Multimodal AI

Lecture 5.1 – Large Multimodal Models

Paul Liang
Assistant Professor
MIT Media Lab & MIT EECS

https://pliang279.github.io
ppliang@mit.edu
@pliang279

mit media lab

MIT EECS

Assignments for This Coming Week

HW2 due tomorrow Wed (3/4).

HW2 presentation this Thurs + in-depth tutorial on implementing multimodal LLMs.

HW3 (multimodal LLMs) out later this week.

Today's lecture

1. Multimodal transformers & foundation models
2. Adapting LLMs into multimodal LLMs
3. From text to multimodal generation
4. Latest directions

Recap: Large Language Models

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]

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.

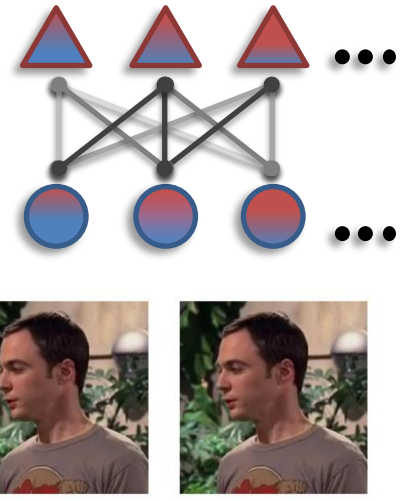
Counterfactual: What if these people were from a different society or culture?

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

multisensory intelligence

From Large Language Models to Multimodal Models

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)

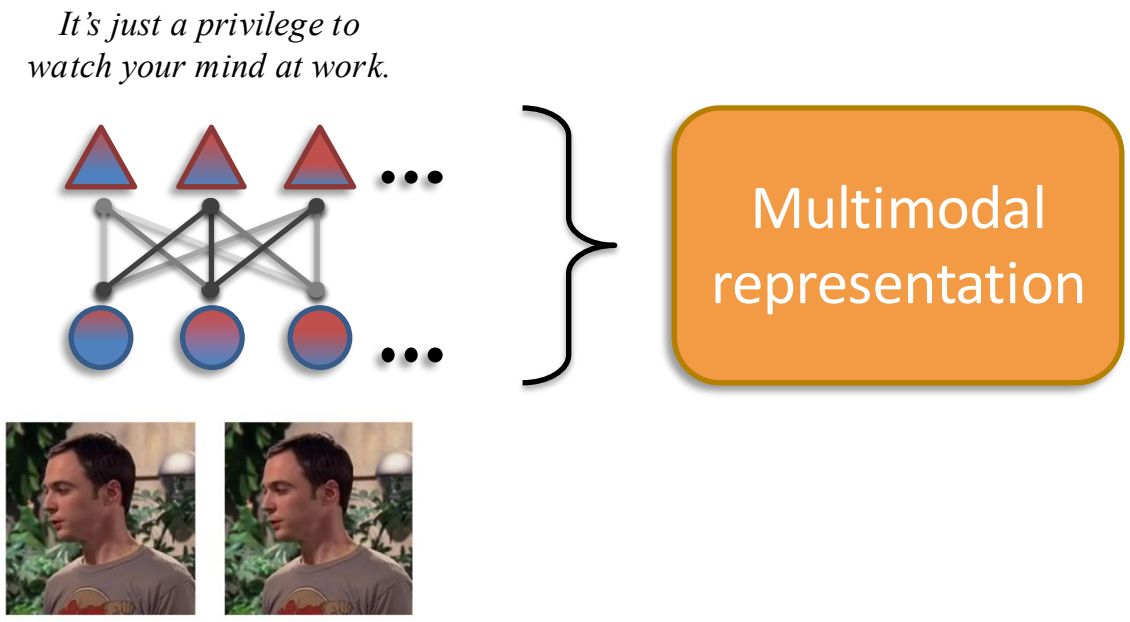
Here's a story of them in a different culture...

(generate future episodes)

