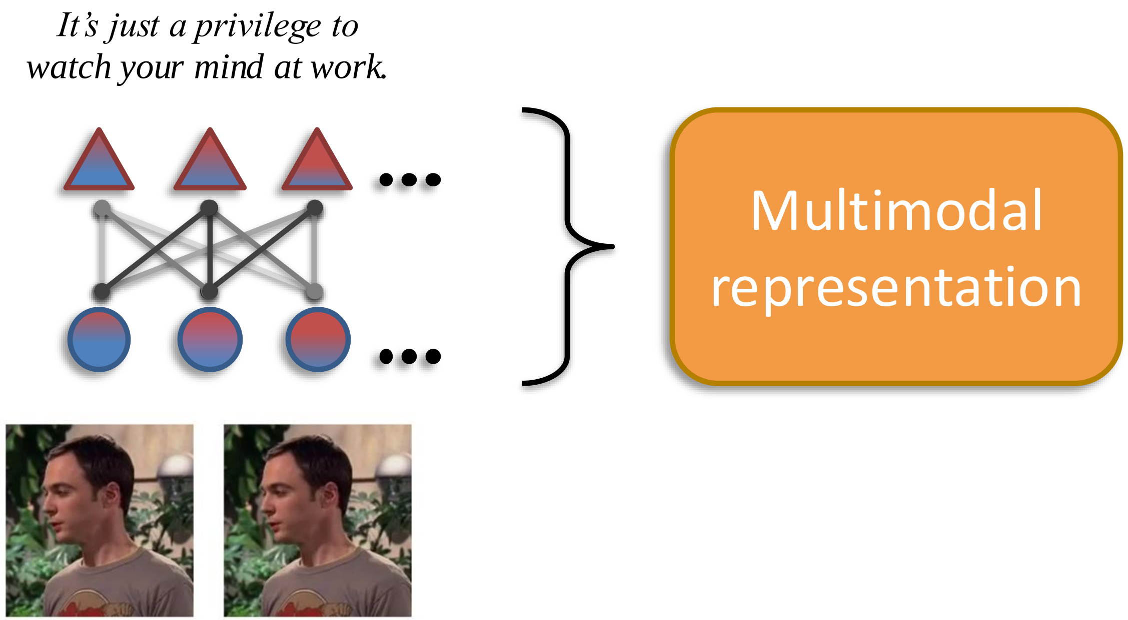
multisensory intelligence

From Large Language Models to Multimodal Models



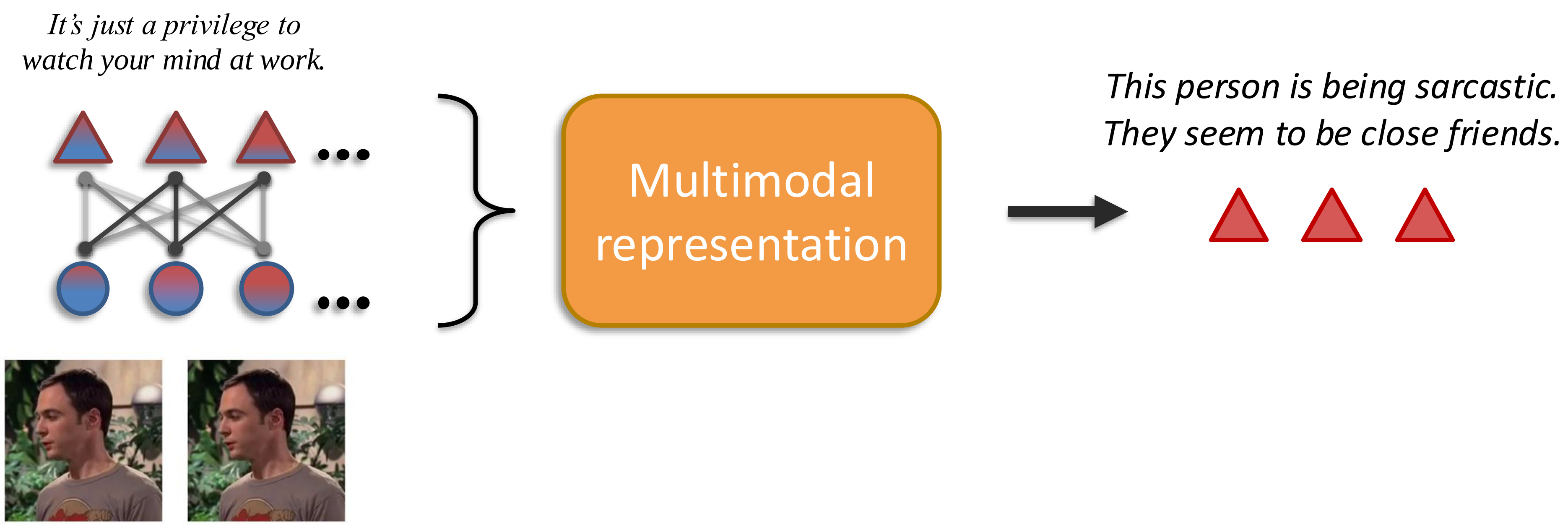
Lecture outline

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



Lecture outline

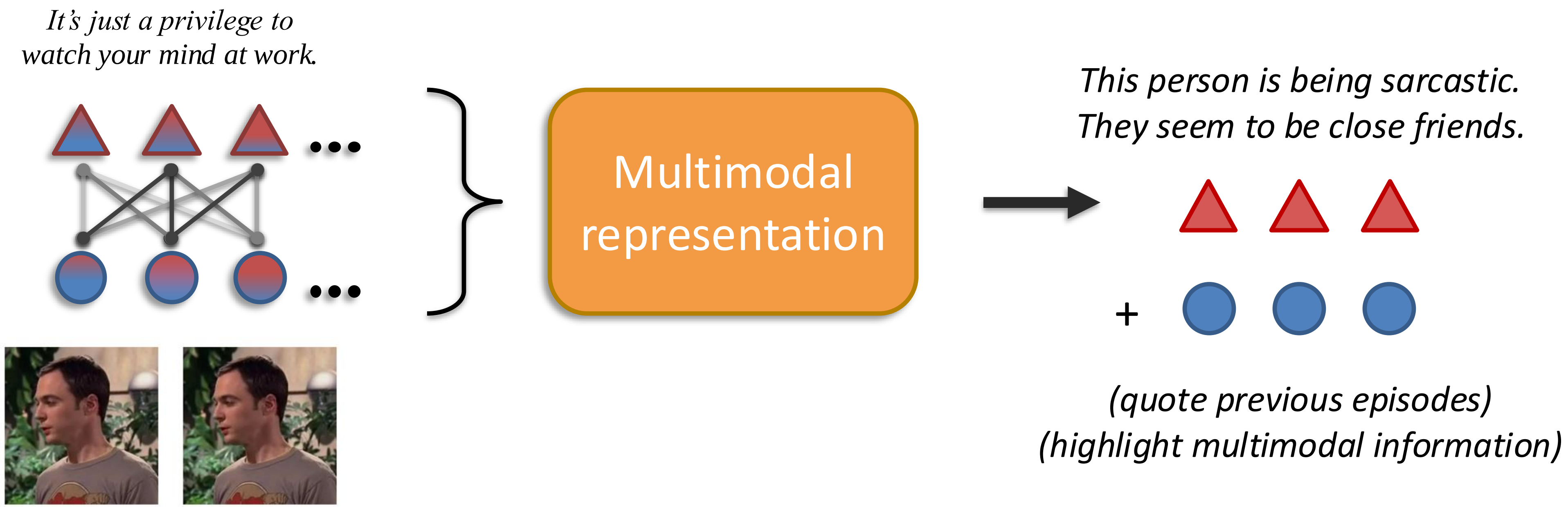
Part 2: Adapting large language models for multimodal text generation



multisensory intelligence

Lecture outline

Part 3: Enabling text and image generation



multisensory
intelligence

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

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

privilege
mind

vocal emphasis
eye-roll

vocal emphasis
eye-roll

Vision-to-language attention

	vocal emphasis			eye-roll
privilege	0.7			0.3
mind	1.0			

privilege
mind

New **language** representation contextualized with vision

Sarcasm

3 x 4

(weighted) outer product

(row) normalize to 0-1

3 x d   d x 4   4 x d

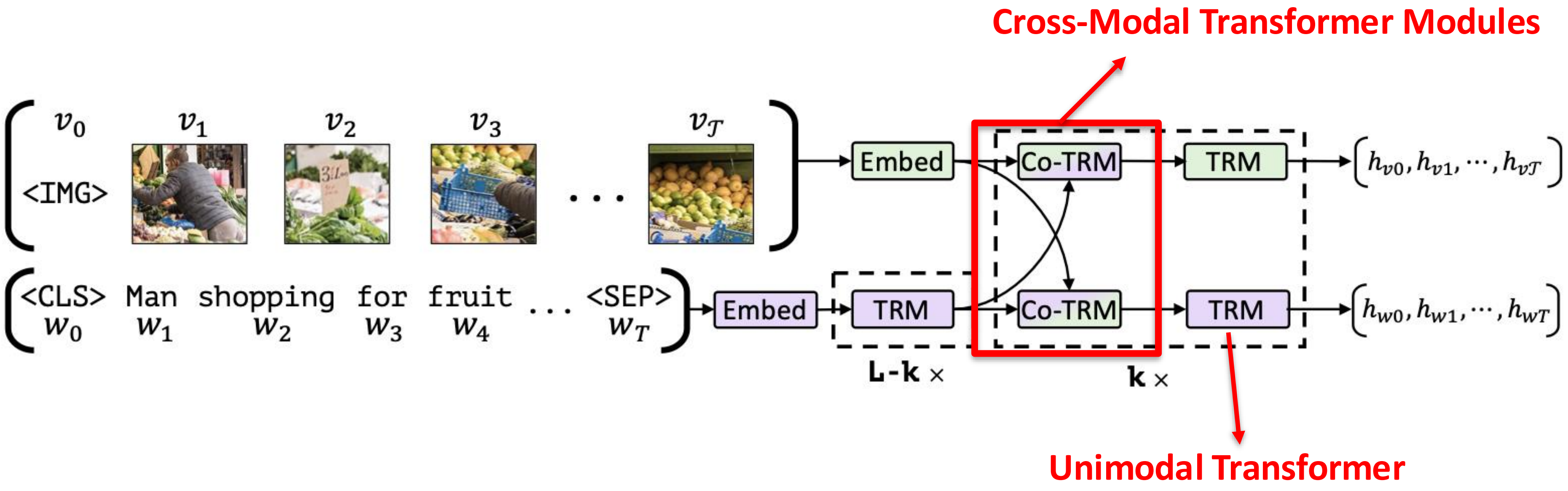
$$h = \mathrm{softmax}\left(\frac{X_1 W_q W_k^T X_2^T}{\sqrt{d}}\right) X_2 W_v$$

normalize wrt dimension *d*

[Liang et al., Multimodal Language Analysis with Recurrent Multistage Fusion. EMNLP 2018]
[Tsai et al., Multimodal Transformer for Unaligned Multimodal Language Sequences. ACL 2019]

