Learning to Understand Our Multimodal World with Minimal Supervision

Yong Jae Lee

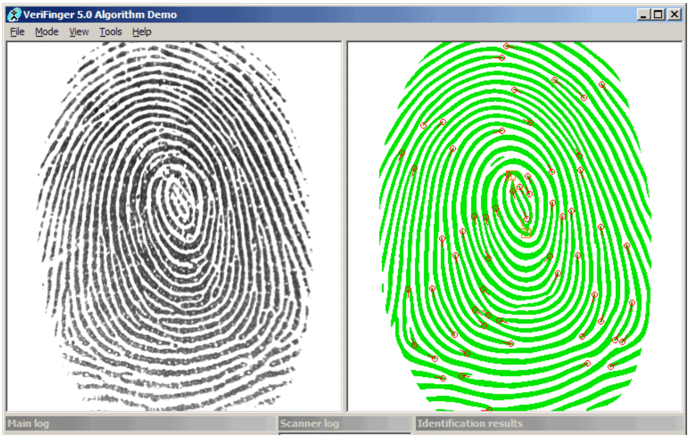
University of Wisconsin-Madison / GivernyAI

Image of LLaVA generated by GLIGEN
"a cute lava llama with glasses" + box prompt

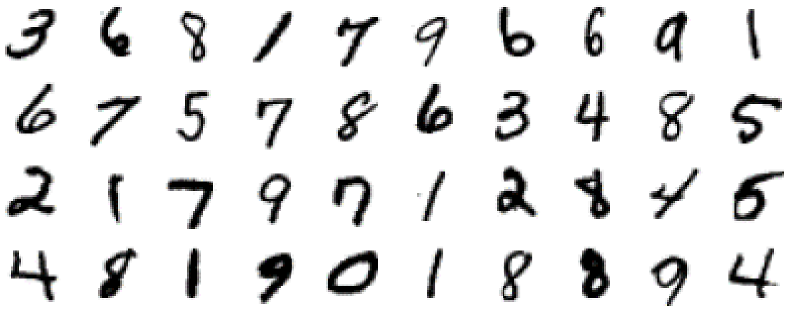
2024 embedded VISION SUMMIT

Once Upon a Time...
When I was a Graduate Student (2006-2012)

Frontal face detection



Fingerprint recognition



Recognizing license plates, zip codes, checks

Very few computer vision systems worked

Computer Vision in the Deep Learning Era (2012 - Present)

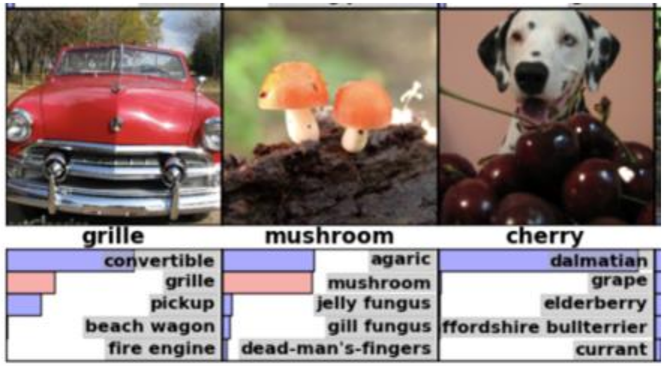
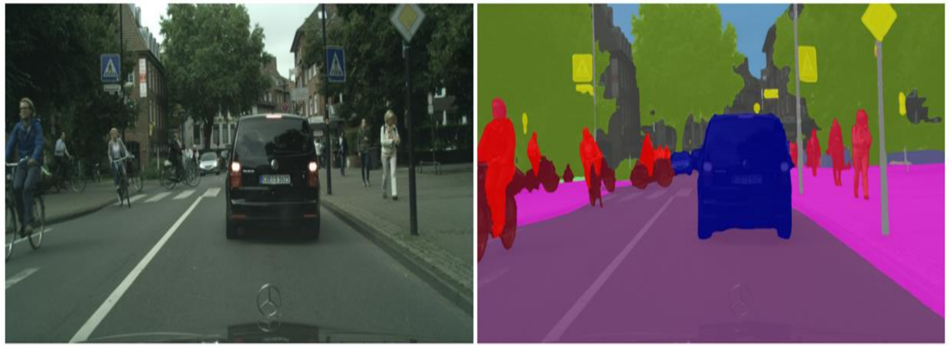
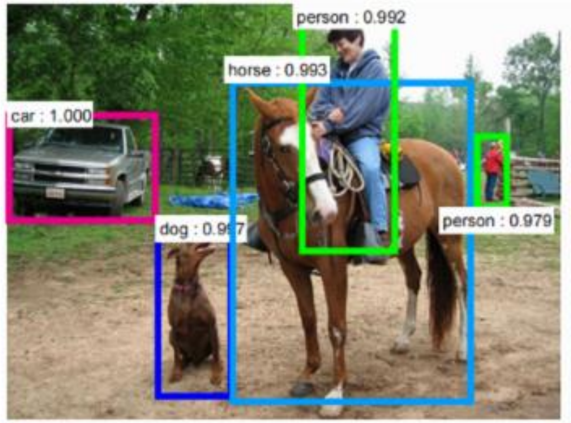