multisensory intelligence

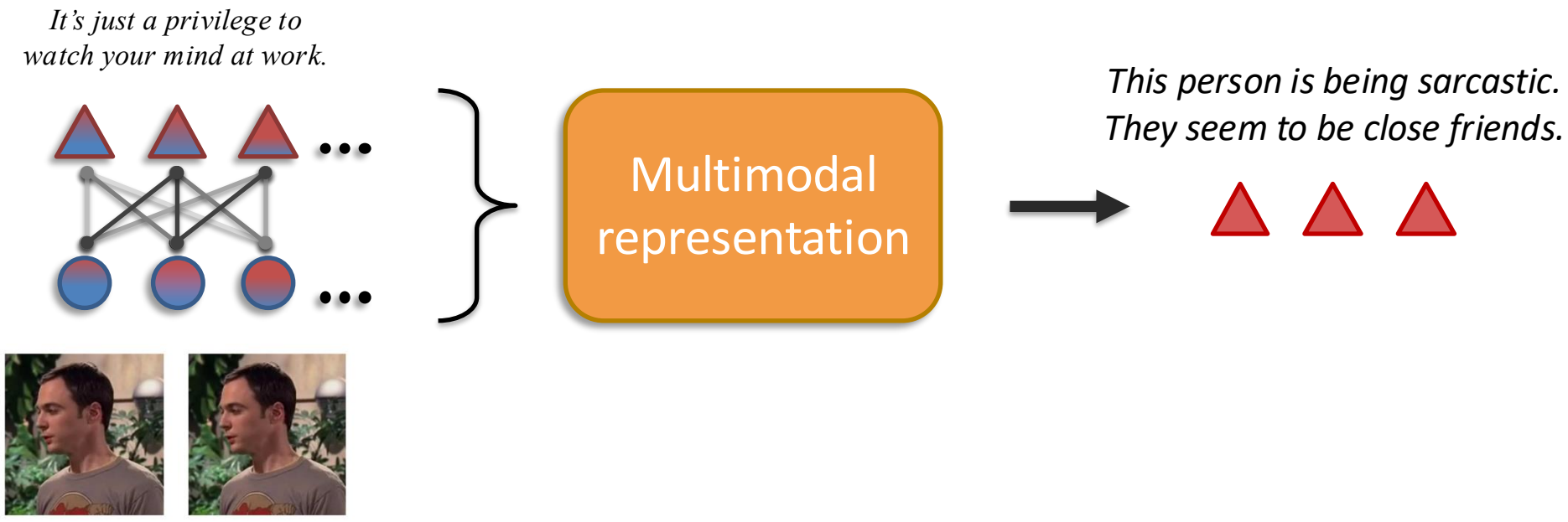
Lecture outline

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



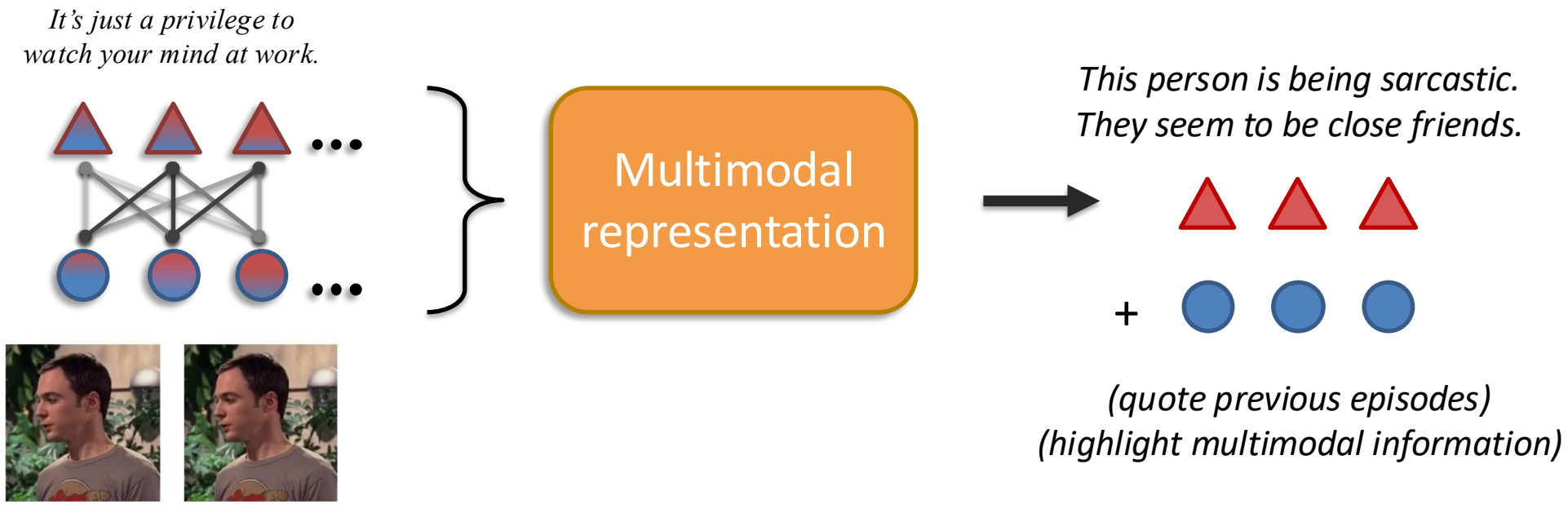
Lecture outline

Part 2: Adapting large language models for multimodal text generation

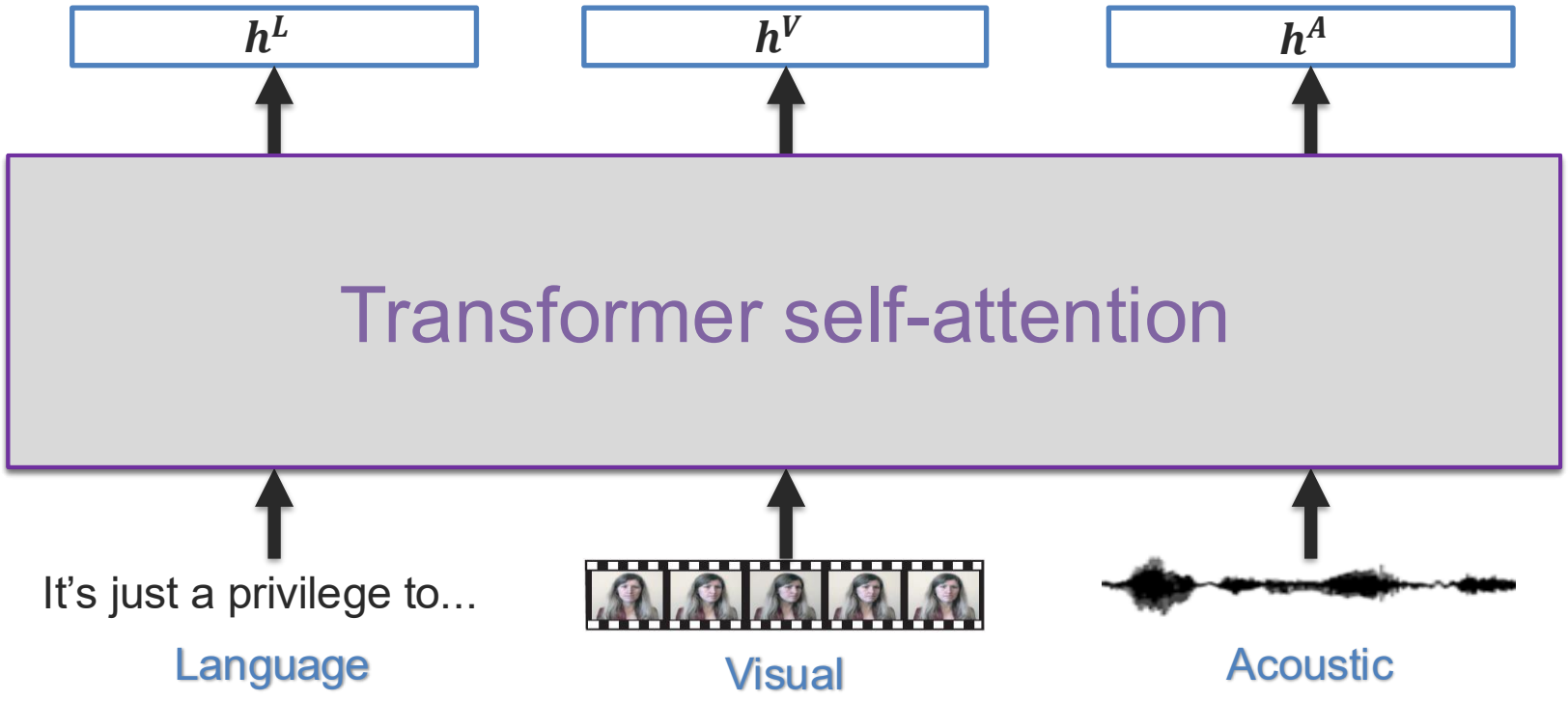


Lecture outline

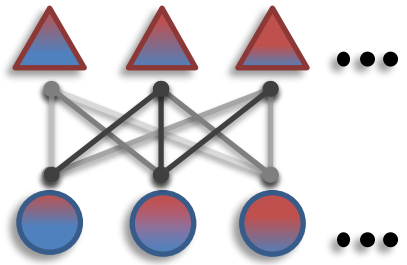
Part 3: Enabling text and image generation



Large Multimodal Models



Contextualized representation



Implicit alignment + representation

multisensory intelligence

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

privilege	0.7			0.3
mind	1.0			

vocal emphasis
eye-roll

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

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

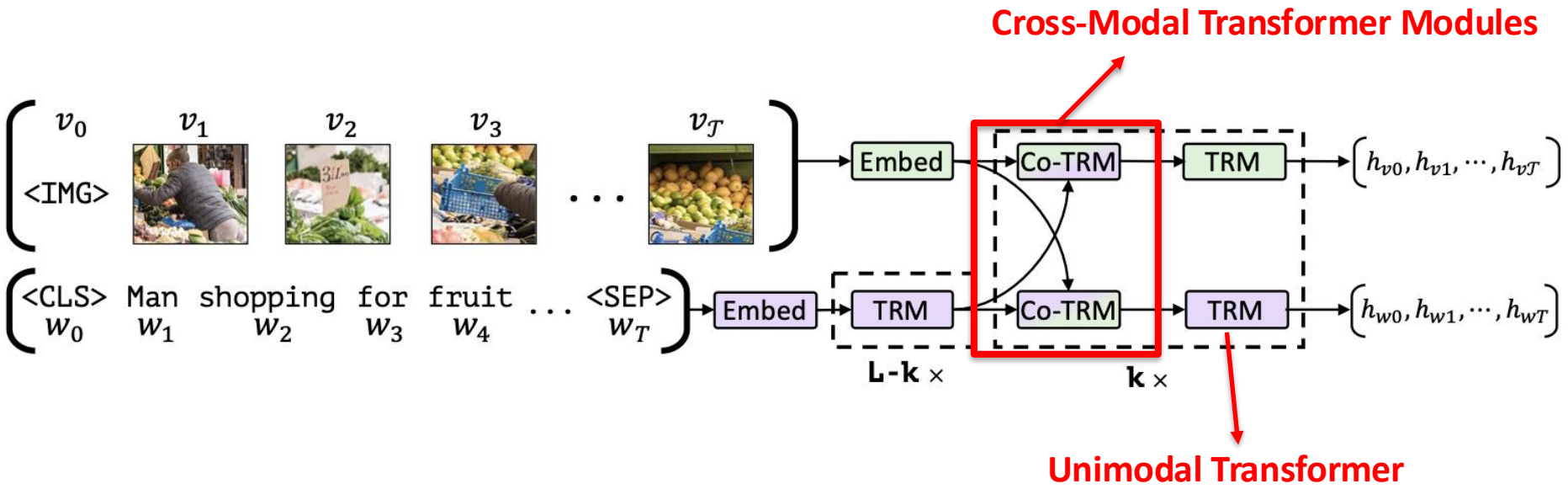
4 x d

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

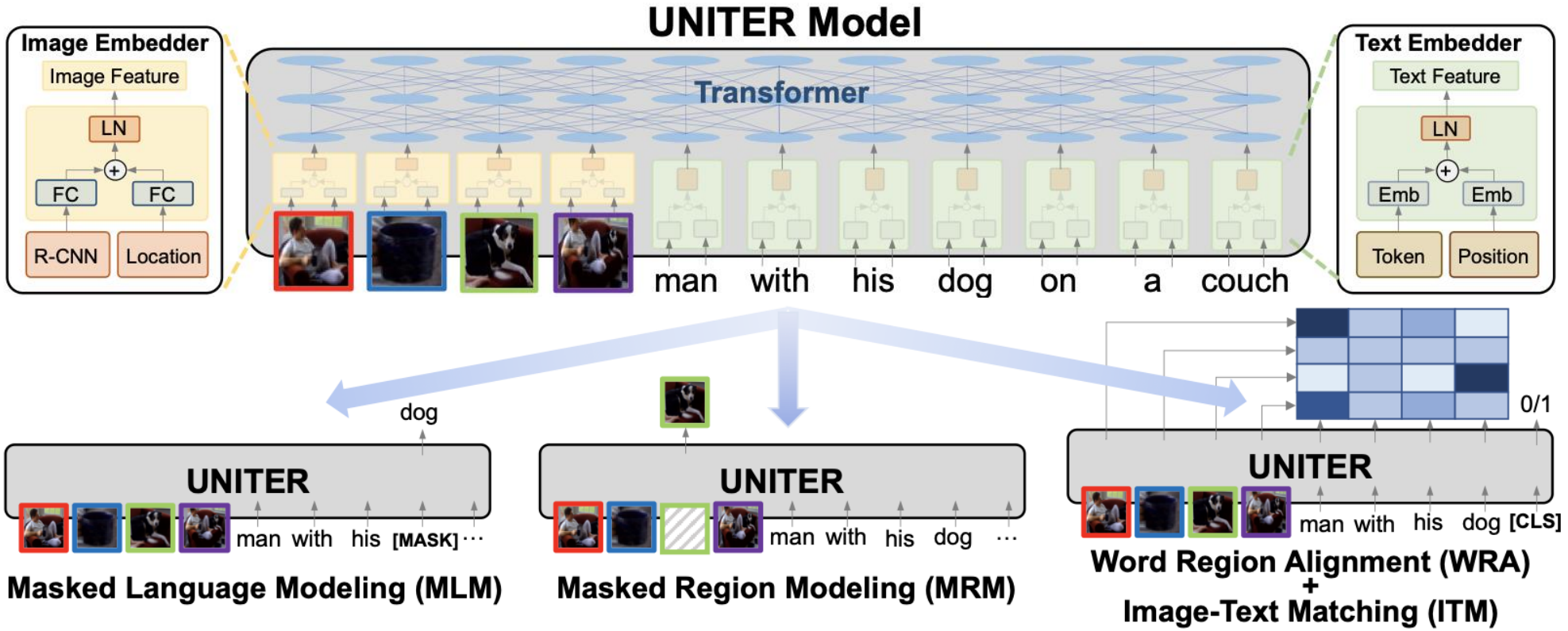


[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

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

multisensory intelligence

Multimodal Cross-attention Transformers



Masking words

Masking images

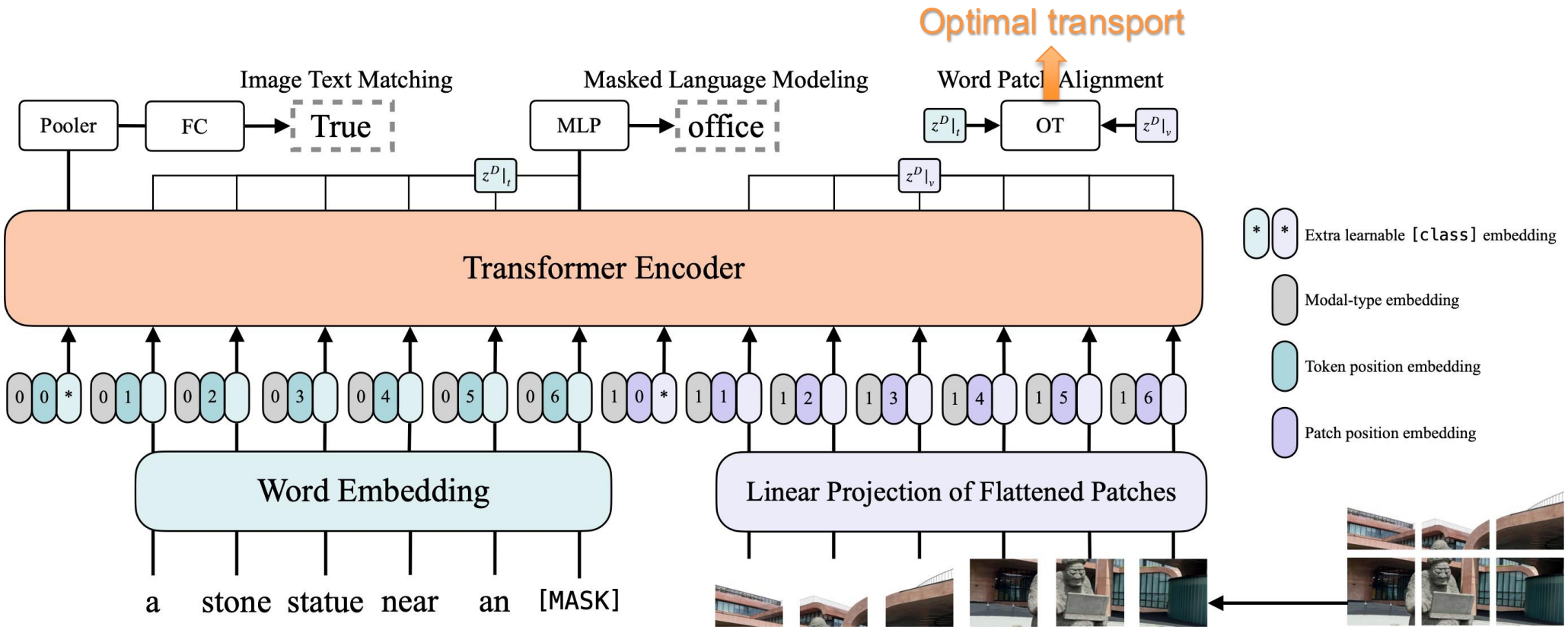
Does the sentence match the image?