multisensory intelligence

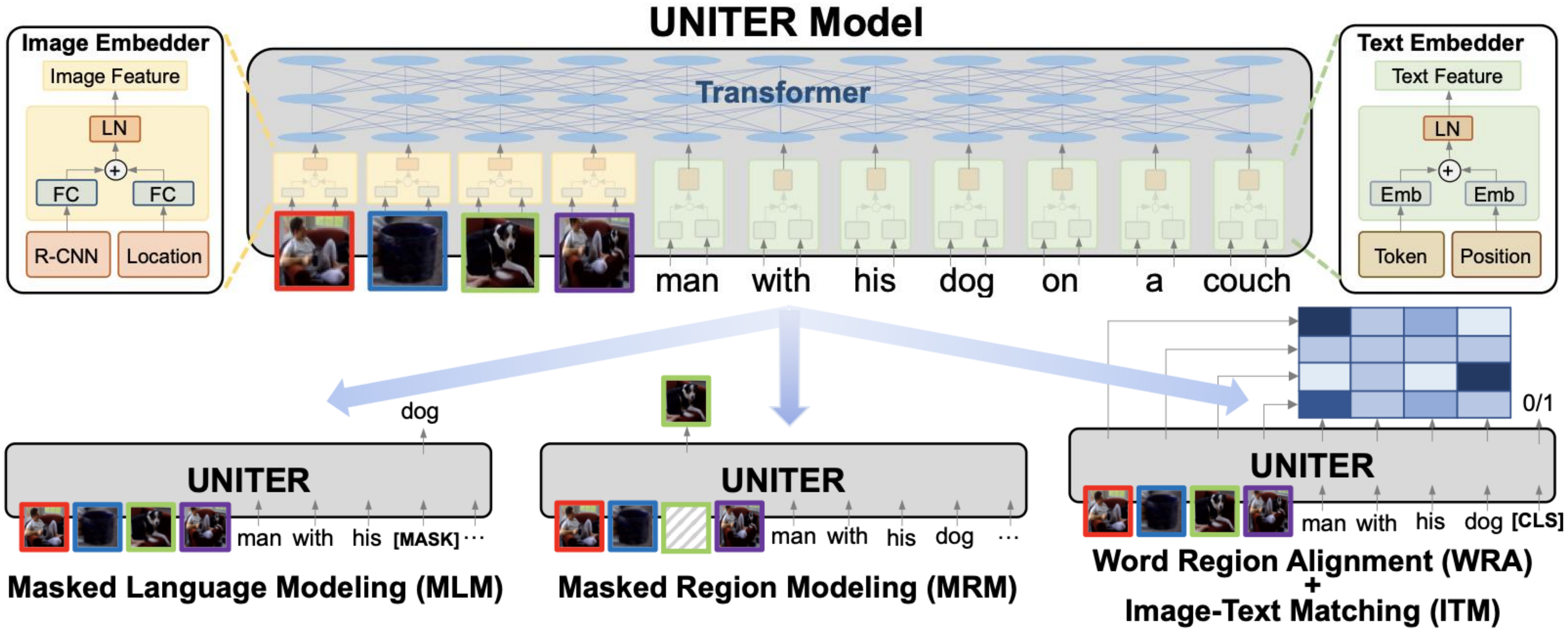
Multimodal Cross-attention Transformers



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

[Lu et al., Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. NeurIPS 2019]

Multimodal Cross-attention Transformers



Masking words

Masking images

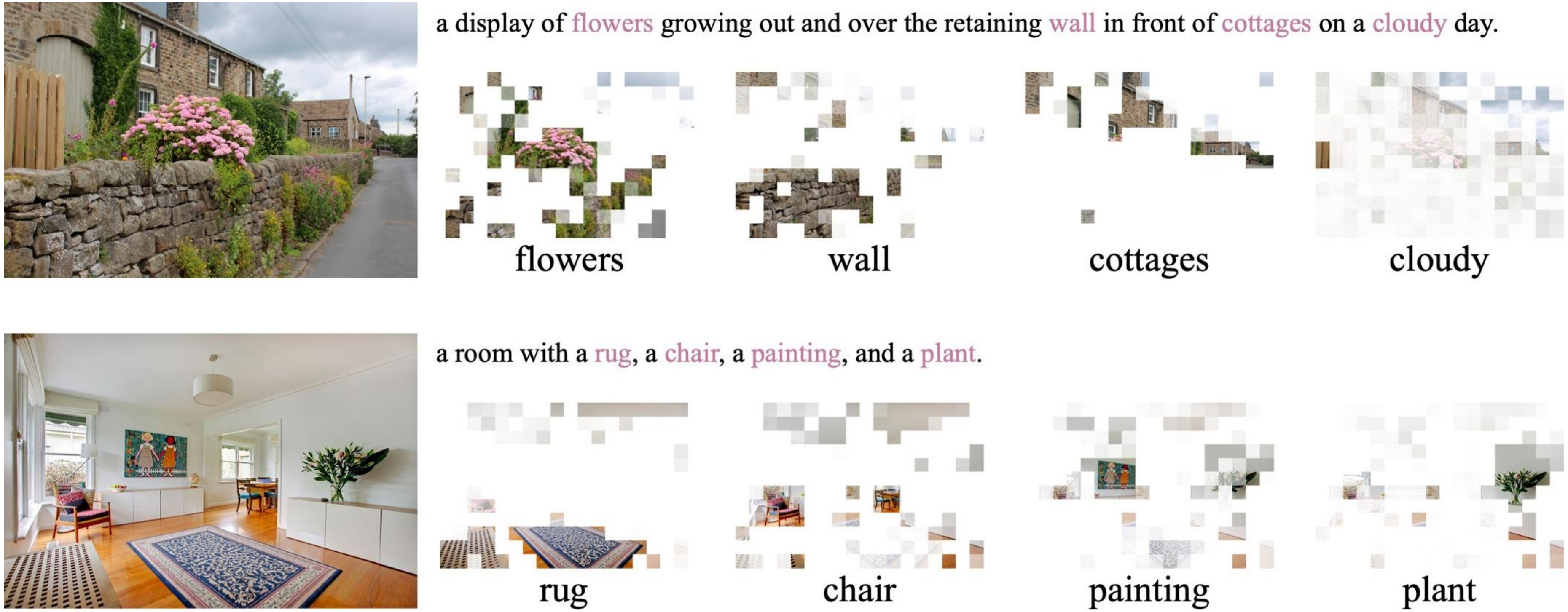
Does the sentence match the image?

[Chen et al., Uniter: Universal Image-Text Representation Learning. ECCV 2020]

[Kim et al., ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. ICML 2021]

Visual-and-Language Transformer (ViLT)

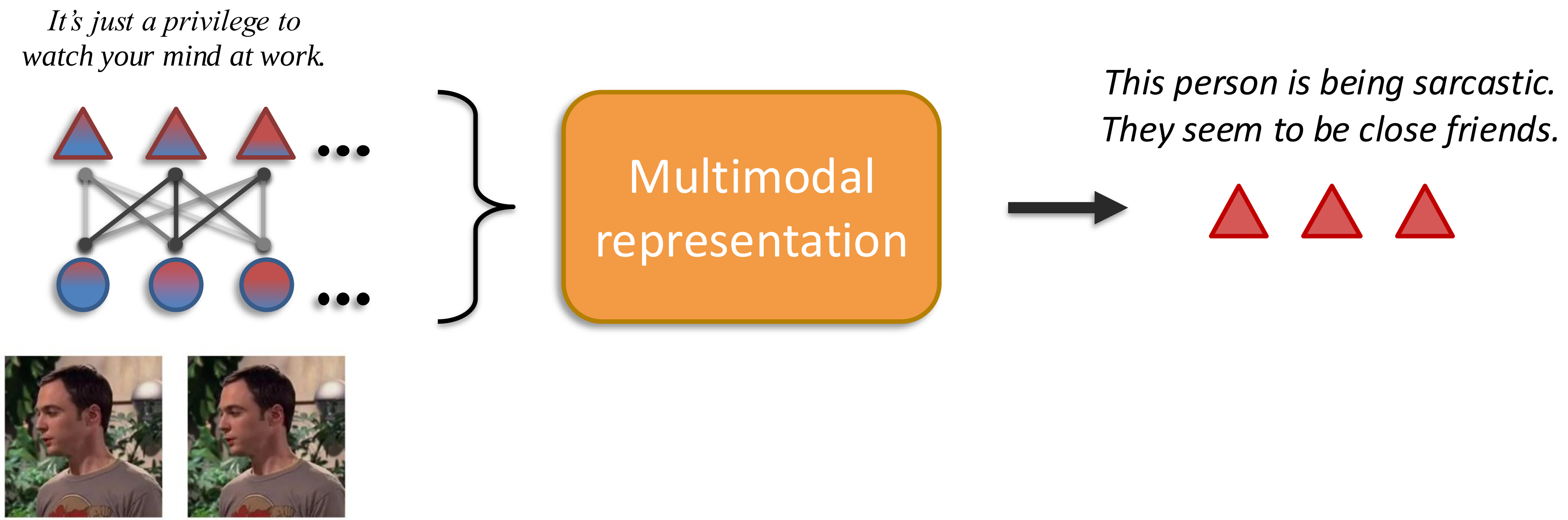
Example of alignment between modalities:



[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]
[Kim et al., ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. ICML 2021]

multisensory intelligence

Adapting Large Language Models to Multimodal

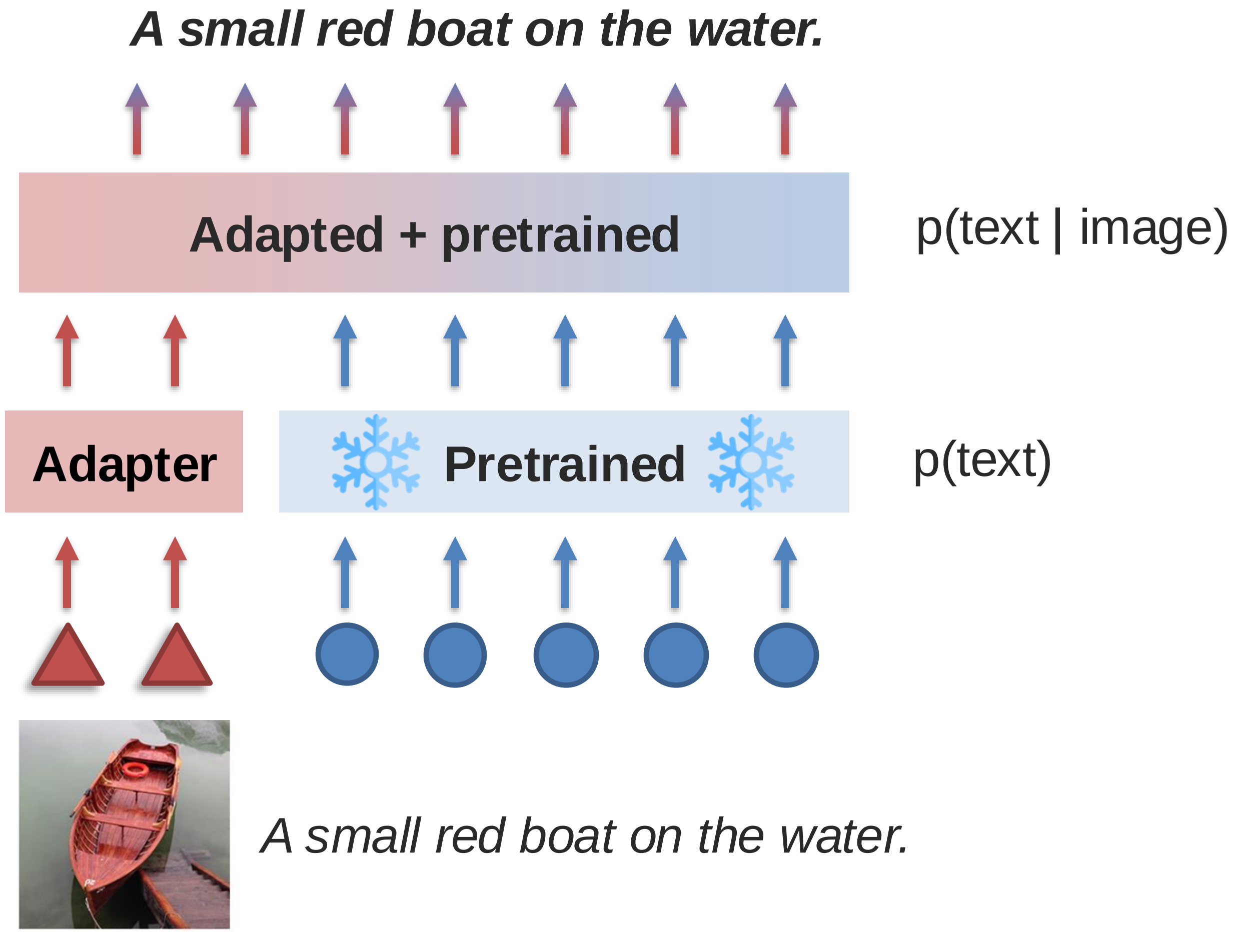


multisensory intelligence

Adapting Large Language Models

Conditioning via prefix tuning

Modeling p(text | image):