Image classification



Semantic segmentation



Object detection

Source: https://www.youtube.com/watch?v=YGO2lwAgrig

Pose recognition

Input Image

Predicted 3D Shape

Predicted 3D Parts

3D prediction

Surface normal prediction

… and many more

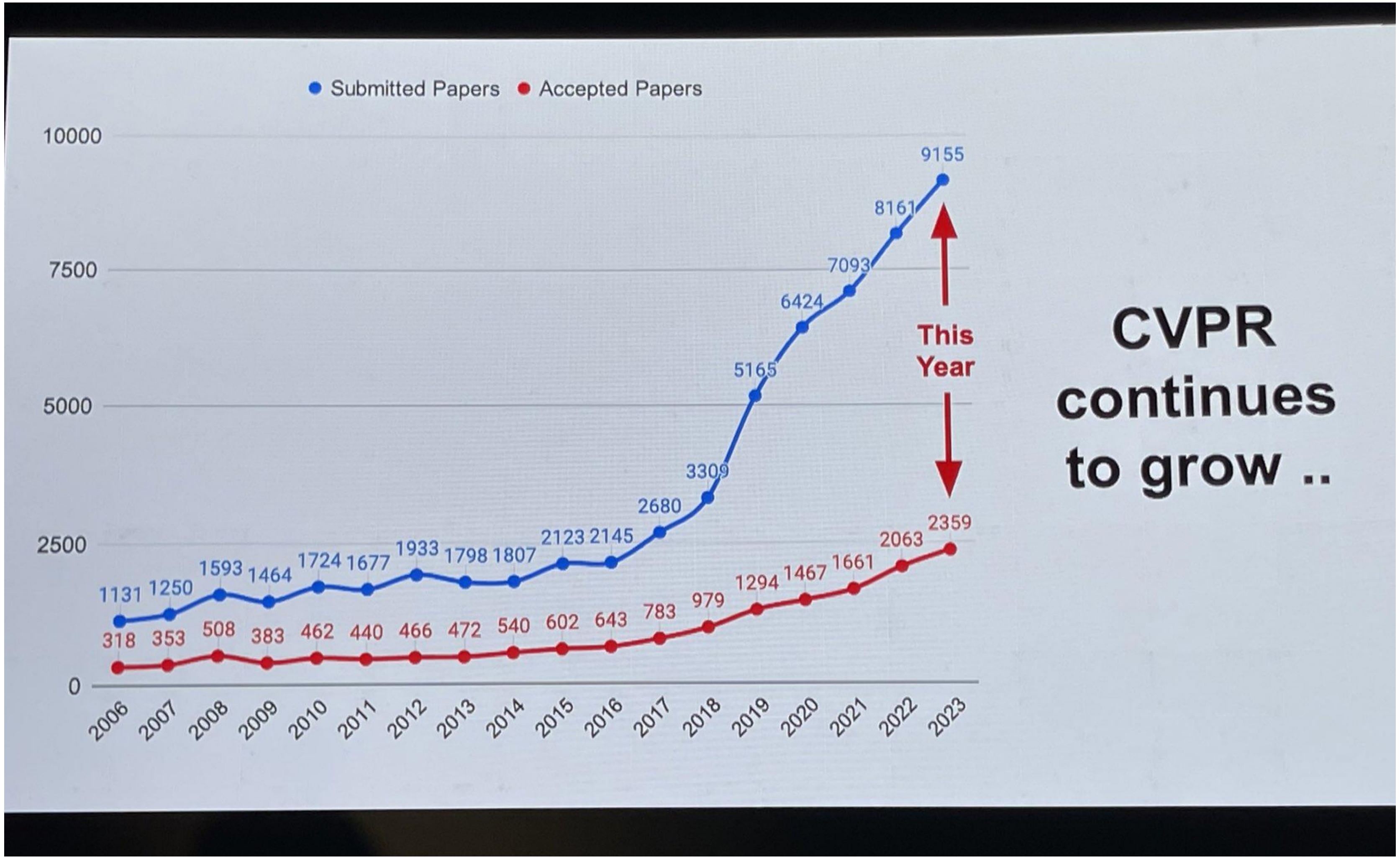
Explosion in …

Students!

Startups!

Funding!

Hiring!

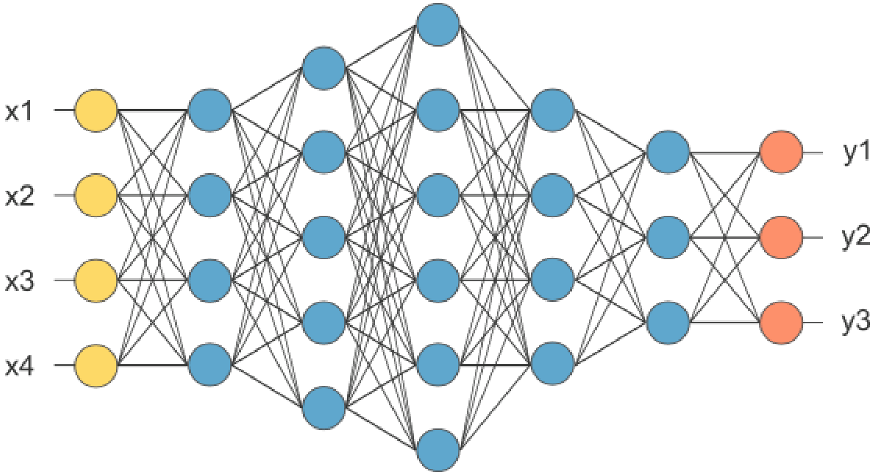


CVPR continues to grow ..