[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]

multisensory intelligence

Visual-and-Language Transformer (ViLT) (≈ BERT + ViT)



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

multisensory intelligence

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

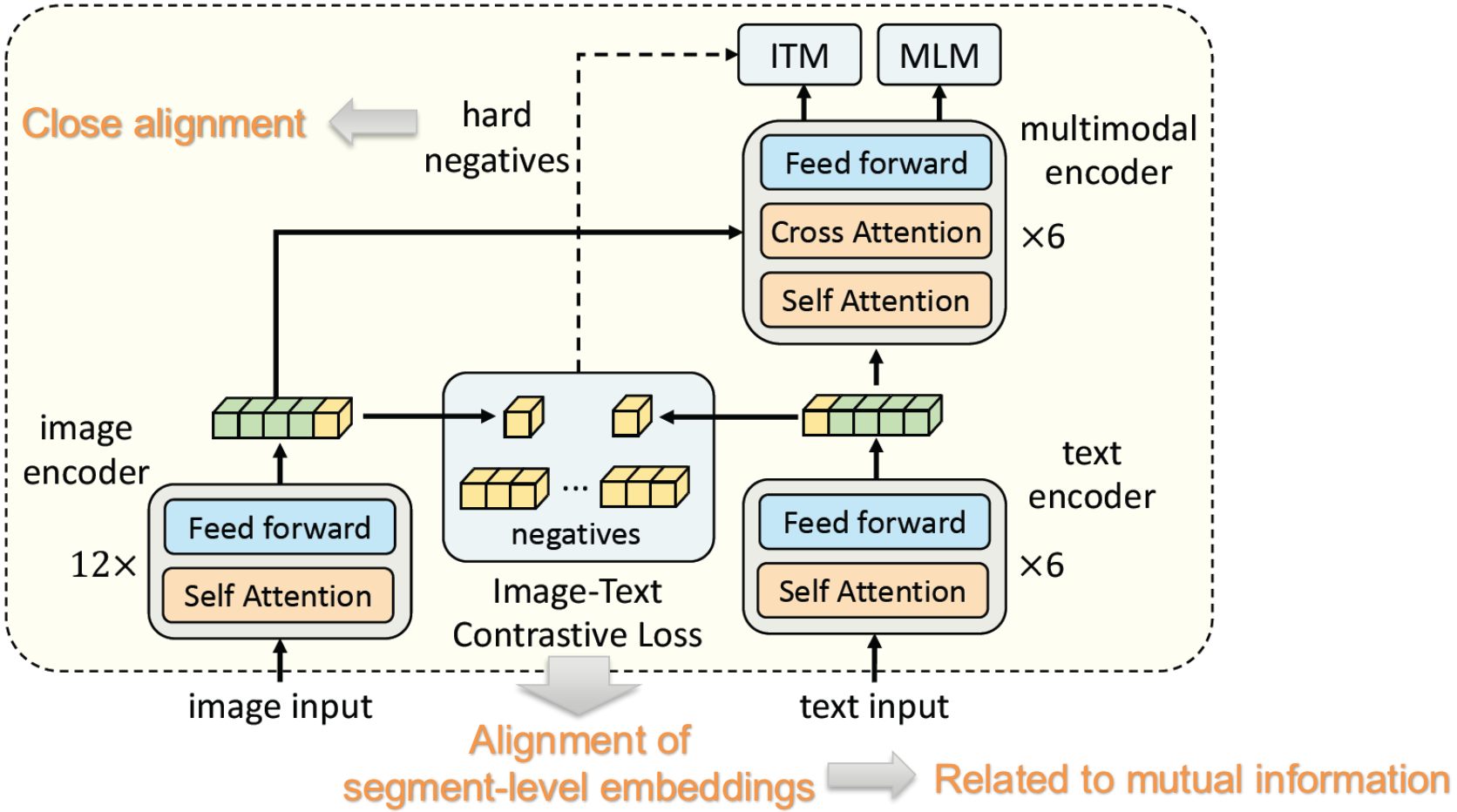

painting


plant

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

multisensory intelligence

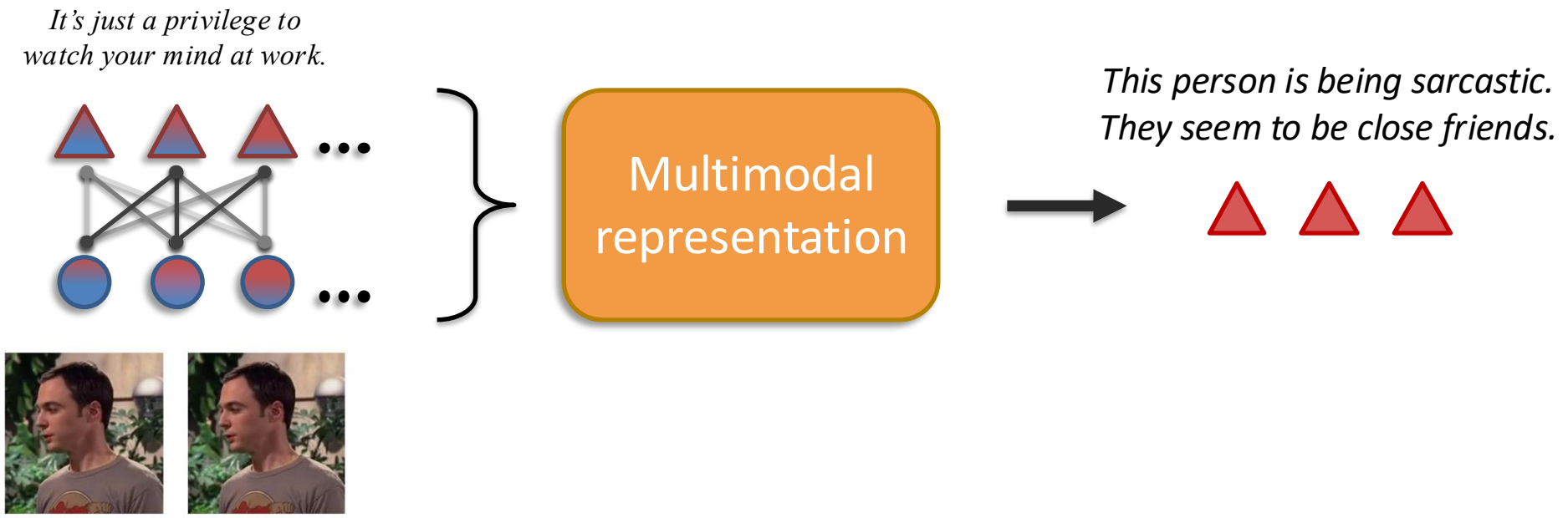
ALBEF: Align Before Fusion (≈ BERT + ViT + CLIP-ish)



[Li et al., Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. NeurIPS 2021]

multisensory intelligence

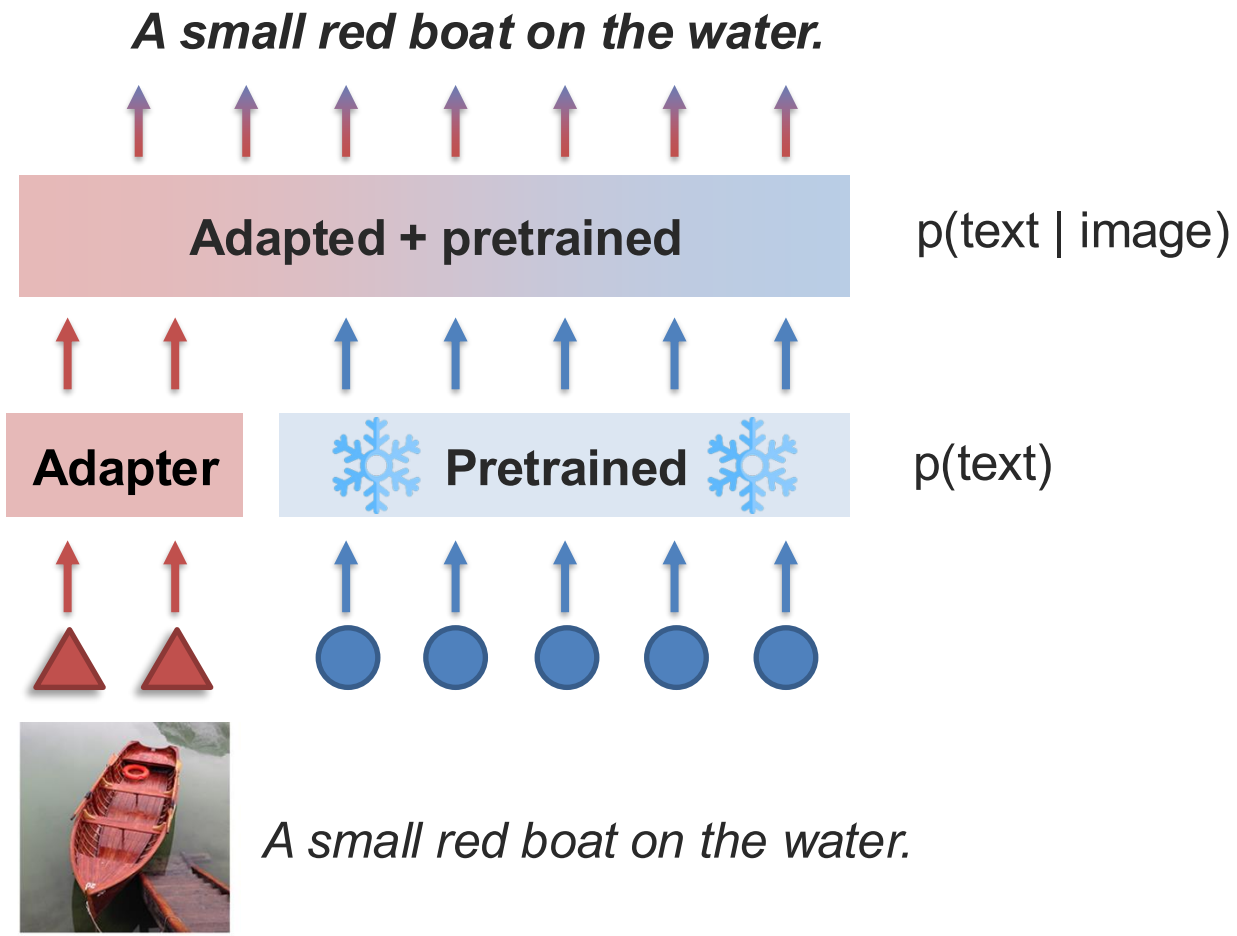
Adapting Large Language Models to Multimodal