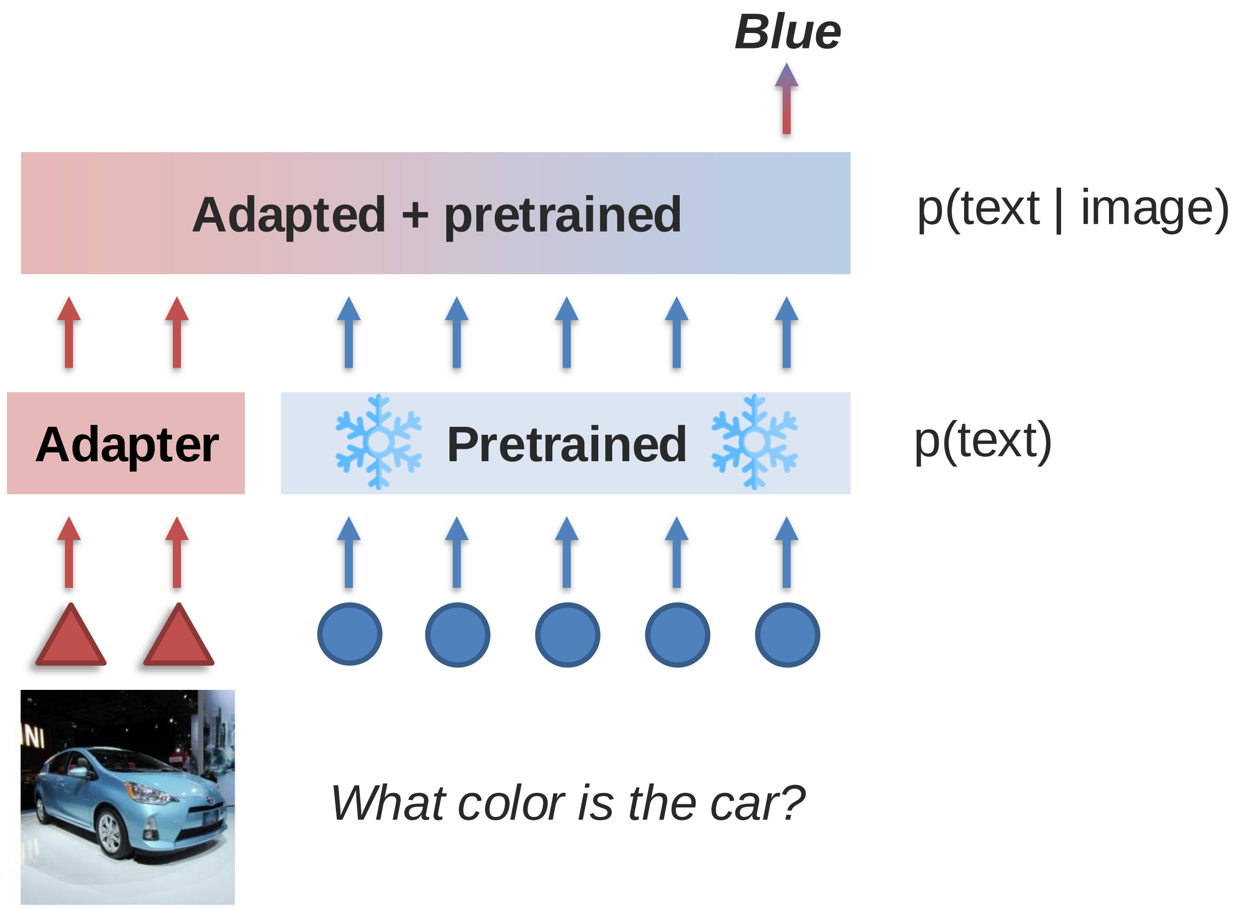


[Tsimpoukelli et al., Multimodal Few-Shot Learning with Frozen Language Models. NeurIPS 2021]

Adapting Large Language Models

Conditioning via prefix tuning

Modeling p(text | image):

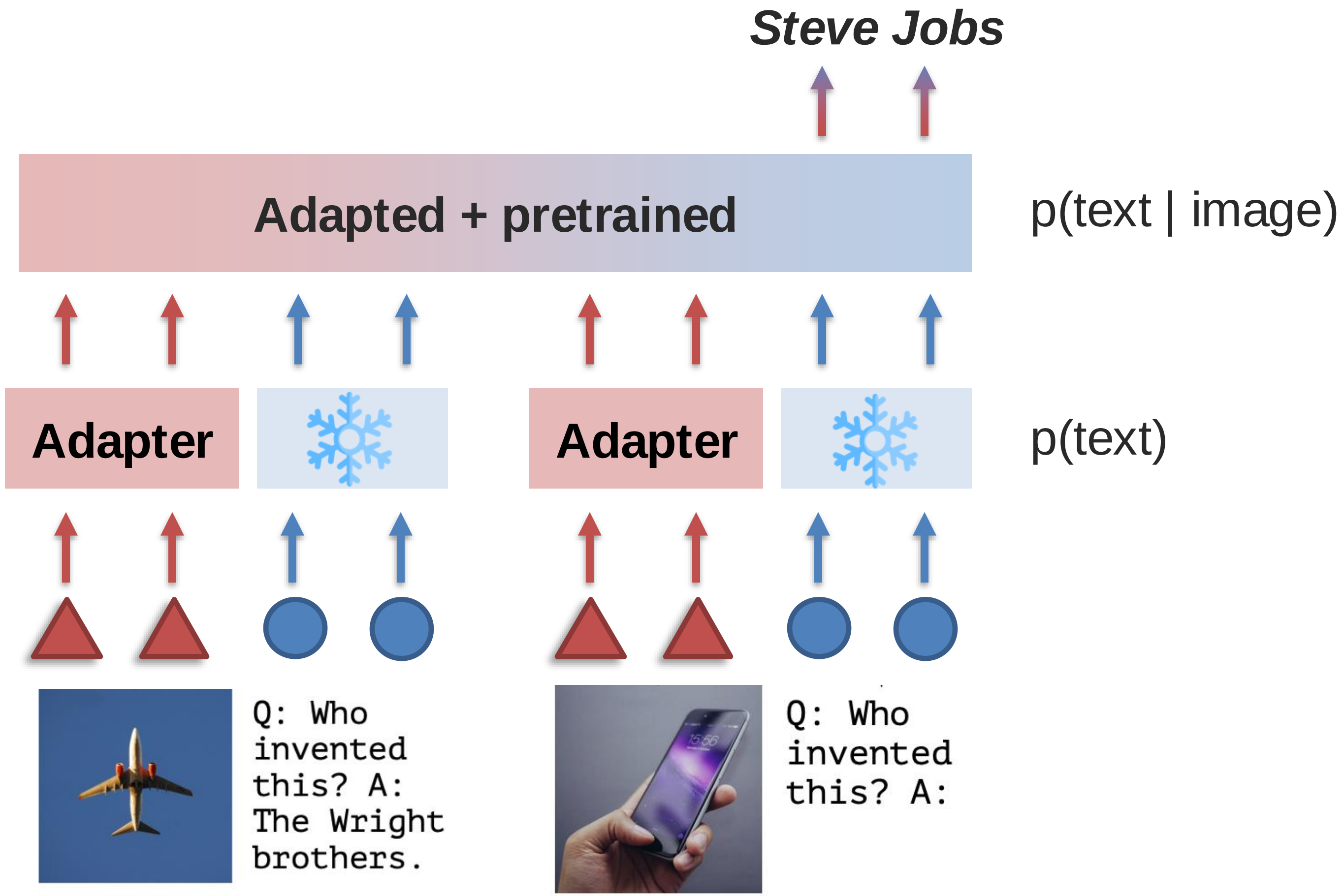


[Tsimpoukelli et al., Multimodal Few-Shot Learning with Frozen Language Models. NeurIPS 2021]

Adapting Large Language Models

Conditioning via prefix tuning

Modeling p(text | image):

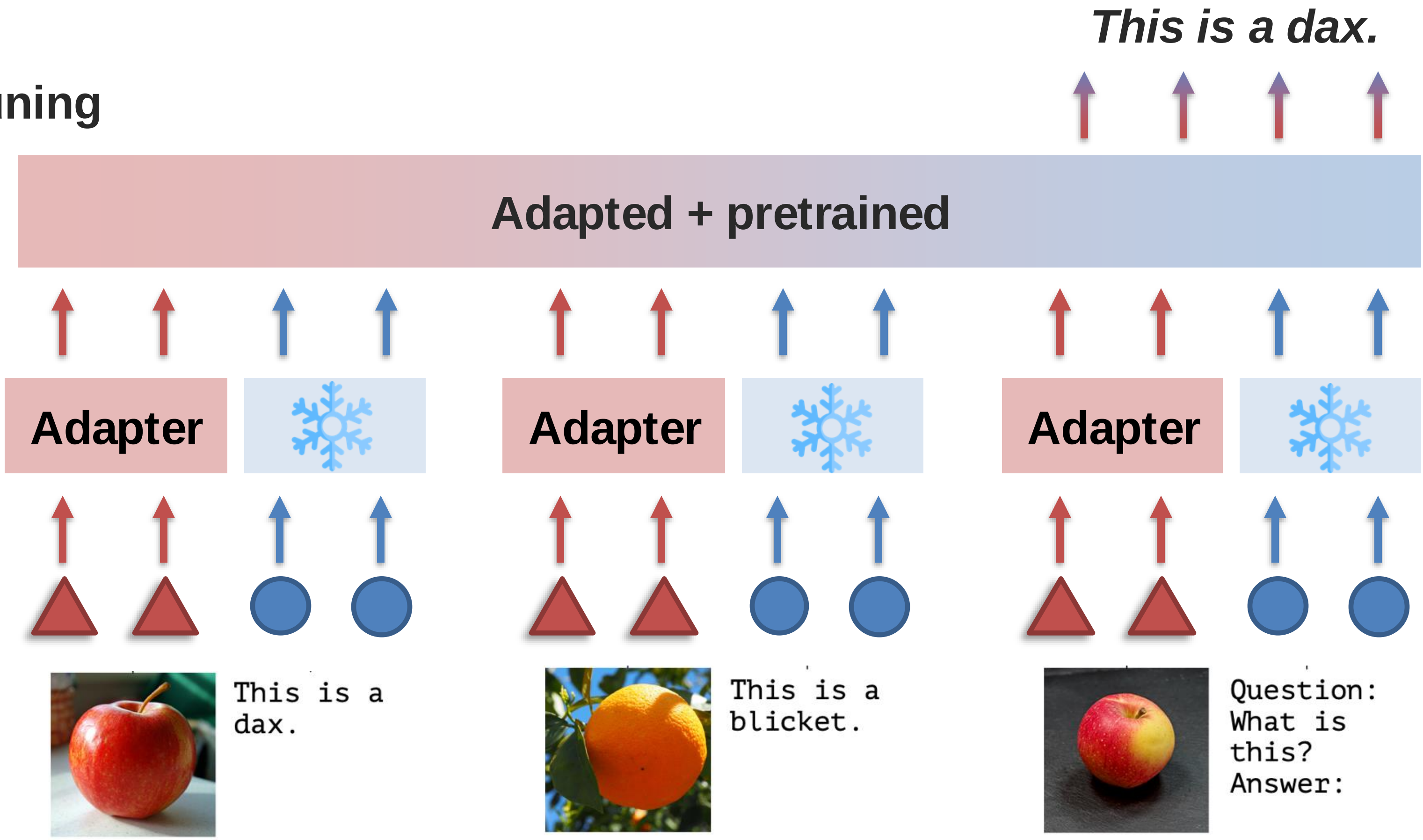


[Tsimpoukelli et al., Multimodal Few-Shot Learning with Frozen Language Models. NeurIPS 2021]