Ingredients for Success Today

1. Big compute (GPUs)

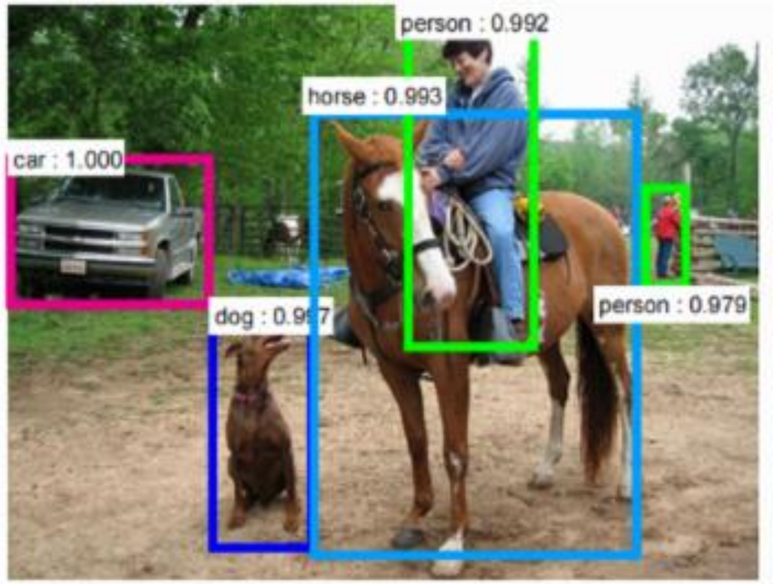
2. Big models (deep neural nets)



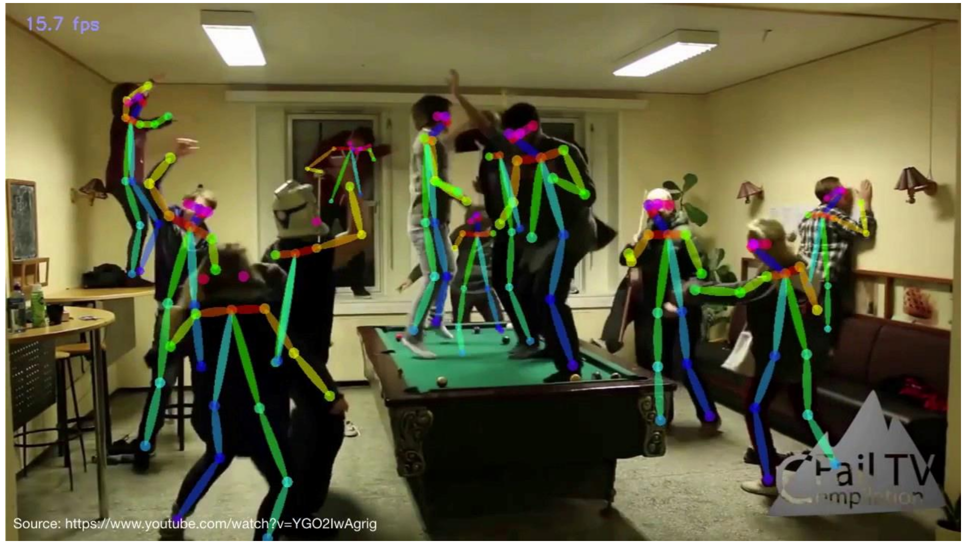
3. Big data

However, Prevailing Paradigm Thus Far:

“Specialist” models: **single-model, single-task**

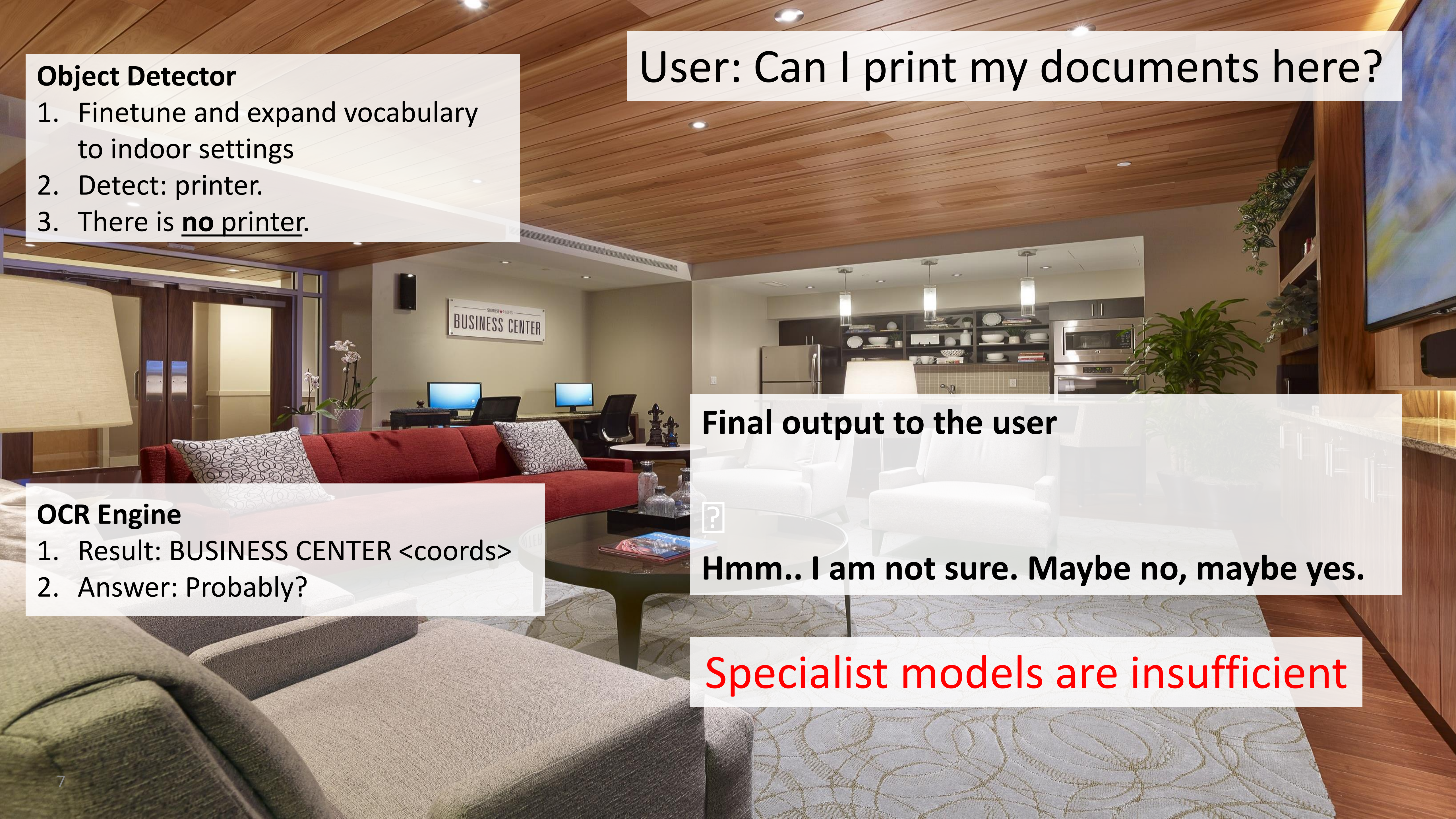


Object Detection *Only*



Source: https://www.youtube.com/watch?v=YGO2IwAgrig

Pose Recognition *Only*



User: Can I print my documents here?

Object Detector

1. Finetune and expand vocabulary to indoor settings
2. Detect: printer.
3. There is **no** printer.

OCR Engine

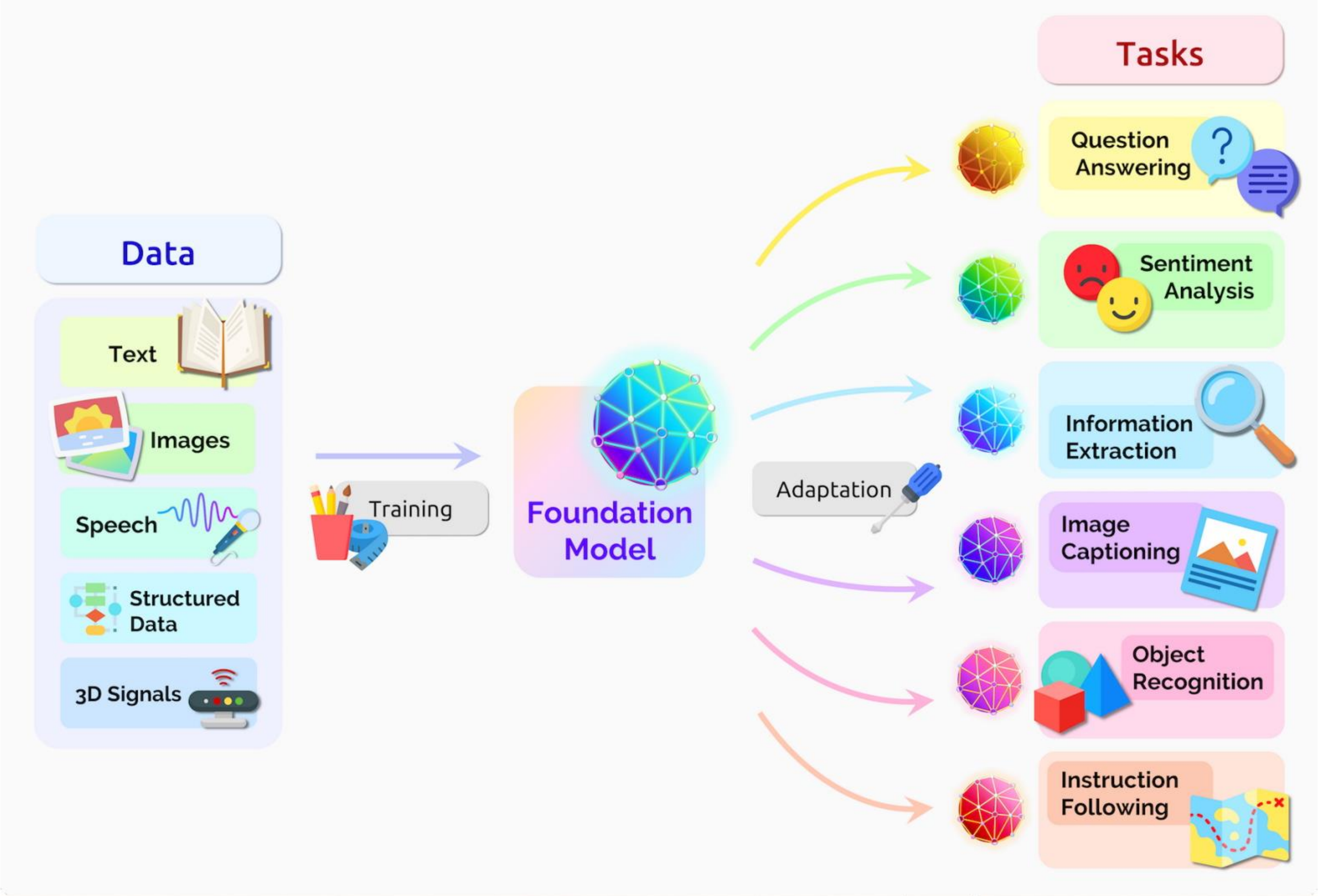
1. Result: BUSINESS CENTER <coords>
2. Answer: Probably?

Final output to the user

Hmm.. I am not sure. Maybe no, maybe yes.