Adapting Large Language Models

Conditioning via prefix tuning

Modeling p(text | image):



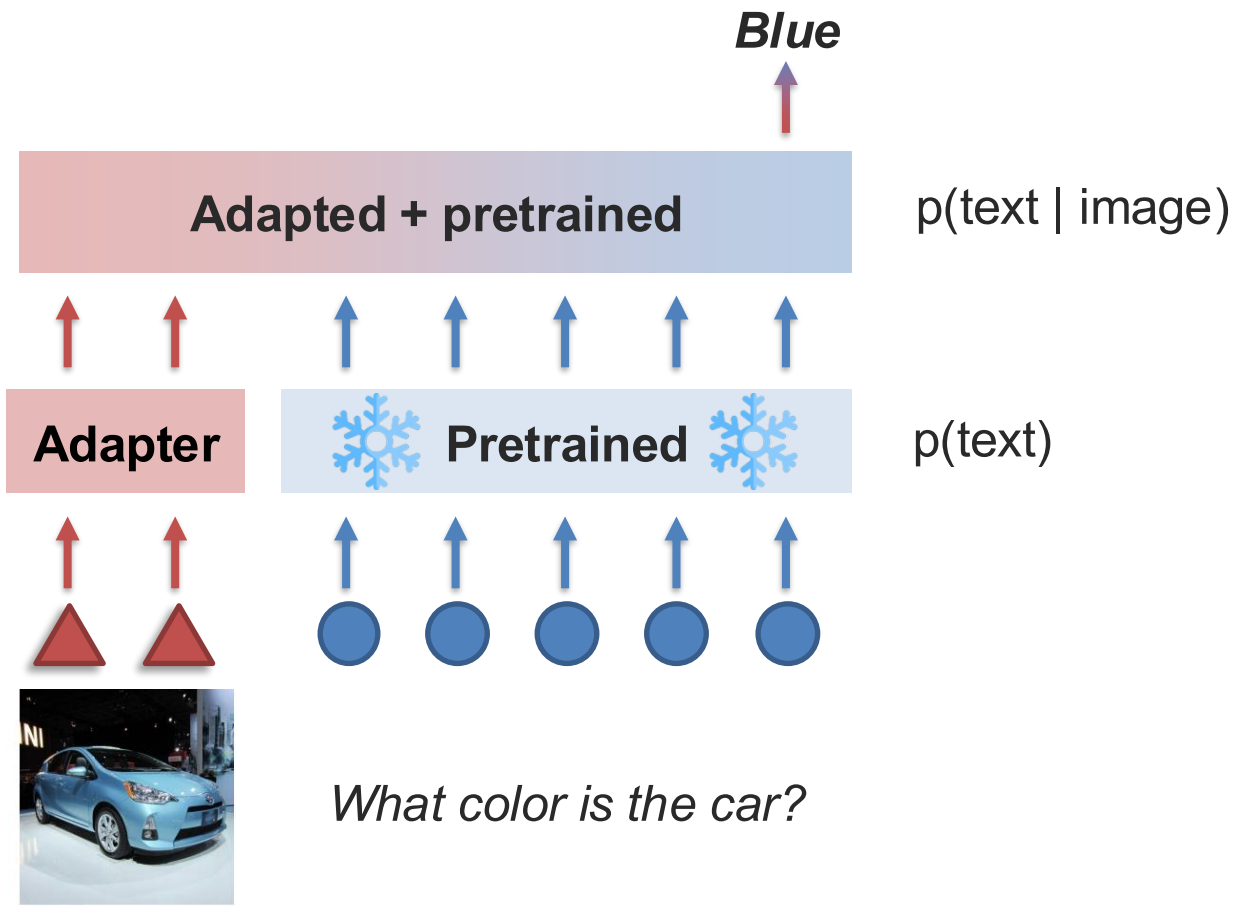
[Tsimpoukelli et al., Multimodal Few-Shot Learning with Frozen Language Models. NeurIPS 2021]

multisensory intelligence

Adapting Large Language Models

Conditioning via prefix tuning

Modeling p(text | image):



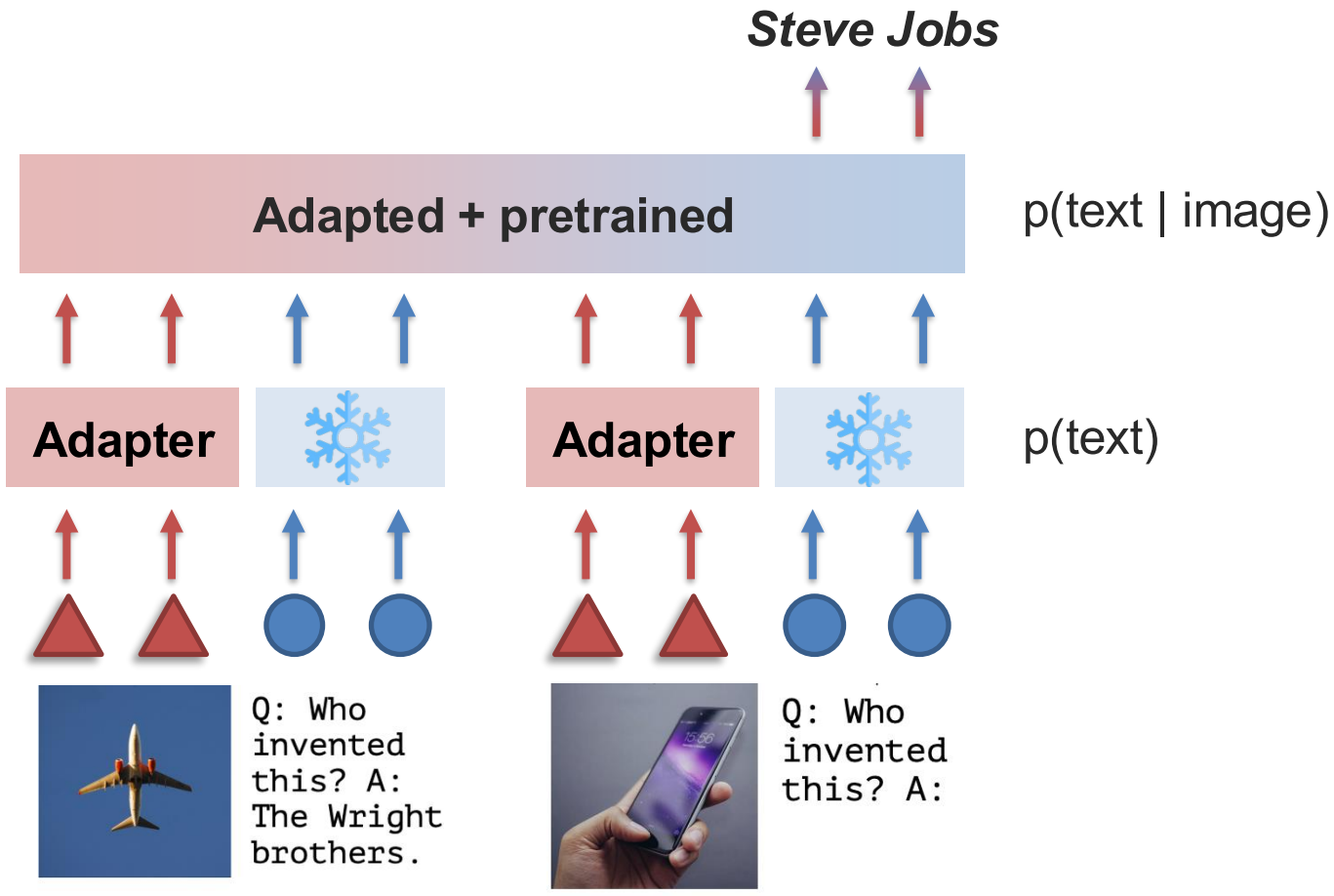
[Tsimpoukelli et al., Multimodal Few-Shot Learning with Frozen Language Models. NeurIPS 2021]

multisensory intelligence

Adapting Large Language Models

Conditioning via prefix tuning

Modeling p(text | image):



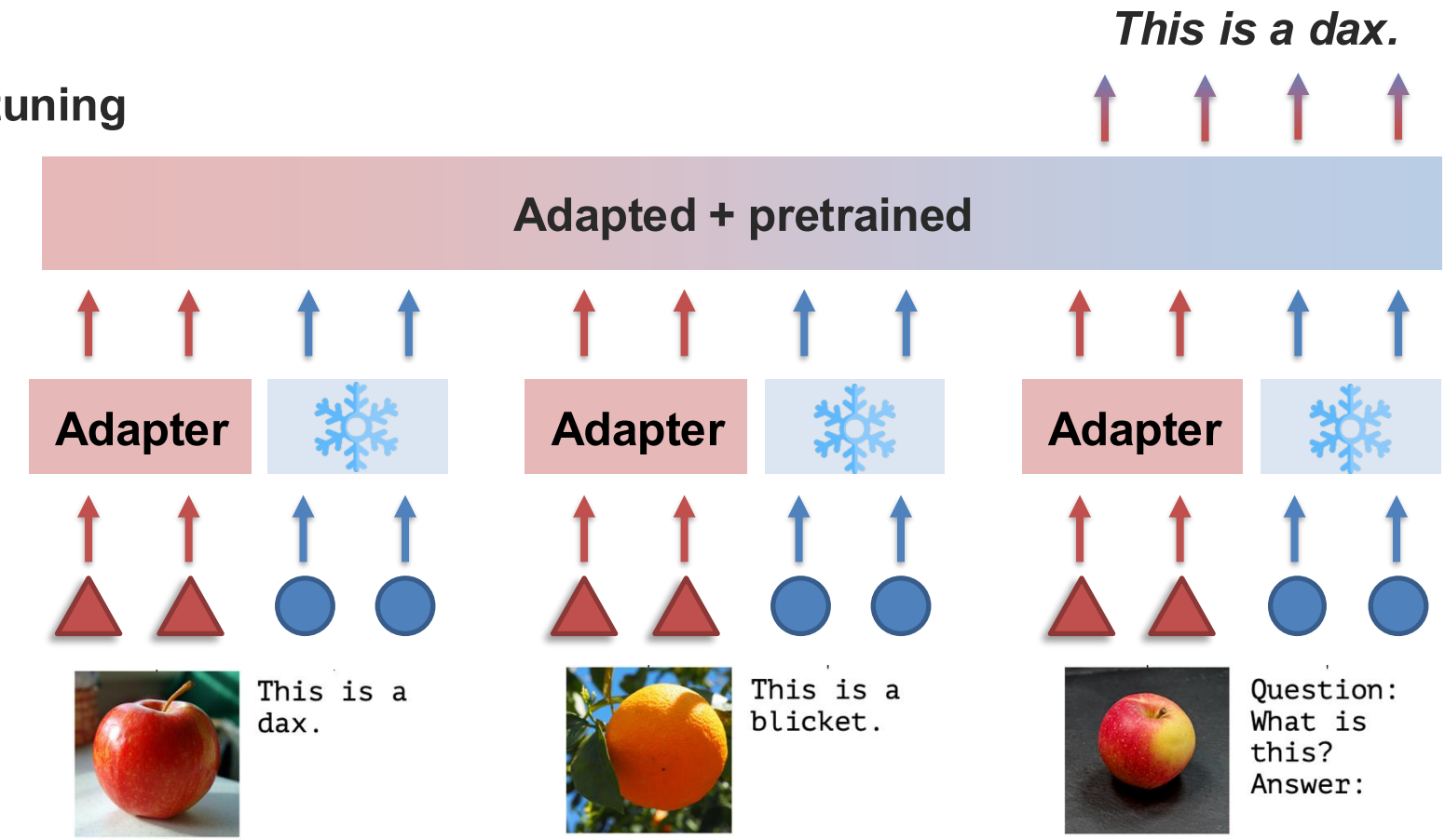
[Tsimpoukelli et al., Multimodal Few-Shot Learning with Frozen Language Models. NeurIPS 2021]

multisensory intelligence

Adapting Large Language Models

Conditioning via prefix tuning

Modeling p(text | image):