Adapting Large Language Models

Conditioning via prefix tuning

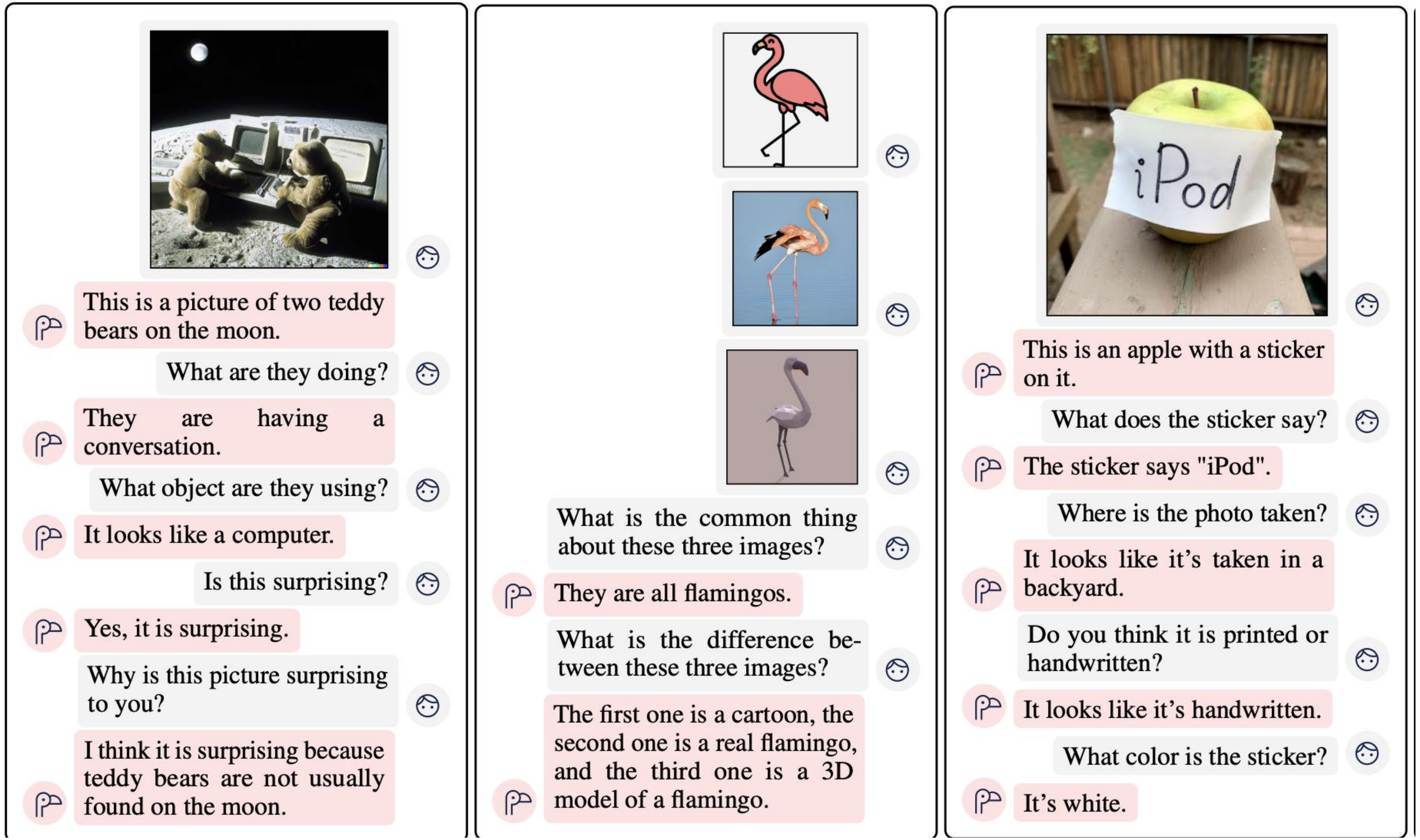
Modeling p(text | image):



[Tsimpoukelli et al., Multimodal Few-Shot Learning with Frozen Language Models. NeurIPS 2021]

Adapting Large Language Models

Flamingo



[Alayrac et al. Flamingo: a Visual Language Model for Few-Shot Learning. NeurIPS 2022]

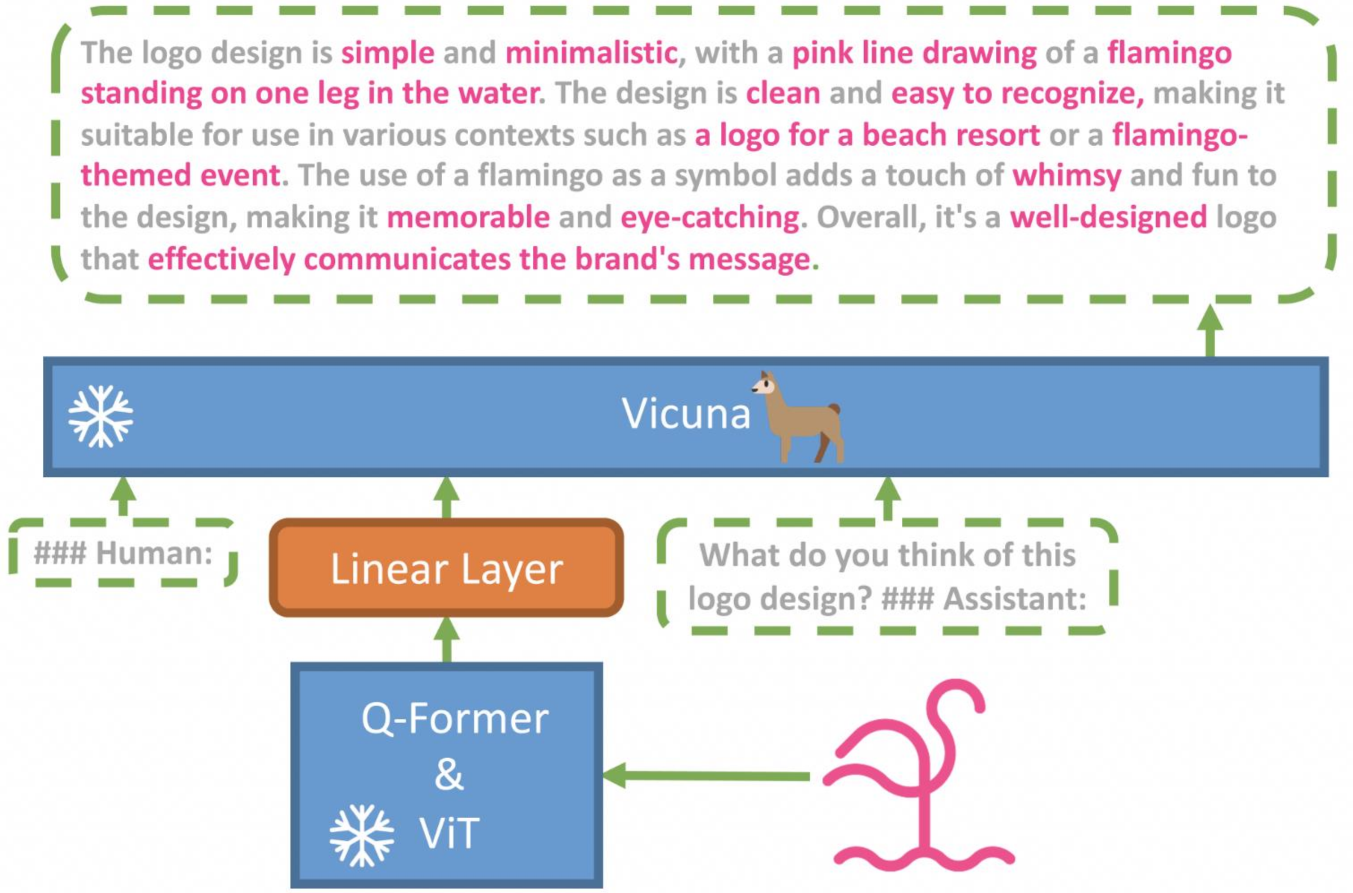
multisensory intelligence

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.

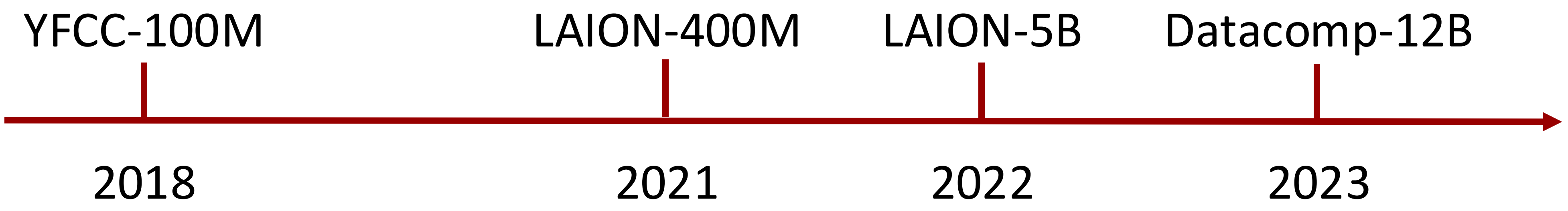
Scaling Large Multimodal Models

LLaMA-Adapter

[Zhang et al., LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention. ICLR 2024]

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)



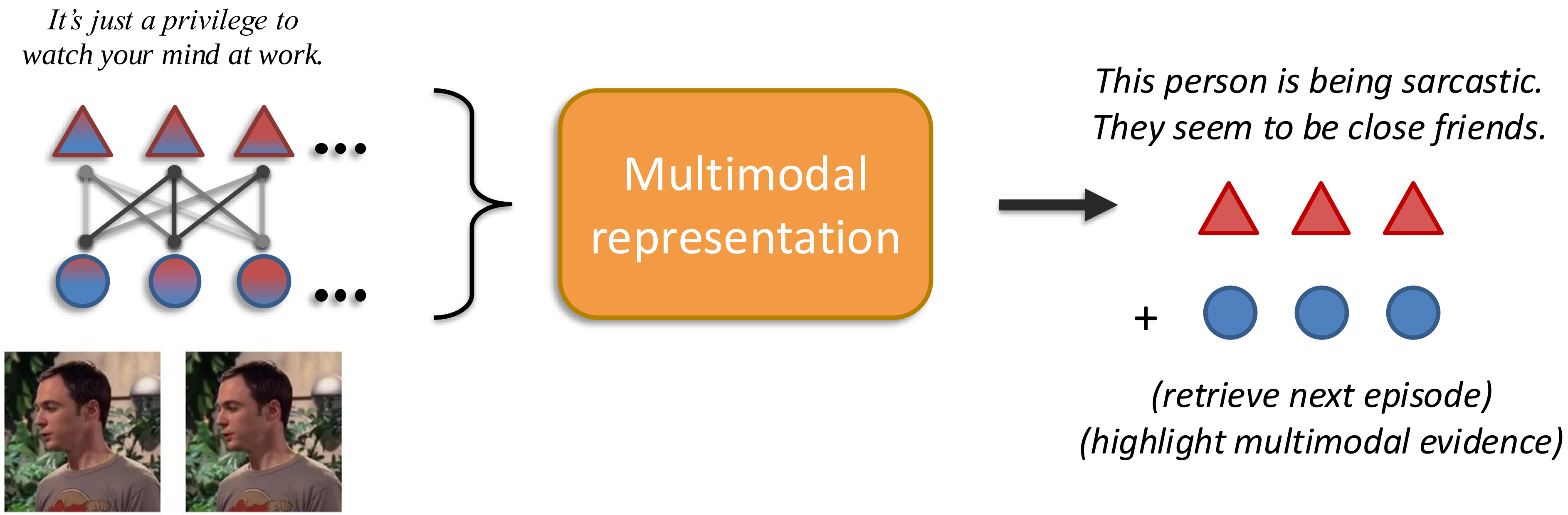
Data type	dataset	#samples	sampling prob.
	DFN [Fang et al., 2023]	2B	27%
Image-Caption	COYO [Byeon et al., 2022]	600M	11.25%
	HQITP	400M	6.75%
Interleaved	Obelics [Laurençon et al., 2024a]	141M Docs	45%
Text	DCLM [Li et al., 2024b]	6.6T Toks	10%

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	✗	$\sim$ 235K	5	✗
InstructBLIP	28	✗	$\sim$ 1.6M	9.7	✗
M^3IT (Ours)	40	✓	2.4M	10	✓