Specialist models are insufficient

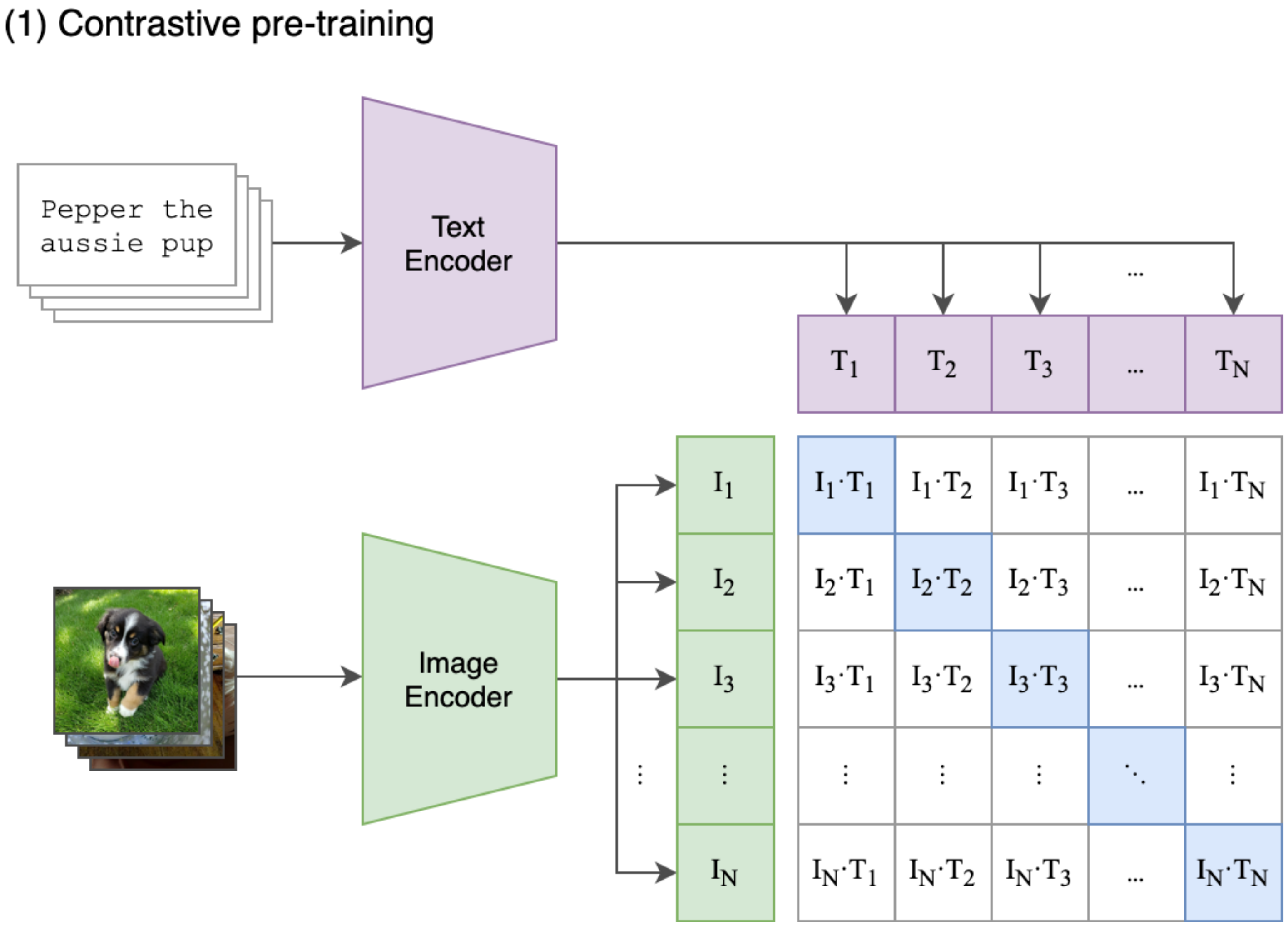
Rise of “Generalist” Foundation Models (2020s)



- **Single-model, many tasks**
- Large Language Models (e.g., GPT4)
- Vision Transformers
- Image-Text Models (e.g., CLIP)