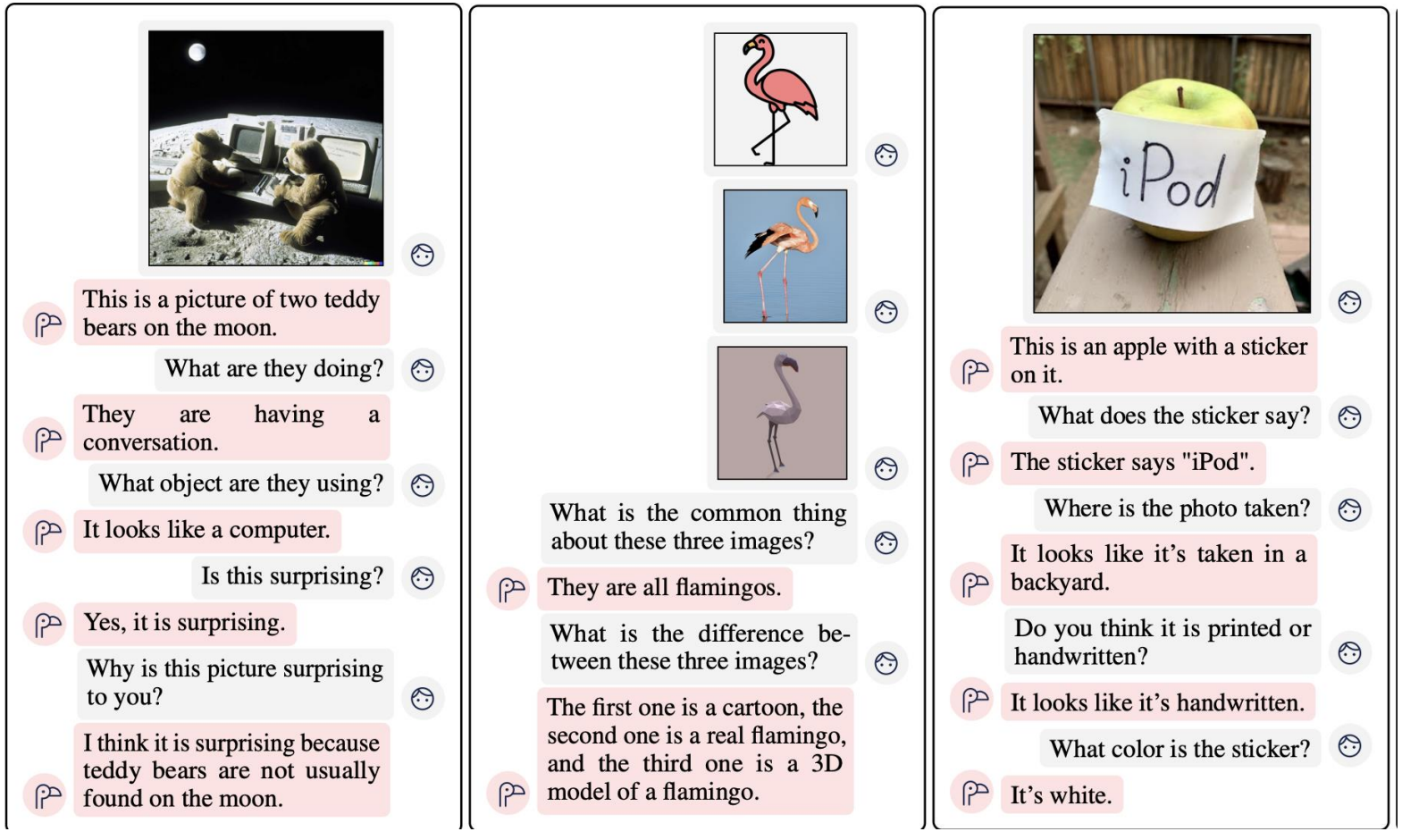


[Tsimpoukelli et al., Multimodal Few-Shot Learning with Frozen Language Models. NeurIPS 2021]

multisensory intelligence

Adapting Large Language Models

Flamingo



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

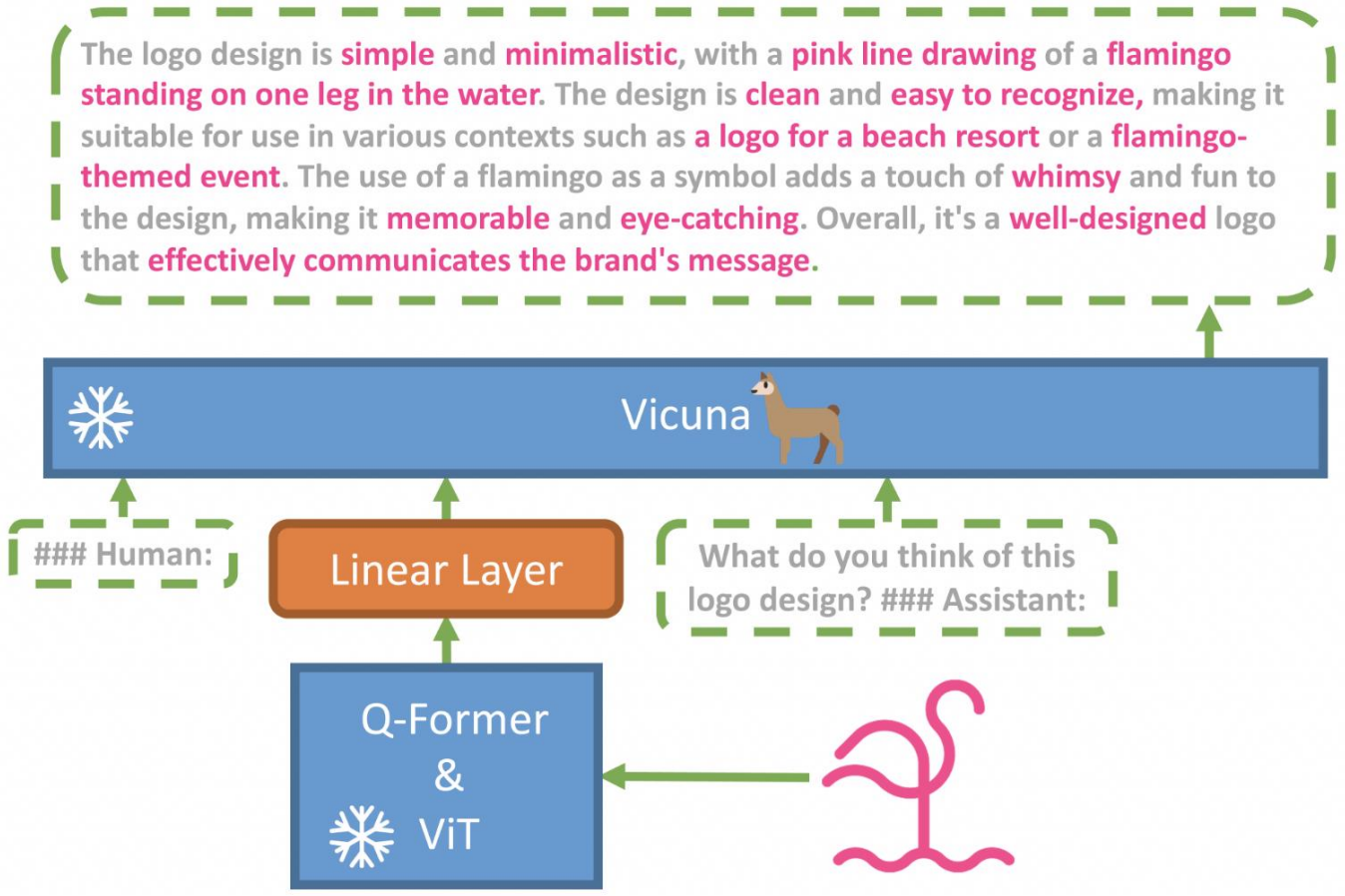
multisensory intelligence

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.