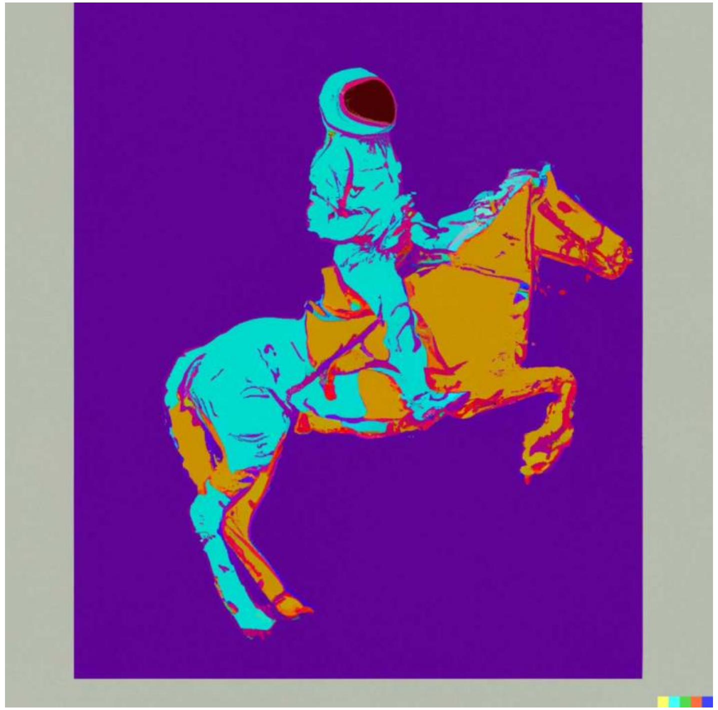
From Text to Multimodal Generation



multisensory intelligence

From Text to Multimodal Generation

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



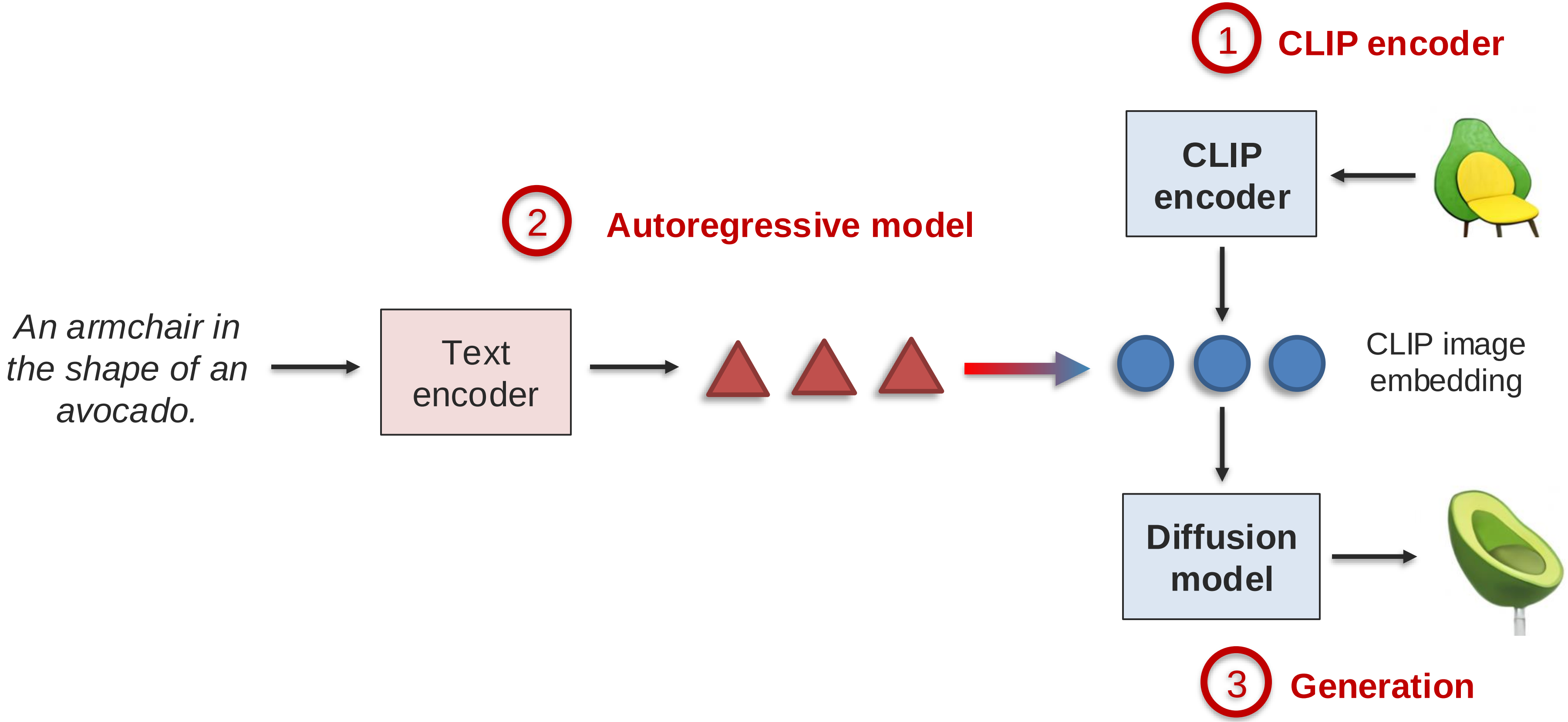
A bowl of soup that is a portal to another dimension as digital art

[Ramesh et al., Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv 2022]

multisensory intelligence

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.



[Ramesh et al., Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv 2022]

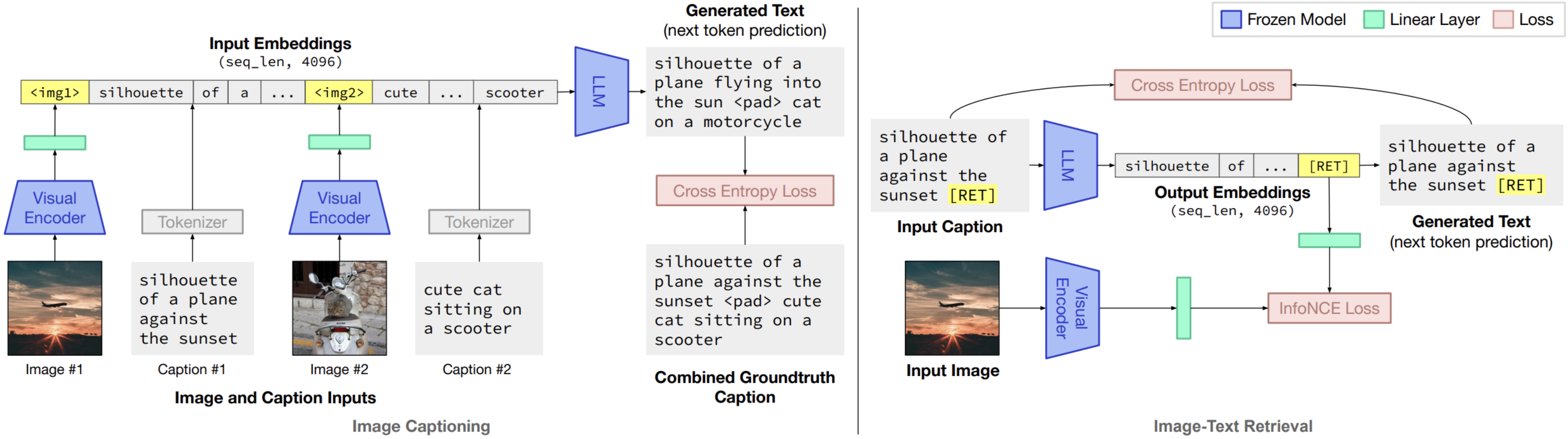
multisensory intelligence

Grounding LMs for Image Retrieval

LIMBeR + CLIP. Trainable in 1 day on 1 GPU

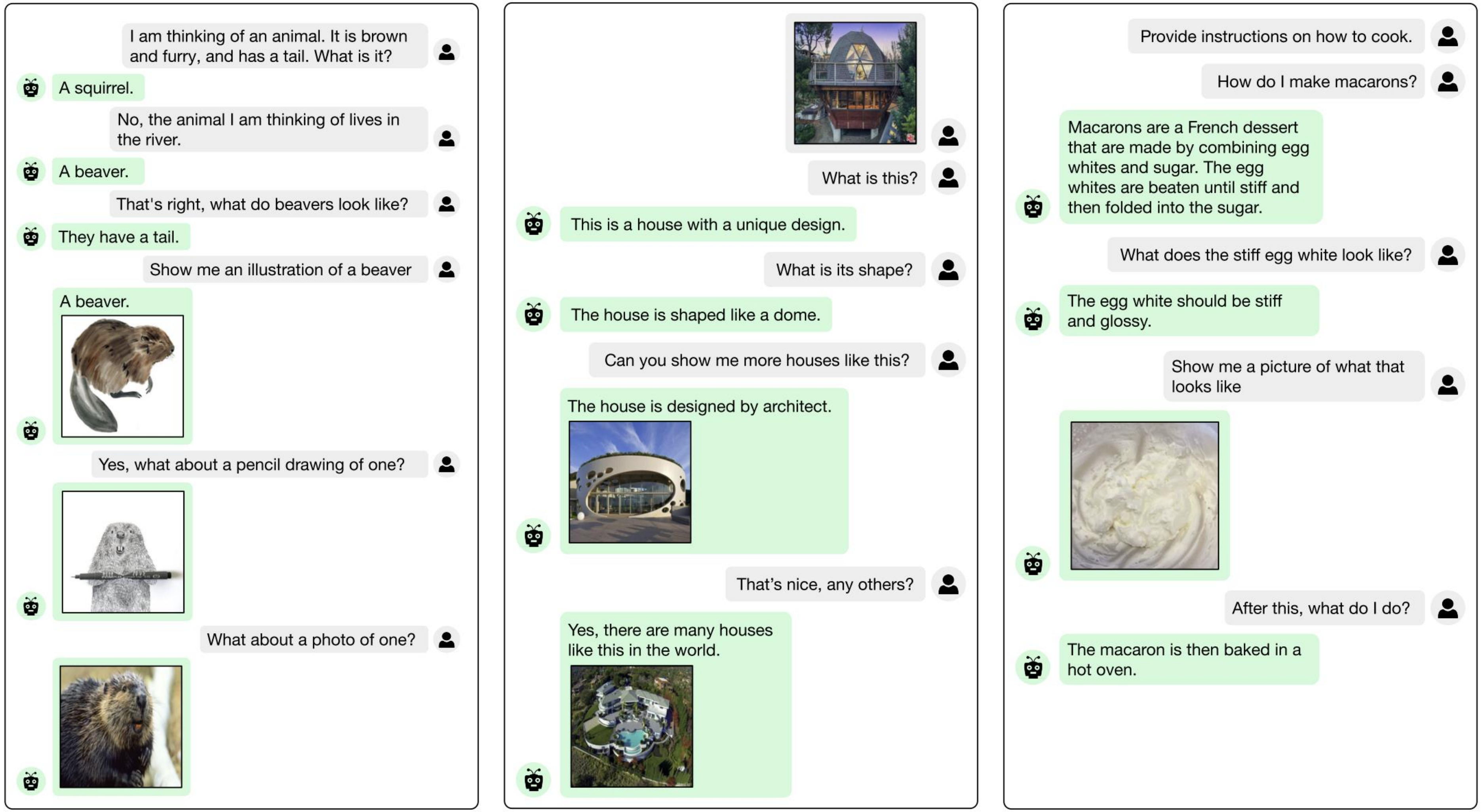
Interleaved images and text

CLIP, with a frozen LLM



[Koh et al., Grounding Language Models to Images for Multimodal Inputs and Outputs. ICML 2023]

Grounding LMs for Image Retrieval

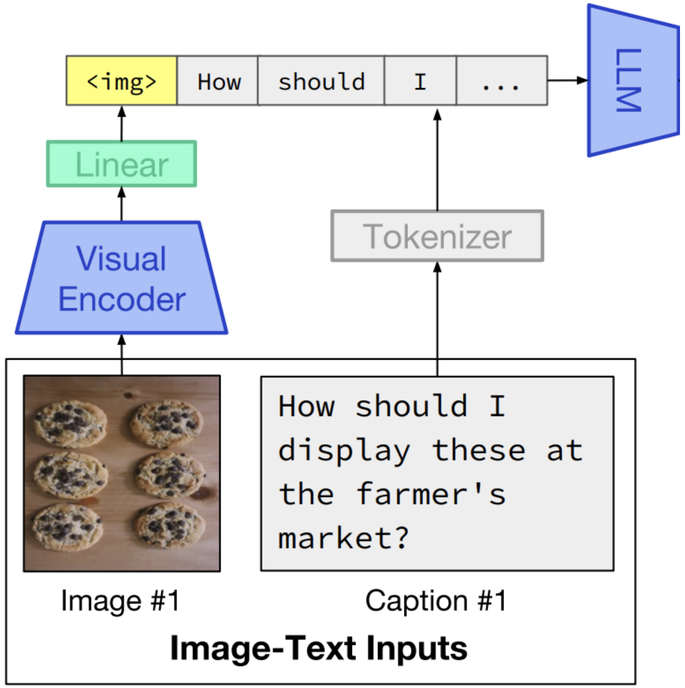


[Koh et al., Grounding Language Models to Images for Multimodal Inputs and Outputs. ICML 2023]