Image credit: https://blogs.nvidia.com/blog/what-are-foundation-models/

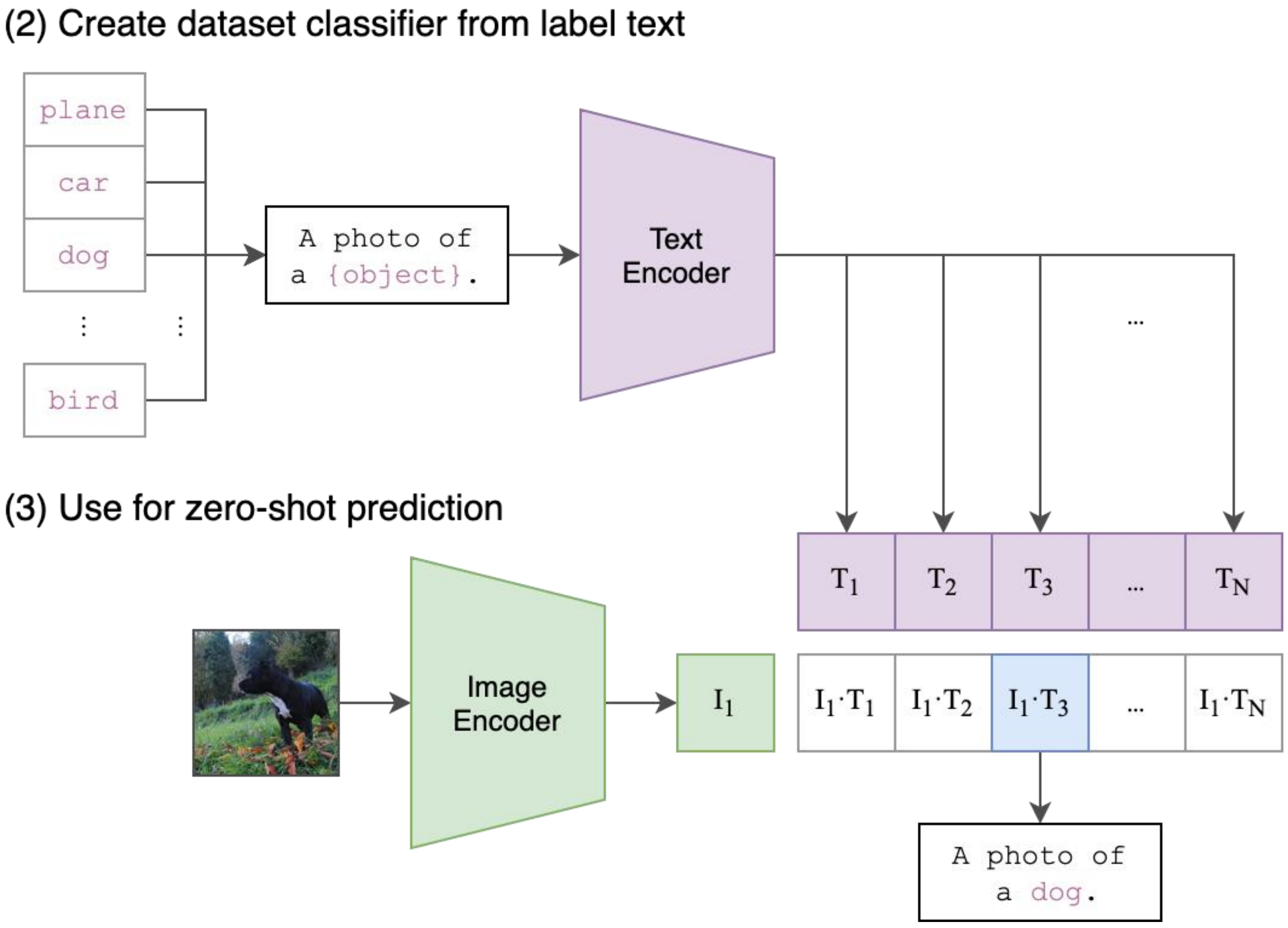
Rise of "Generalist" Foundation Models (2020s)



- Contrastive Language-Image Pretraining (CLIP)
- Trained using 400M image-text pairs
- Zero-shot recognition

"Learning Transferable Visual Models From Natural Language Supervision" Alec Radford et al. 2021

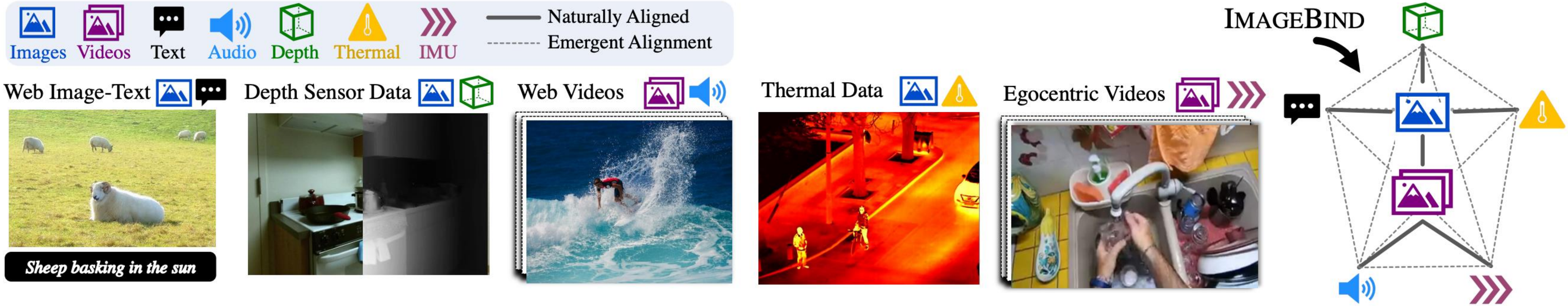
Rise of "Generalist" Foundation Models (2020s)



- Contrastive Language-Image Pretraining (CLIP)
- Trained using 400M image-text pairs
- Zero-shot recognition

"Learning Transferable Visual Models From Natural Language Supervision" Alec Radford et al. 2021

Rise of “Generalist” Foundation Models (2020s)