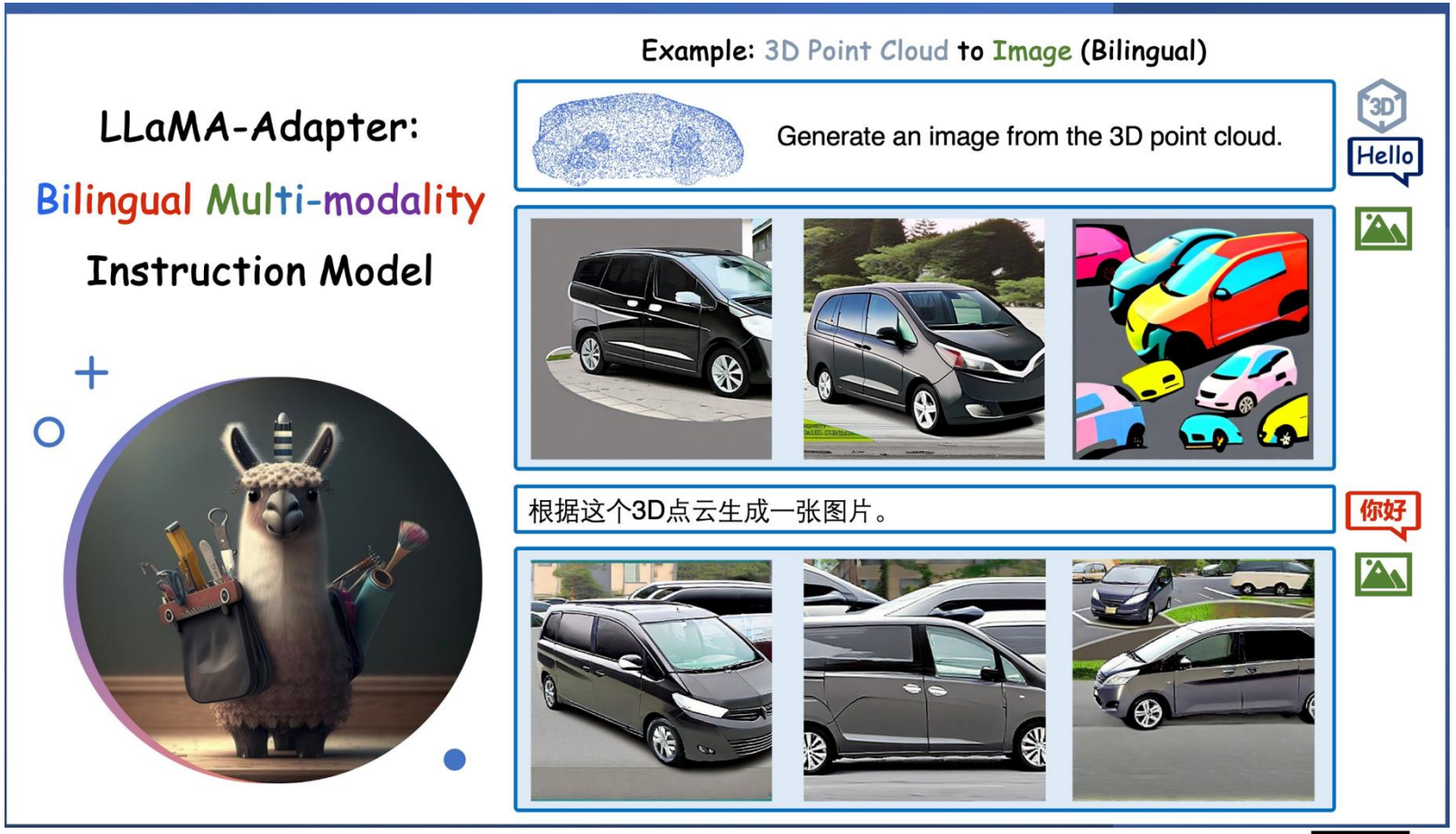
The architecture of MiniGPT-4.

[Zhu et al., MiniGPT-4: Enhancing Vision-language Understanding with Advanced Large Language Models. 2023]

multisensory intelligence

Scaling Large Multimodal Models

LLaMA-Adapter

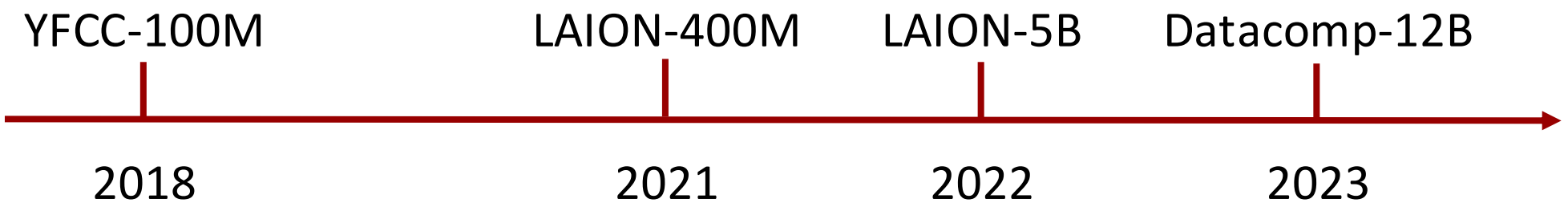


[Zhang et al., LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention. ICLR 2024]

multisensory intelligence

Multimodal 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	Obelics [Laurençon et al., 2024a]	141M Docs	45%
Text	DCLM [Li et al., 2024b]	6.6T Toks	10%

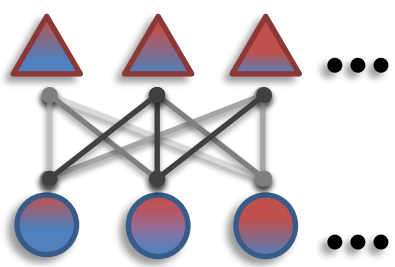
Multimodal Instruction Tuning Datasets

- 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	✗	~ 235K	5	✗
InstructBLIP	28	✗	~ 1.6M	9.7	✗
M^3IT (Ours)	40	✓	2.4M	10	✓

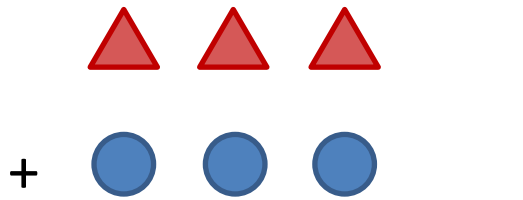
From Text to Multimodal Generation

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



Multimodal representation

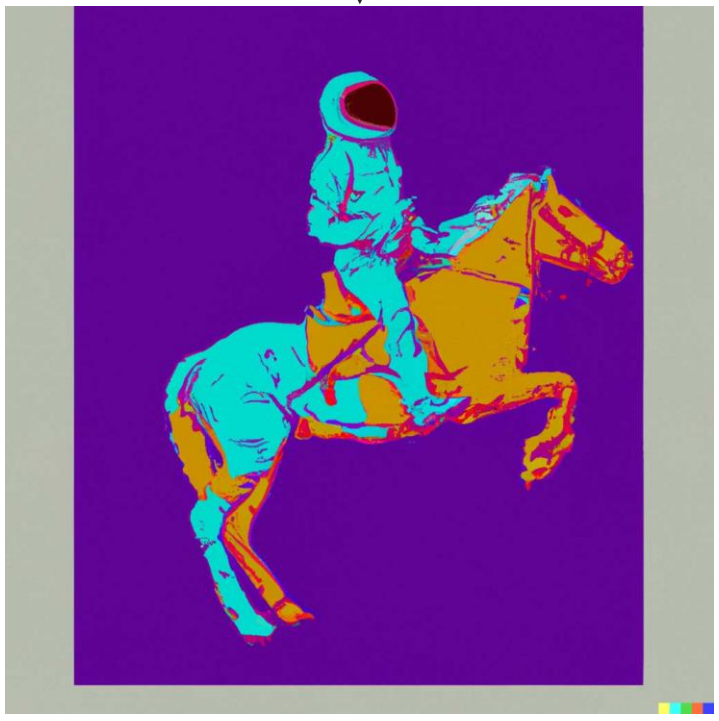
This person is being sarcastic.
They seem to be close friends.



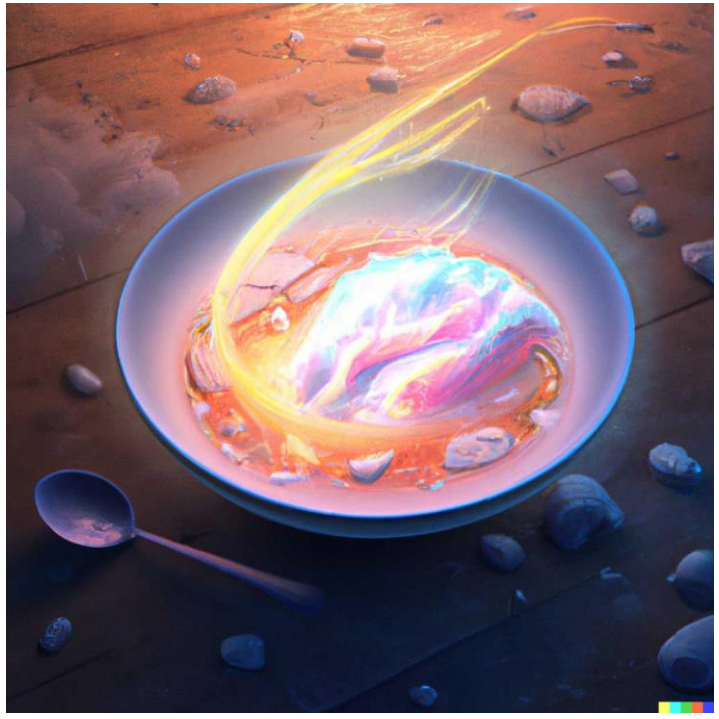
(retrieve next episode)
(highlight multimodal evidence)

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



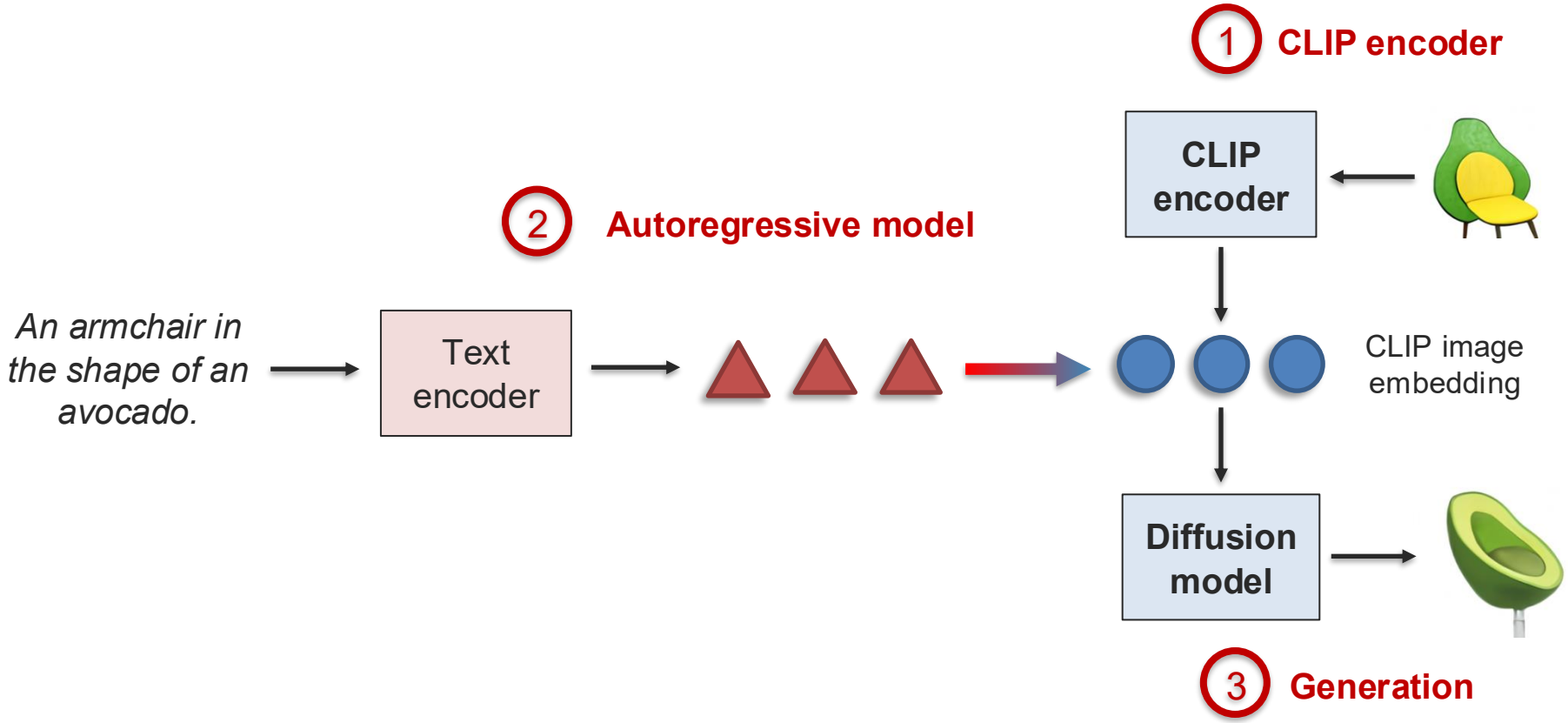
[Ramesh et al., Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv 2022]

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.