Grounding LMs for Multimodal Generation

Large multimodal models with image generation

Adapter layer (part 2)

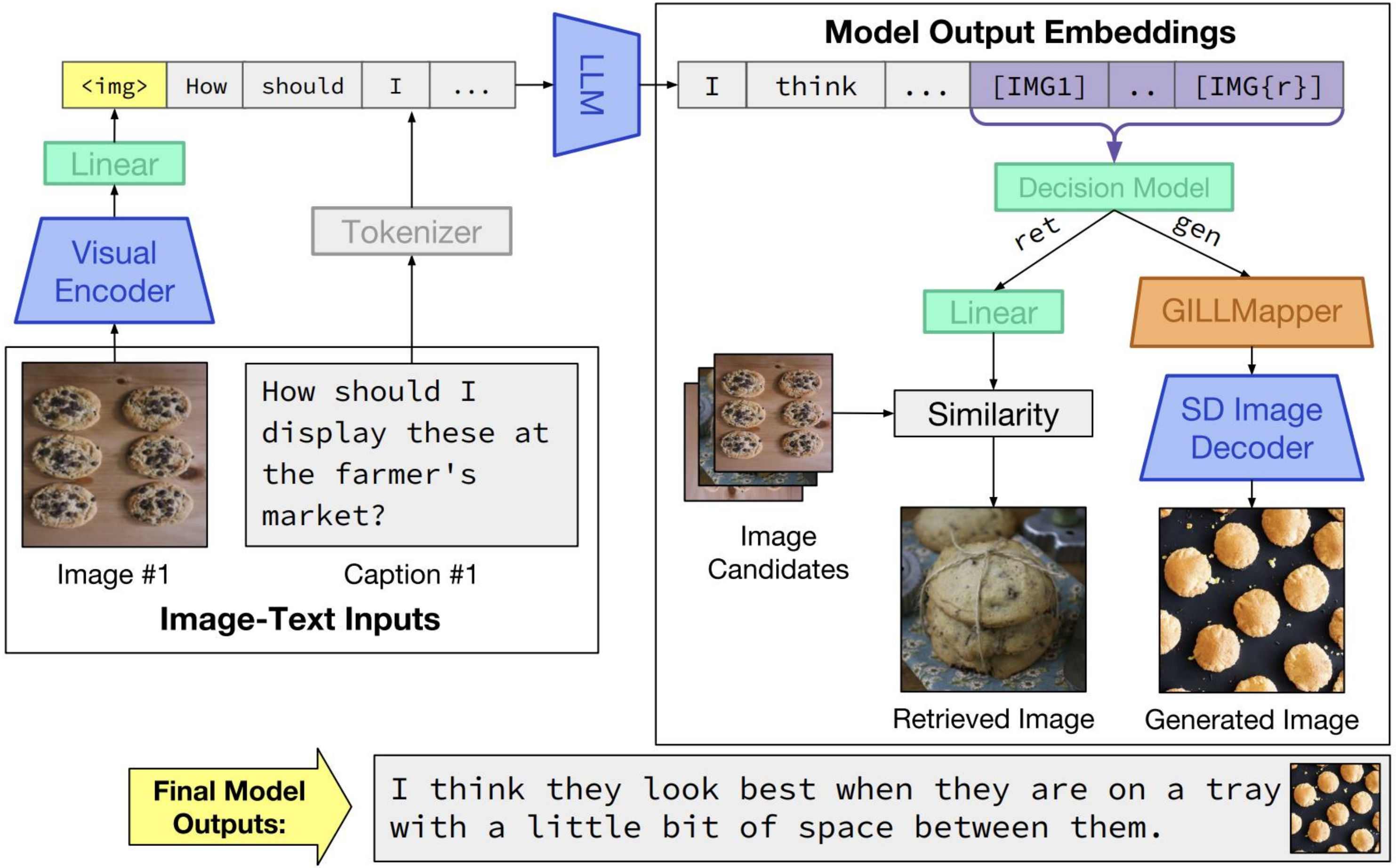


[Koh et al., Generating Images with Multimodal Language Models. NeurIPS 2023]

Grounding LMs for Multimodal Generation

Large multimodal models with image generation

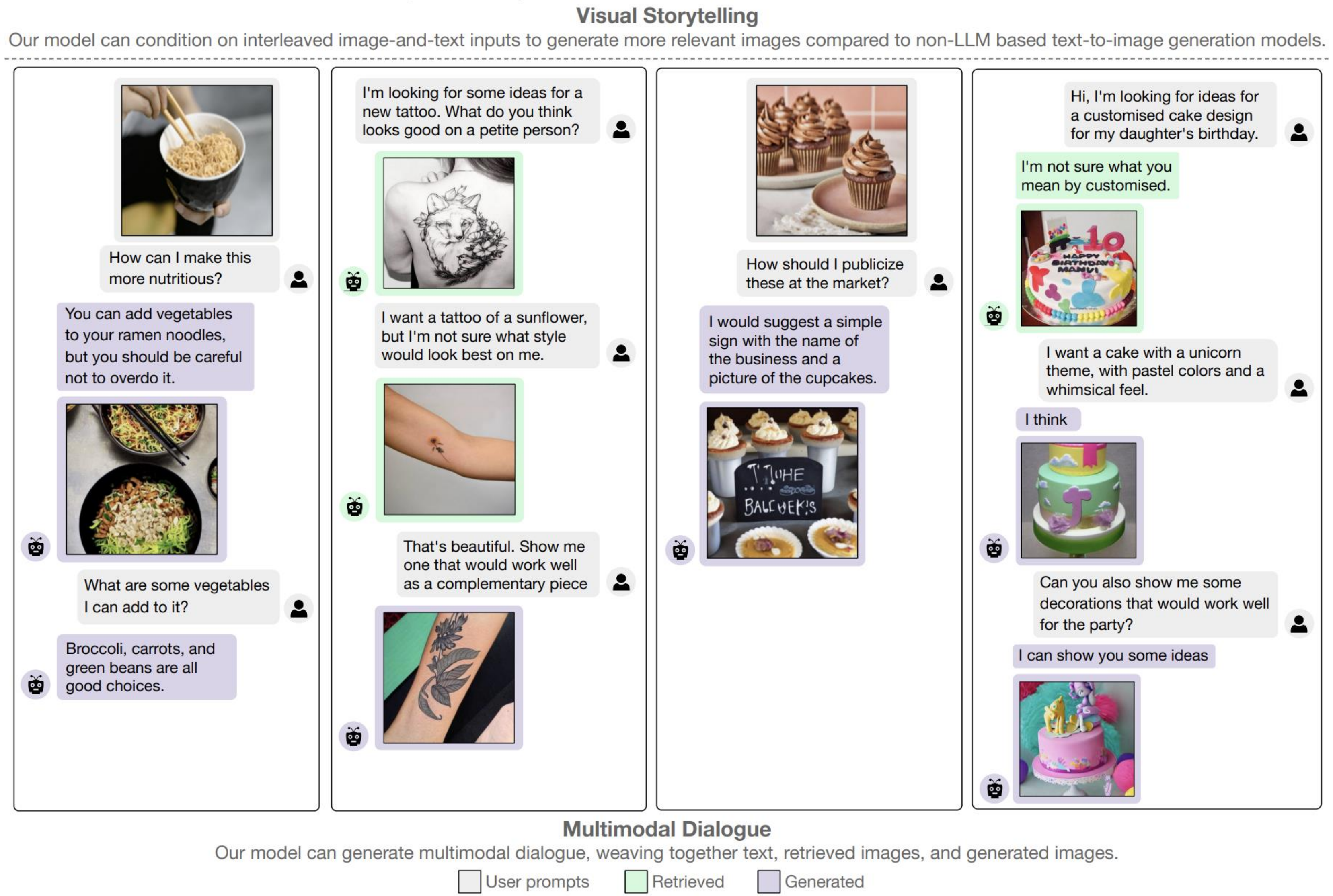
Adapter layer (part 2)



Generation model (part 3)

[Koh et al., Generating Images with Multimodal Language Models. NeurIPS 2023]

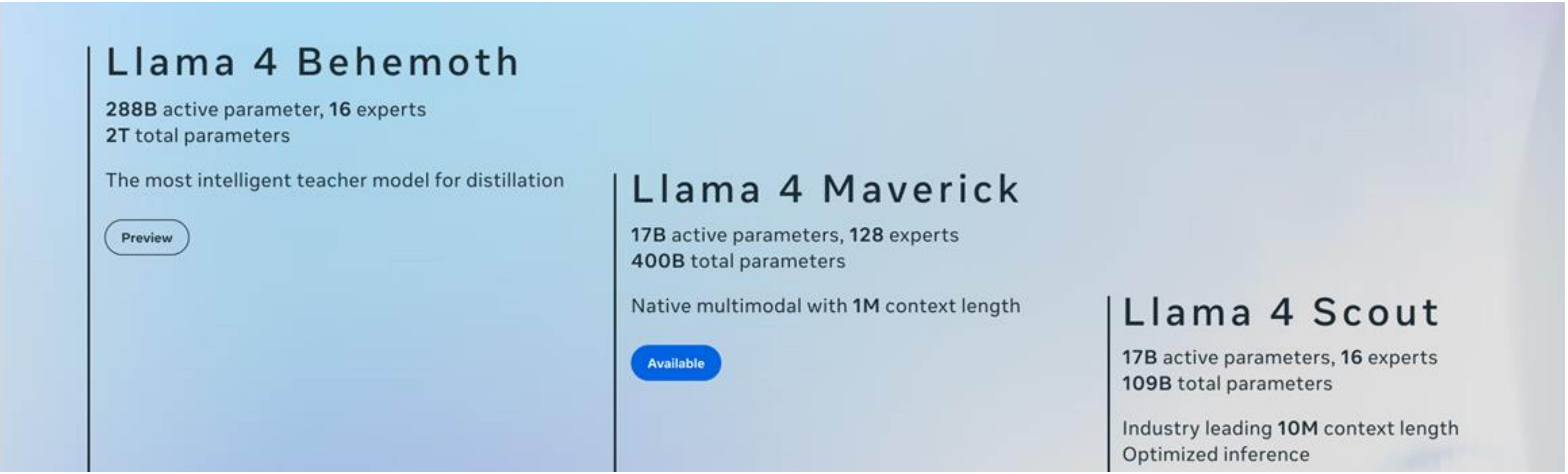
Grounding LMs for Multimodal Generation



multisensory intelligence

Native Multimodal Models

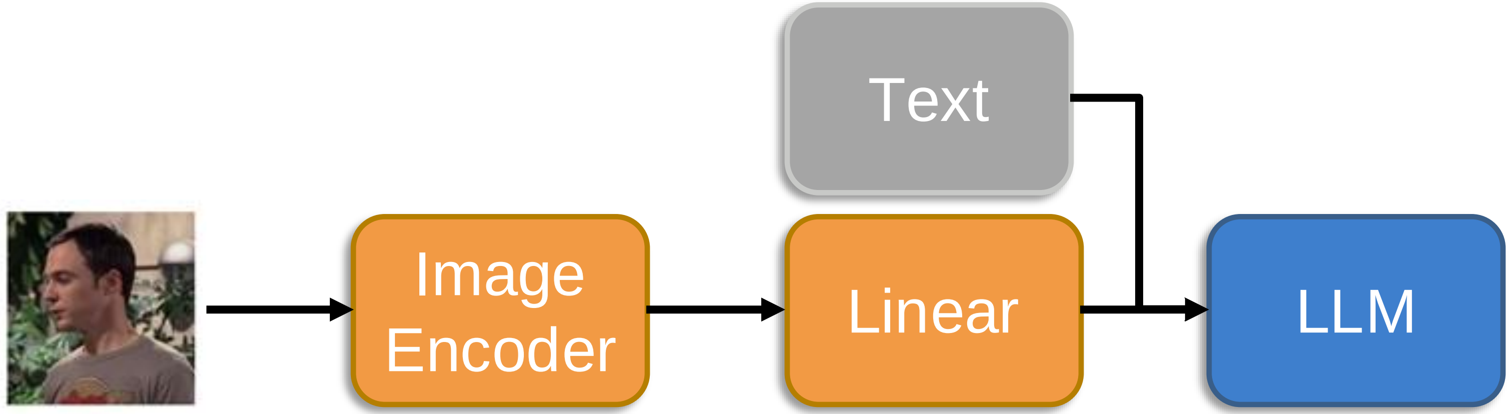
- *Native Multimodal Modals*: LLMs Trained from scratch with multimodal input (instead of finetuning a trained unimodal LLM)
- Largest public model now: 109B - 2T parameters



[The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation, https://ai.meta.com/blog/llama-4-multimodal-intelligence/]

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.