- ImageBind aligns multiple modalities
- Emergent alignment

“ImageBind: One Embedding Space To Bind Them All” Rohit Girdhar et al. 2023

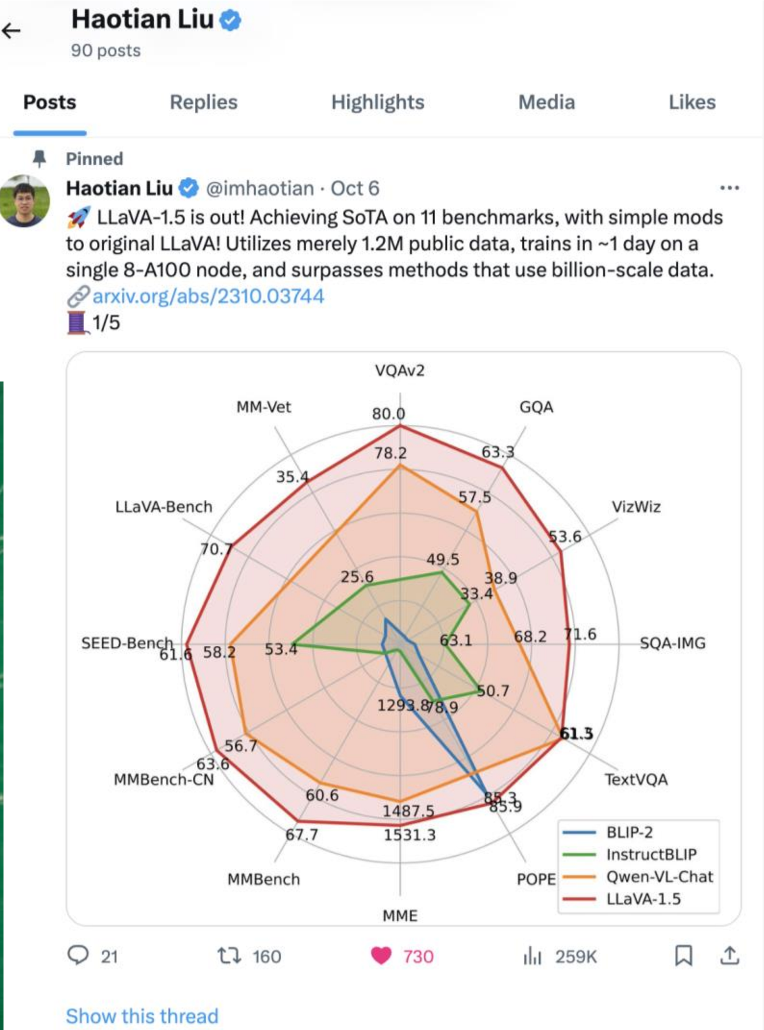
Today's talk:
Large Multimodal Generalist Models

- *Generalist* vision-language models that understand visual data and communicate in natural language

- *Controllable* ("aligned") models that produce desirable outputs for wide-concept knowledge

- Challenge: How to effectively train such models with minimal supervision?
- Solution: Adapt pre-trained foundation models, and design semi-automatic methods for data collection

Humans See and Reason about the Visual World; Express and Interact with Natural Language