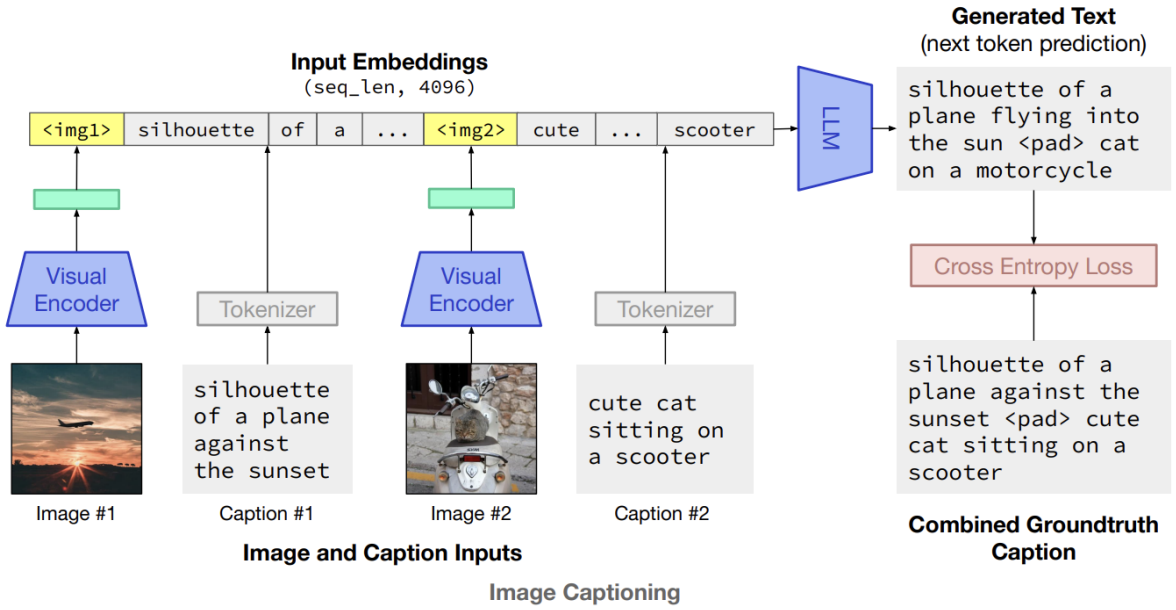


[Ramesh et al., Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv 2022]

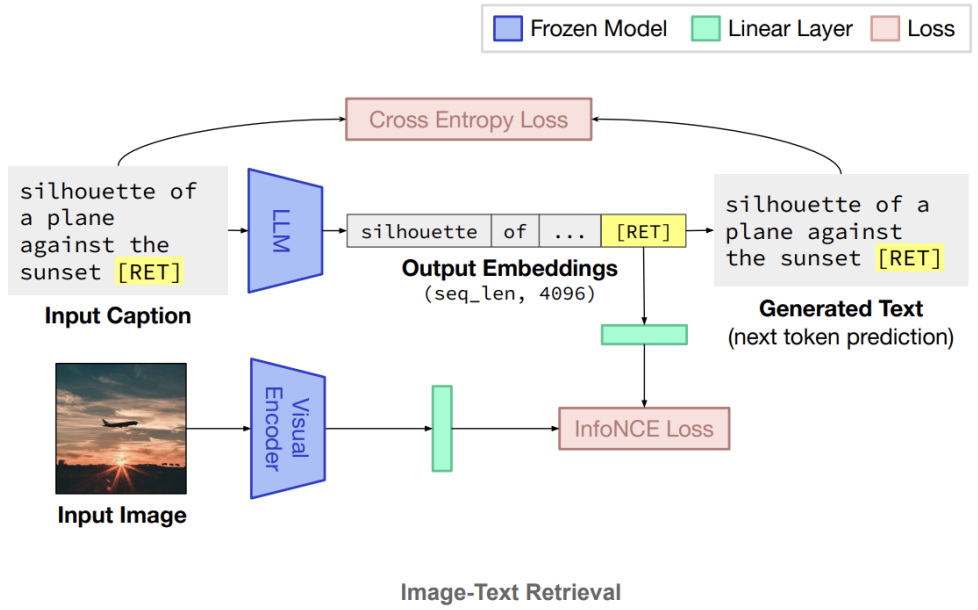
Grounding LMs for Image Retrieval

LIMBeR + CLIP. Trainable in 1 day on 1 GPU

Interleaved images and text



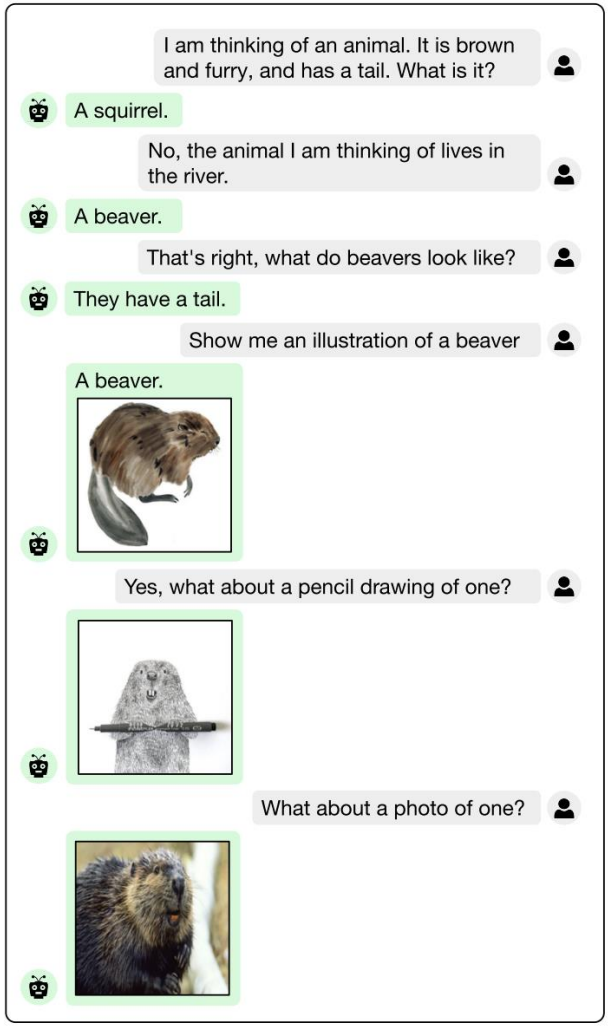
CLIP, with a frozen LLM

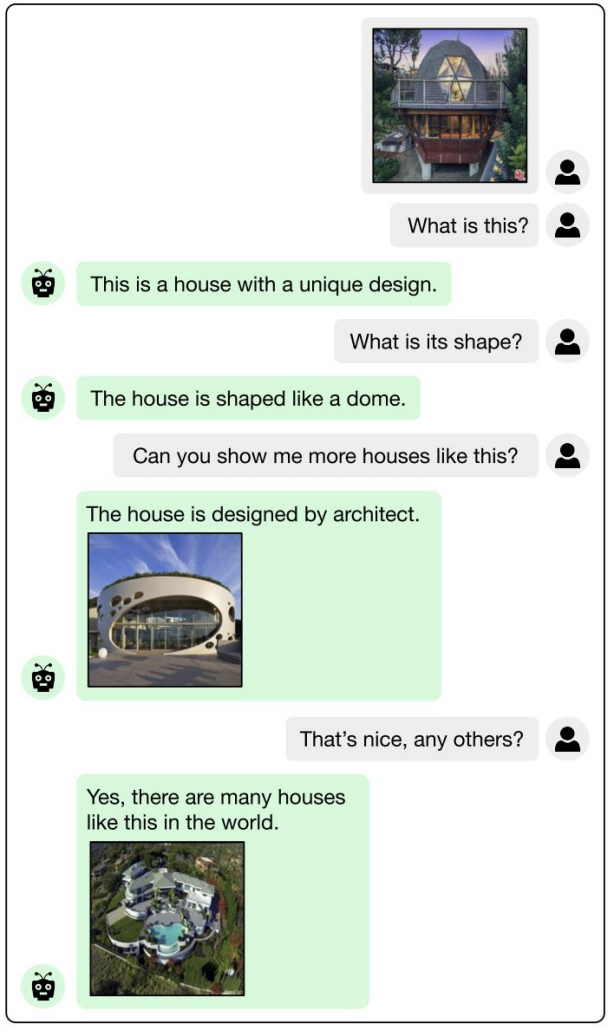


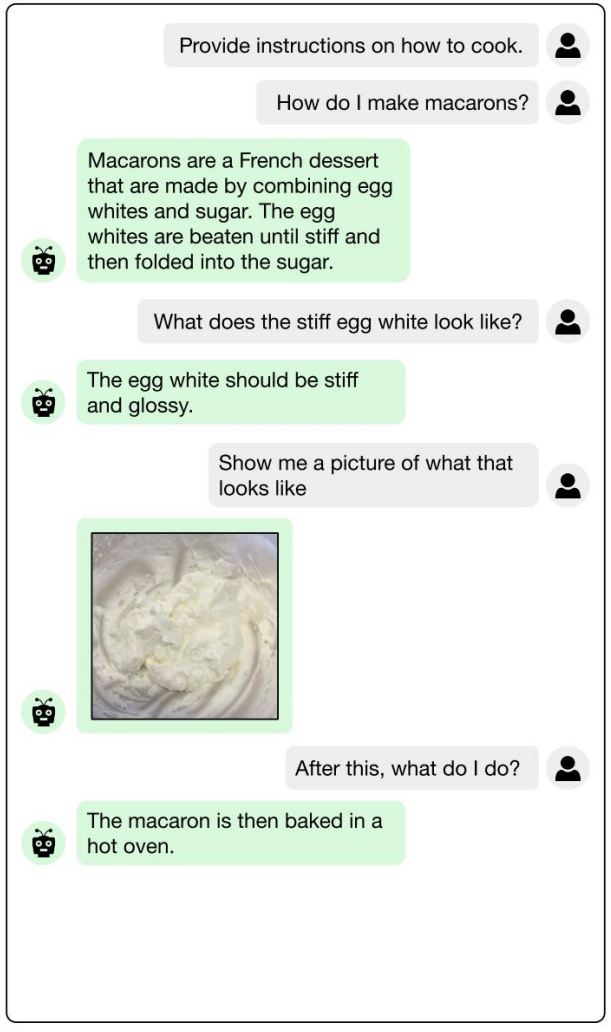
[Koh et al., Grounding Language Models to Images for Multimodal Inputs and Outputs. ICML 2023]