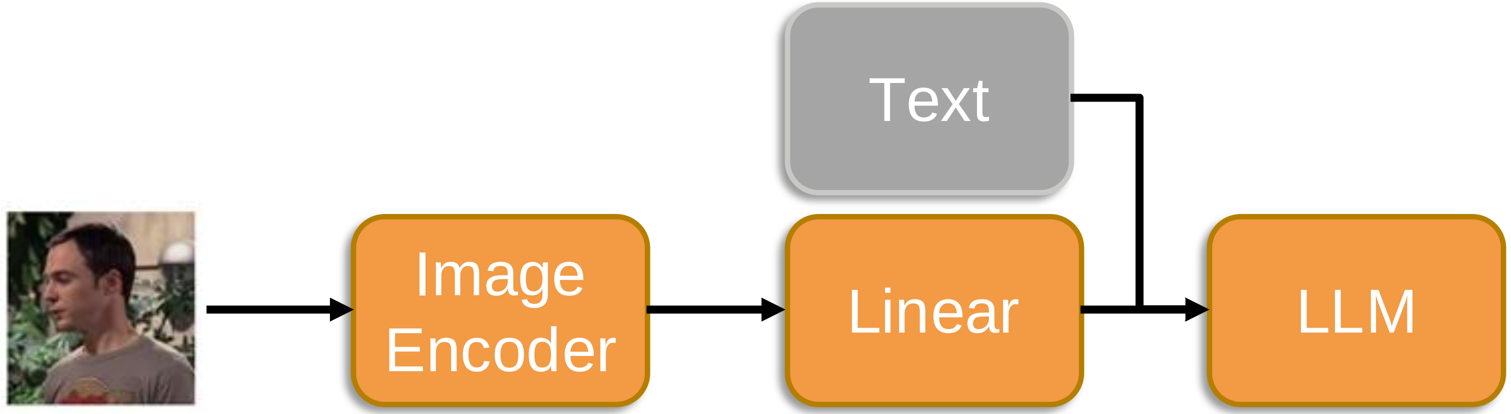
Non-Native VLM

Most current VLMs use this architecture.

[Scaling Laws for Native Multimodal Models Scaling Laws for Native Multimodal Models, https://arxiv.org/abs/2504.07951]

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 Modals: LLMs Trained from scratch with multimodal input
    - **Late fusion**: Image patches -> Image Encoder -> Linear -> LLM.
    - Early fusion: Image patches -> Linear -> LLM (No image encoder!)

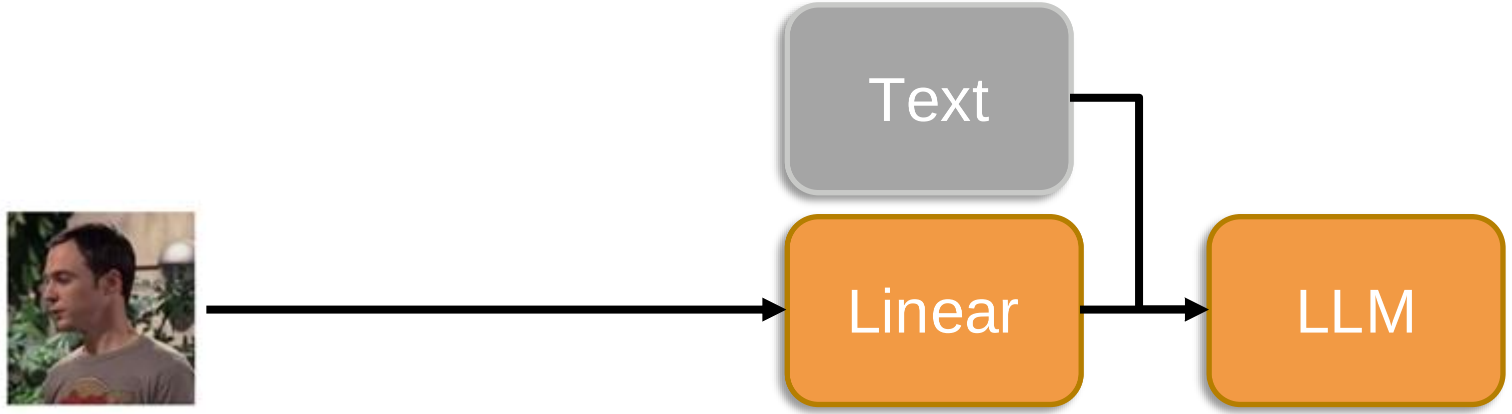


Late fusion

[Scaling Laws for Native Multimodal Models Scaling Laws for Native Multimodal Models, https://arxiv.org/abs/2504.07951]

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 Modals: LLMs Trained from scratch with multimodal input
    - Late fusion: Image patches -> Image Encoder -> Linear -> LLM.
    - **Early fusion**: Image patches -> Linear -> LLM (No image encoder!)

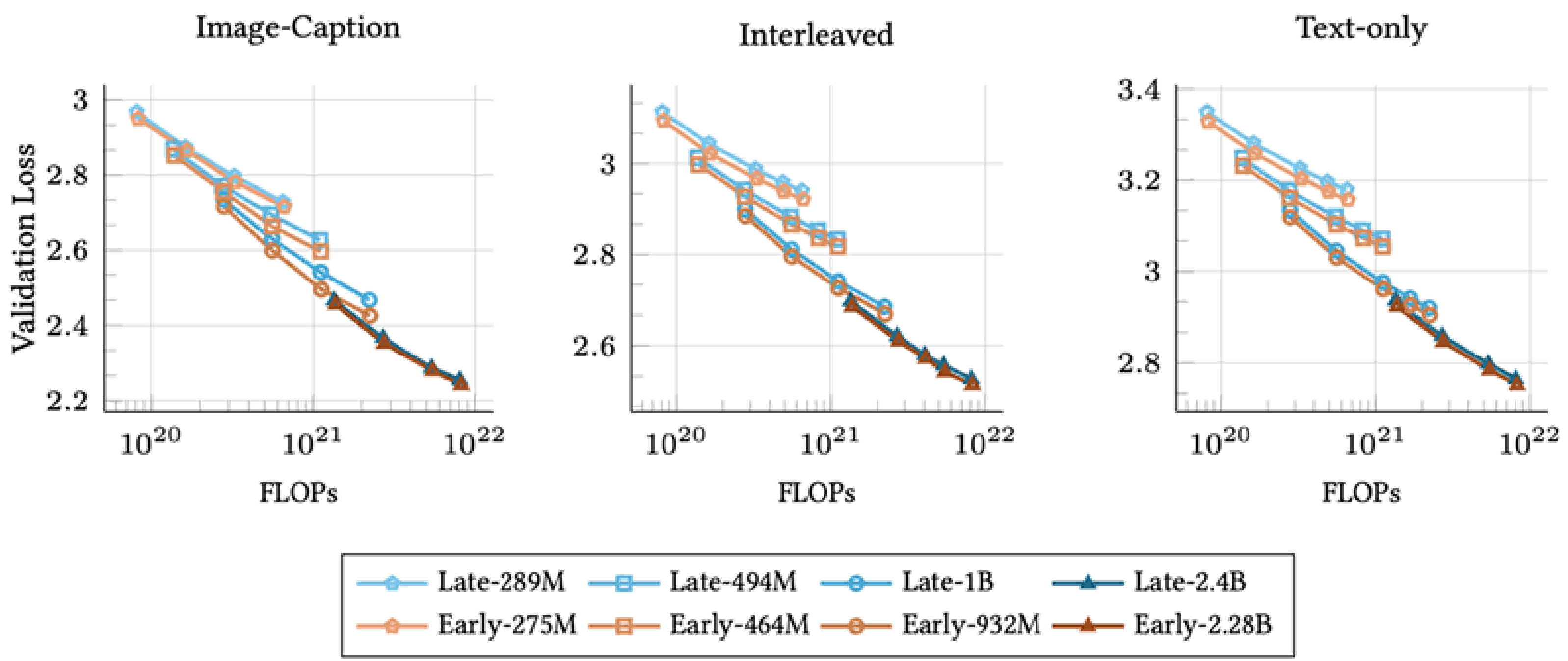


Early fusion

[Scaling Laws for Native Multimodal Models Scaling Laws for Native Multimodal Models, https://arxiv.org/abs/2504.07951]

Scaling Laws for Native Multimodal Models

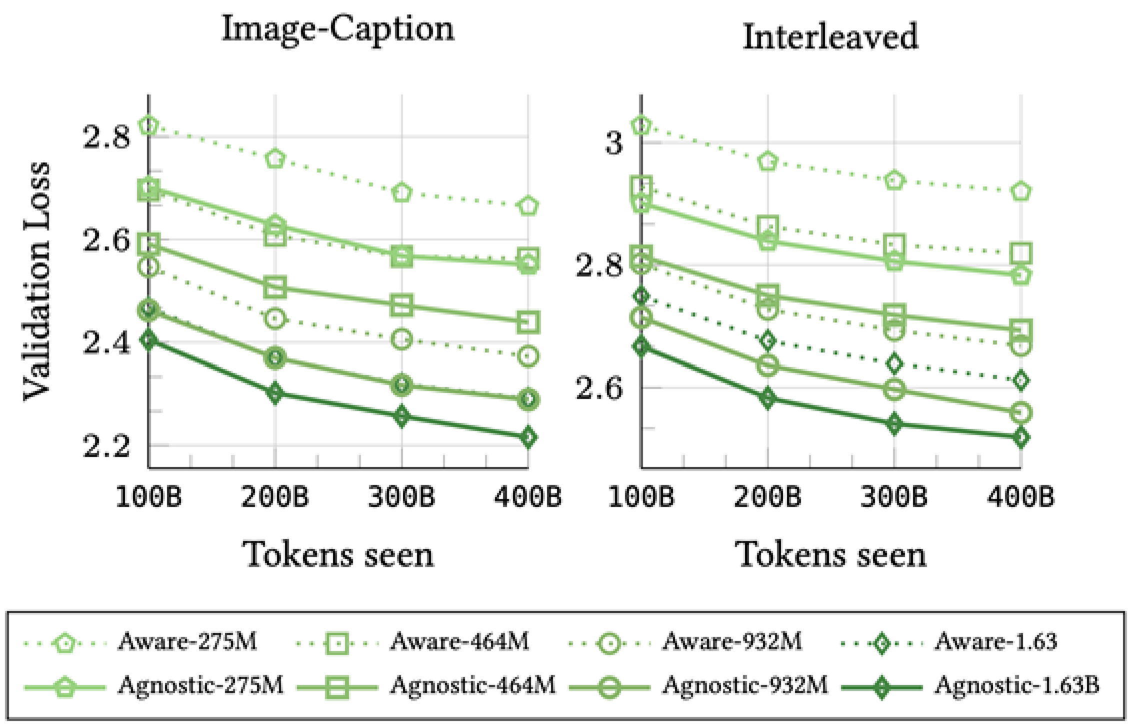
- Early fusion models hold small advantage on small scales.
- On larger scales, **both architectures perform similarly.** (We don't actually need image encoders!)
- **NMMs scale similarly to unimodal LLMs**, with slightly varying scaling exponents depending on the target data type and training mixture



[Scaling Laws for Native Multimodal Models Scaling Laws for Native Multimodal Models, https://arxiv.org/abs/2504.07951]

Scaling Laws for Native Multimodal Models

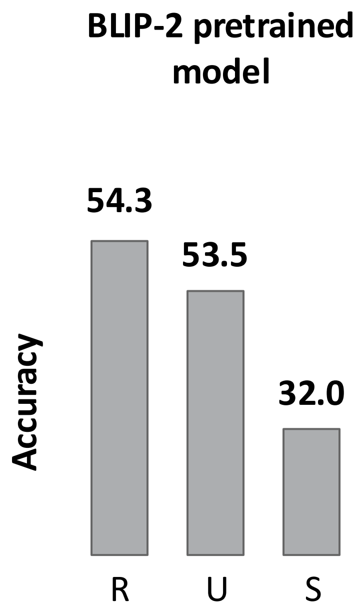
- Sparse structure like **MOE significantly benefits NMMs** at the same inference cost
- In an MOE structure, Modality-aware design (having separate image/text experts) performs **worse** than modality-agnostic design (unified experts for both image/text tokens)



[Scaling Laws for Native Multimodal Models Scaling Laws for Native Multimodal Models, https://arxiv.org/abs/2504.07951

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

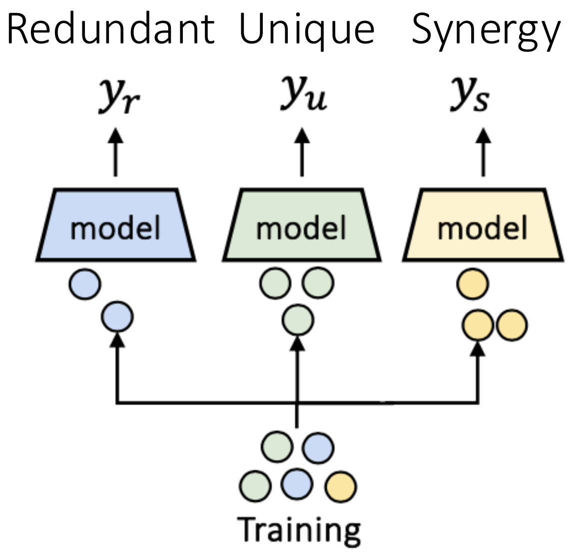
Neutral tone + straight face

[Yu et al., MMOE: Enhancing Multimodal Language Models with Mixture of Multimodal Interaction Experts. EMNLP 2024]

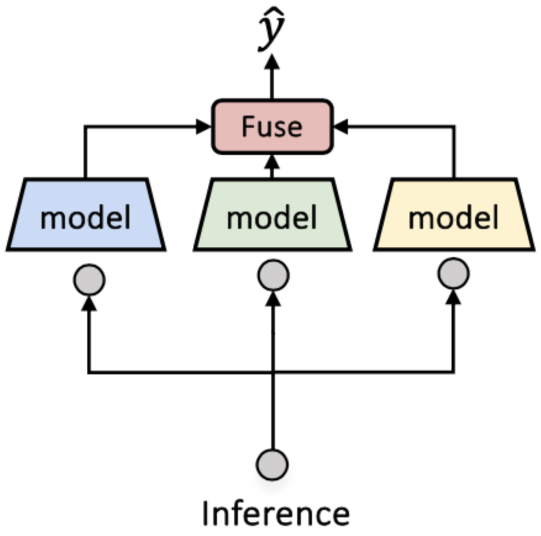
multisensory intelligence

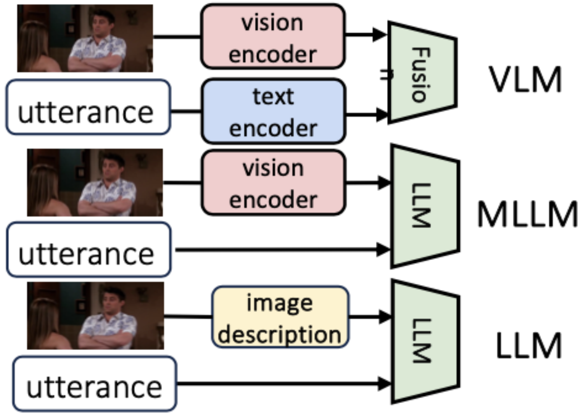
Mixture of Multimodal Interaction Experts

One model for everything -> specialized models for each interaction



Datapoint-level
quantification
of interactions



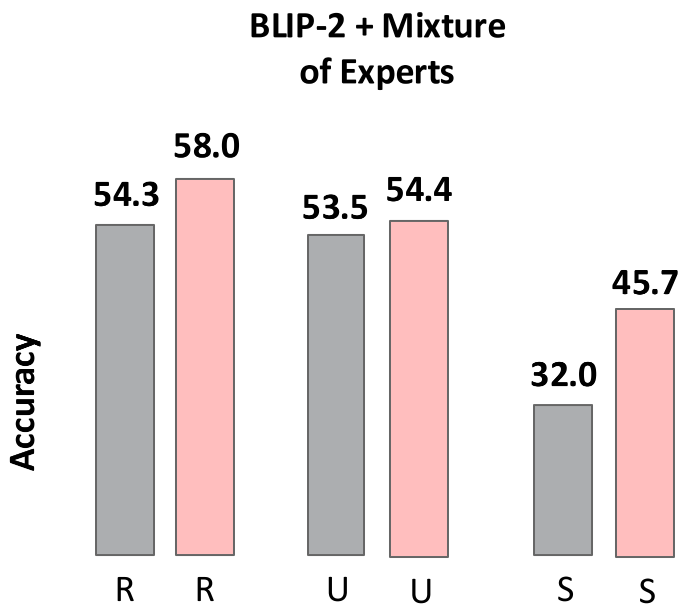


[Yu et al., MMOE: Enhancing Multimodal Language Models with Mixture of Multimodal Interaction Experts. EMNLP 2024]

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]

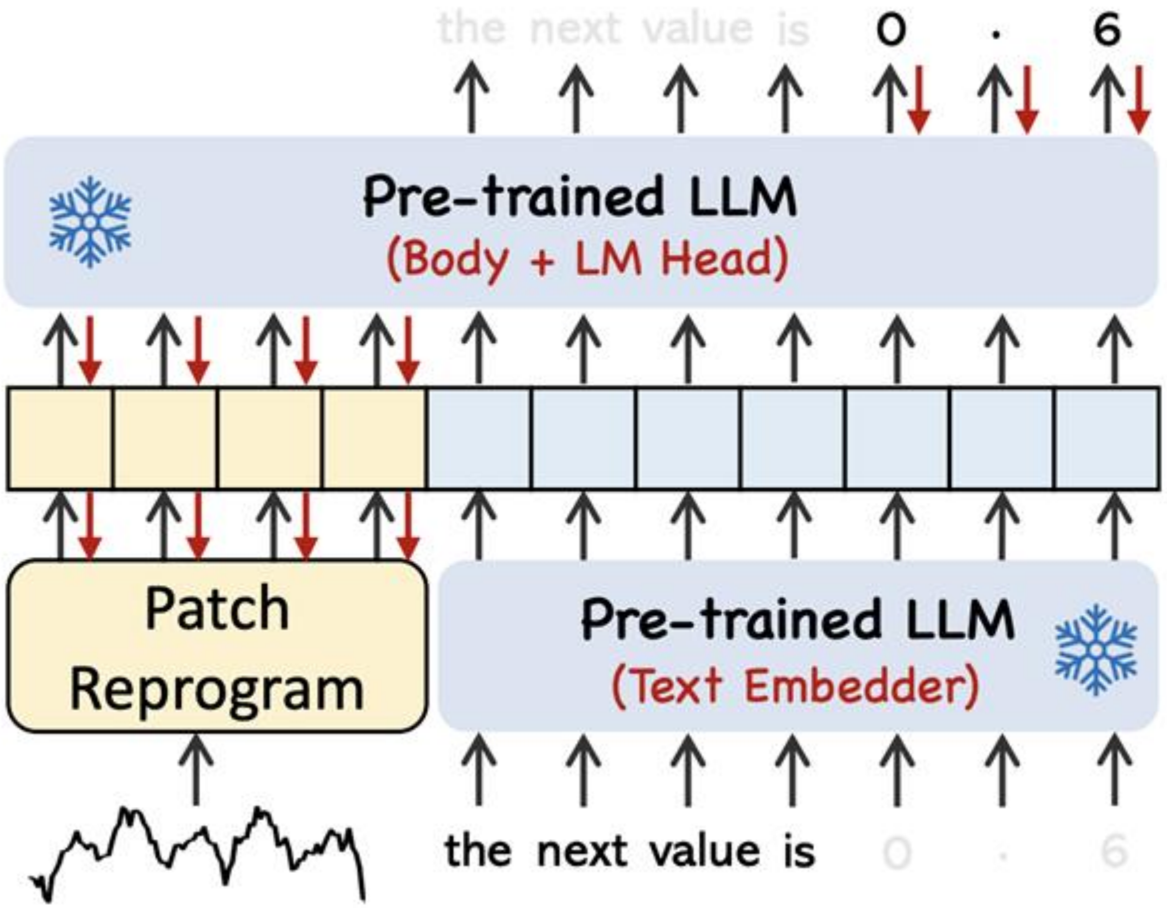
[Hessel et al., ACL 23]

[Yu et al., MMOE: Enhancing Multimodal Language Models with Mixture of Multimodal Interaction Experts. EMNLP 2024]

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

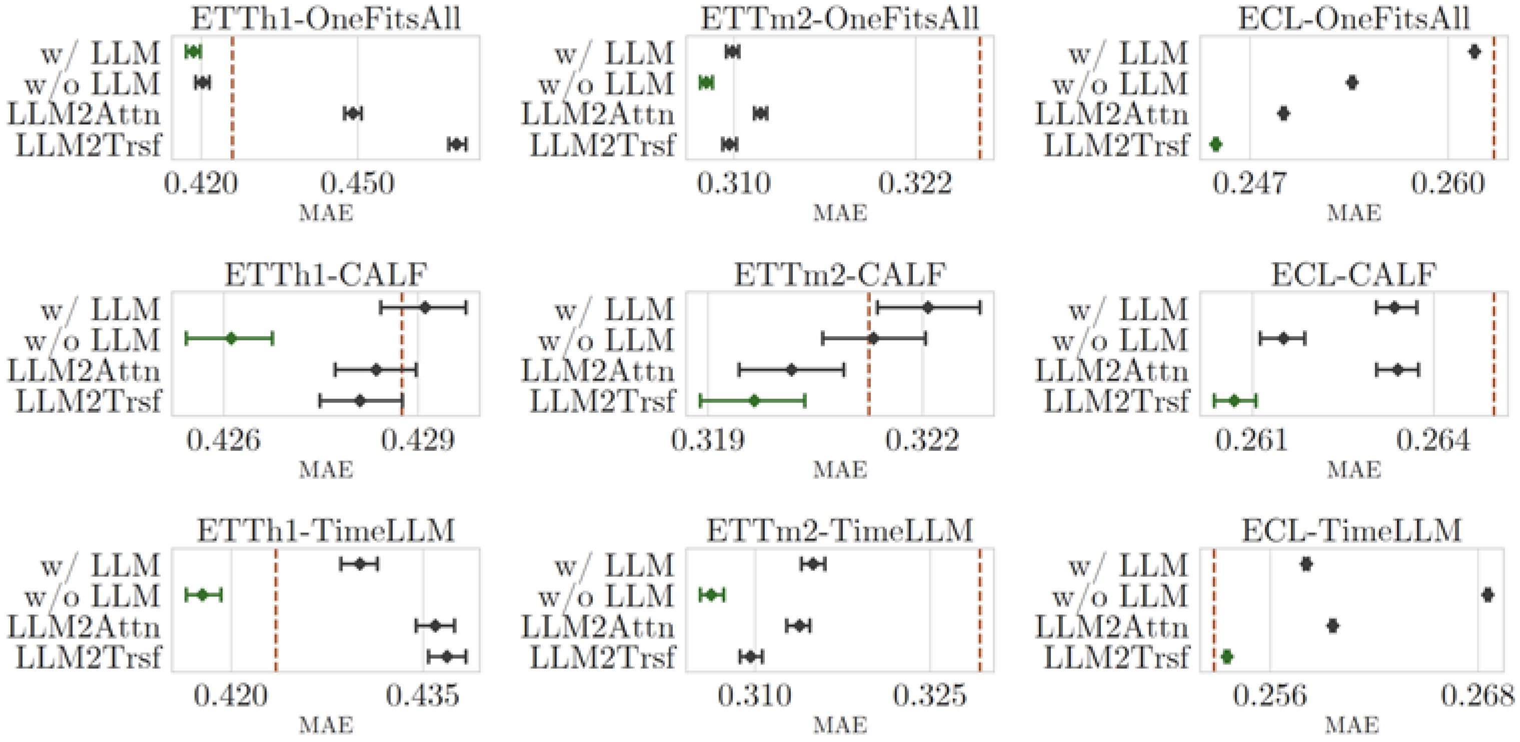


[Time-LLM: Time Series Forecasting by Reprogramming Large Language Models, https://arxiv.org/abs/2310.01728]

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



[Are Language Models Actually Useful for Time Series Forecasting? https://arxiv.org/abs/2406.16964]

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