multisensory intelligence

Grounding LMs for Image Retrieval







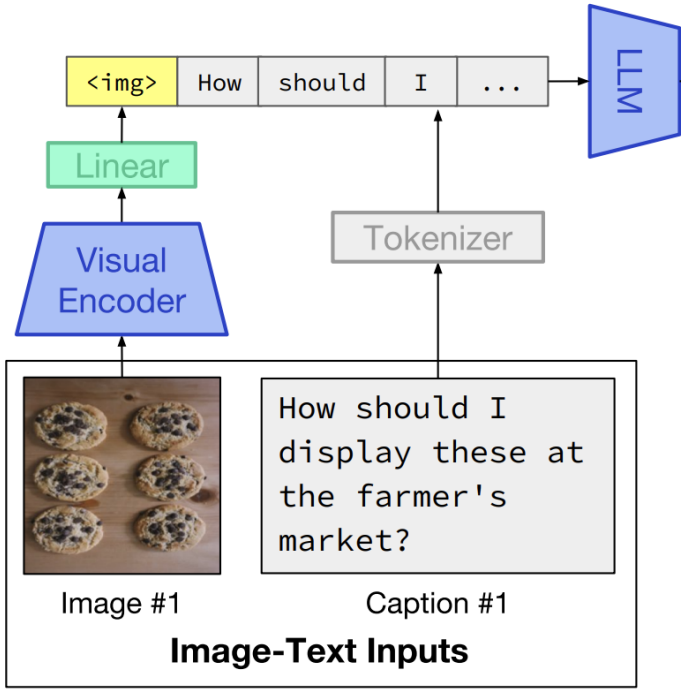
[Koh et al., Grounding Language Models to Images for Multimodal Inputs and Outputs. ICML 2023]

multisensory intelligence

Grounding LMs for Multimodal Generation

Large multimodal models with image generation

Adapter layer (part 2)



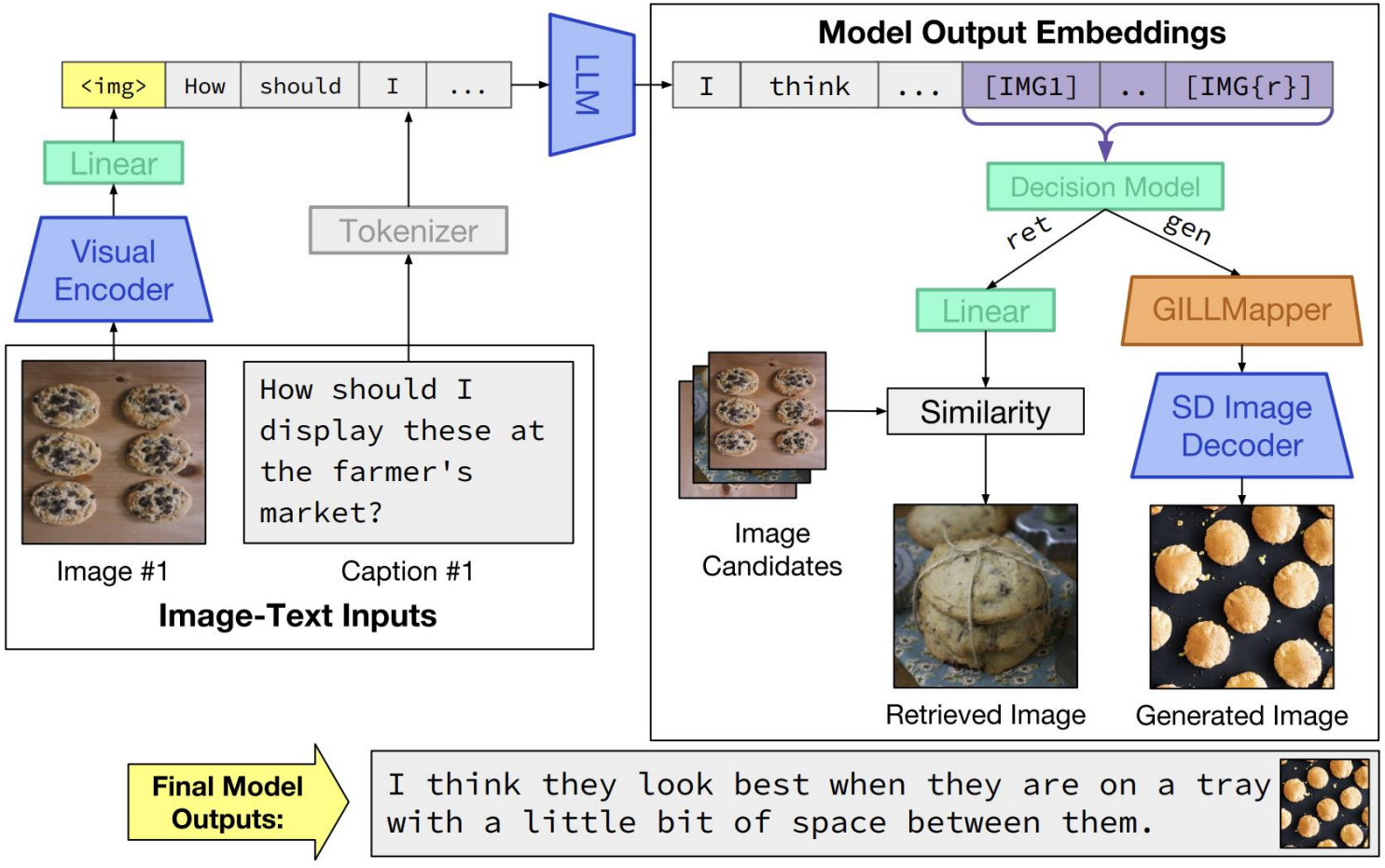
[Koh et al., Generating Images with Multimodal Language Models. NeurIPS 2023]

multisensory intelligence

Grounding LMs for Multimodal Generation

Large multimodal models with image generation

Adapter layer (part 2)



Generation model (part 3)

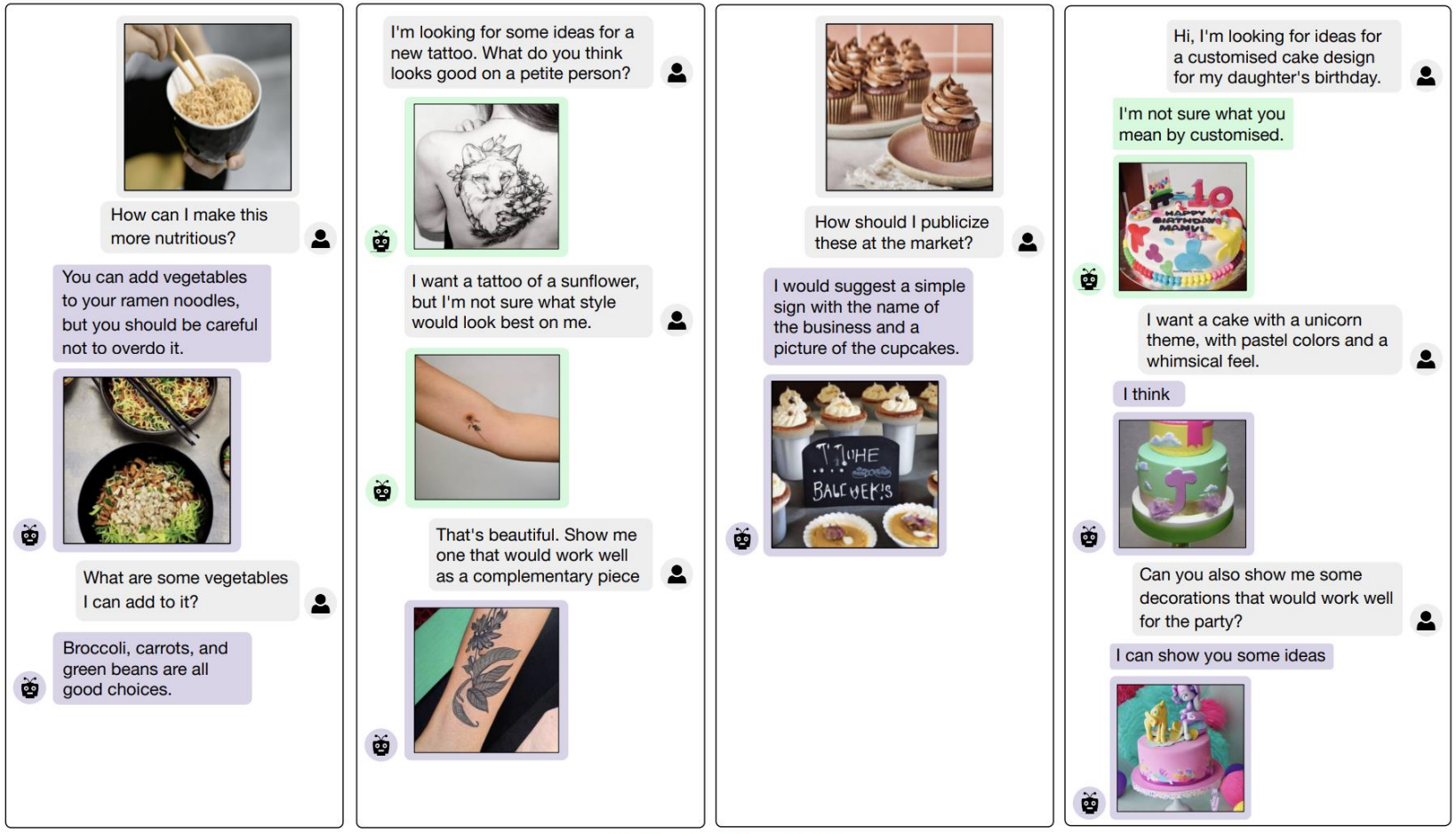
[Koh et al., Generating Images with Multimodal Language Models. NeurIPS 2023]

multisensory intelligence

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.



Multimodal Dialogue

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

User prompts Retrieved Generated

multisensory intelligence

Summary

1. Multimodal transformers & foundation models
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

HW2 due tomorrow Wed (3/4).

HW2 presentation this Thurs + in-depth tutorial on implementing multimodal LLMs.

HW3 (multimodal LLMs) out later this week.