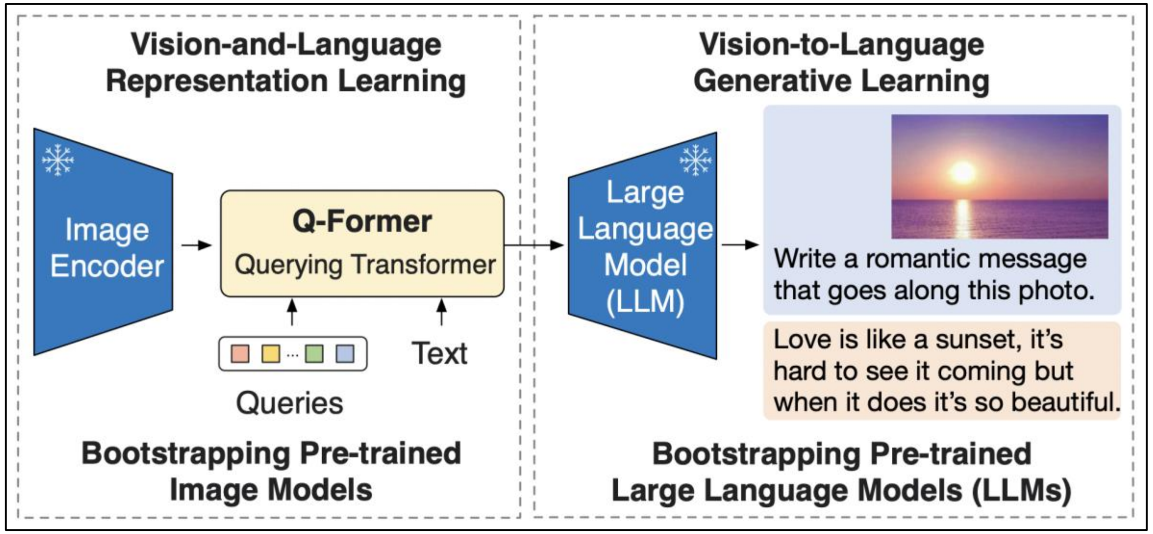
Haotian Liu
90 posts

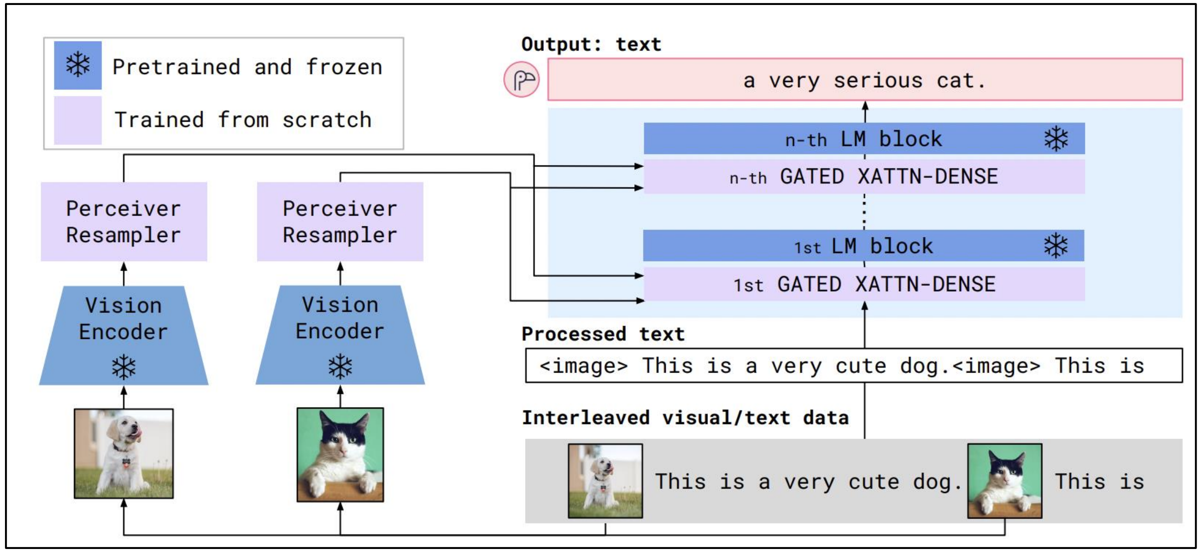
Haotian Liu @imhaotian · Oct 6

🚀 LLaVA-1.5 is out! Achieving SoTA on 11 benchmarks, with simple mods to original LLaVA! Utilizes merely 1.2M public data, trains in ~1 day on a single 8-A100 node, and surpasses methods that use billion-scale data.
🔗arxiv.org/abs/2310.03744
🧵1/5

Show this thread

How to Build Generalist Multimodal Models?


Vision-and-Language Representation Learning
Image Encoder
Q-Former
Querying Transformer
Queries
Text
Bootstrapping Pre-trained Image Models
Vision-to-Language Generative Learning
Large Language Model (LLM)
Write a romantic message that goes along this photo.
Love is like a sunset, it's hard to see it coming but when it does it's so beautiful.
Bootstrapping Pre-trained Large Language Models (LLMs)



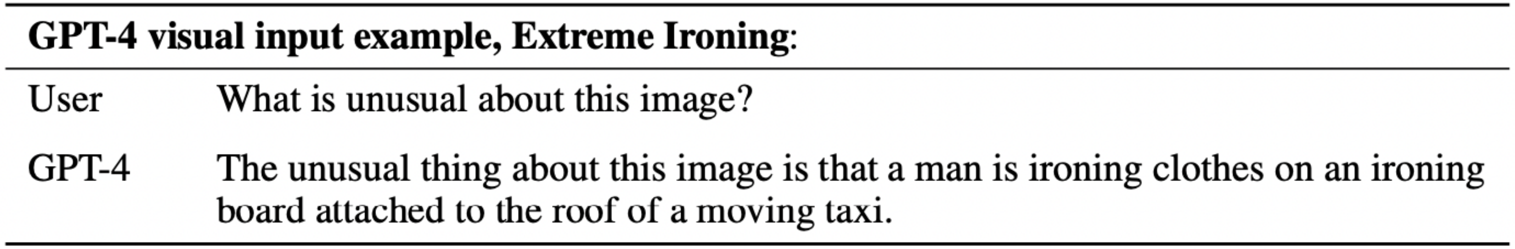
- Prior methods generally lack instruction following capabilities

[J. Wang et al. 2022. GIT: A Generative Image-to-text Transformer for Vision and Language]
[J. Li et al. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models]
[J.-B. Alayrac. 2022. Flamingo: a visual language model for few-shot learning]

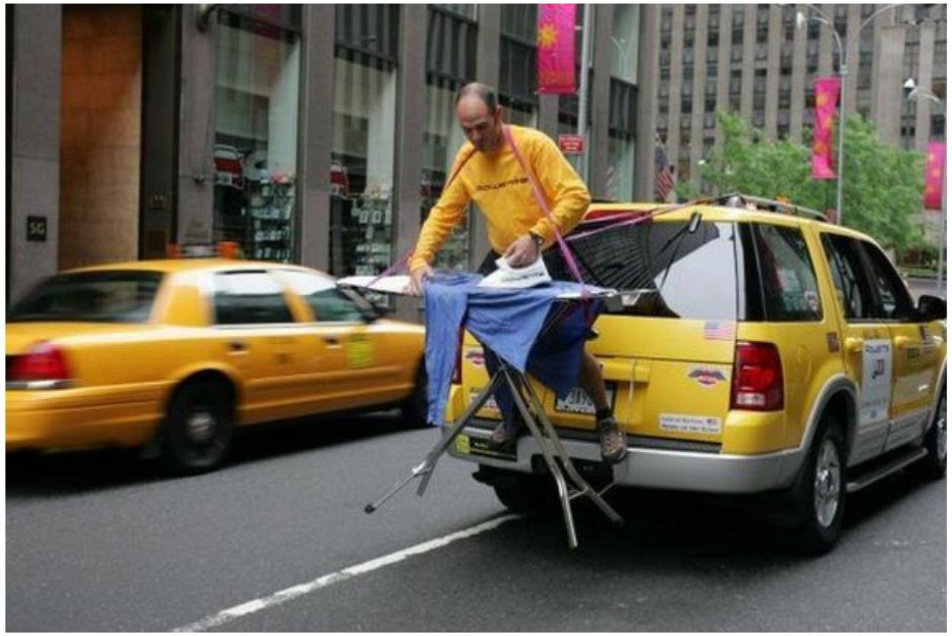
…

How to Build Generalist Multimodal Models?

- GPT-4V: Strong language and visual reasoning, but closed-source



GPT-4 visual input example, Extreme Ironing:	
User	What is unusual about this image?
GPT-4	The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.



Source: `https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg`

[GPT-4 Technical Report, OpenAI. March 2023.]

How to Build Generalist Multimodal Models?

Output

The unusual aspect of this image is …

Language Decoder

Cross-modal Connector

Visual Encoder

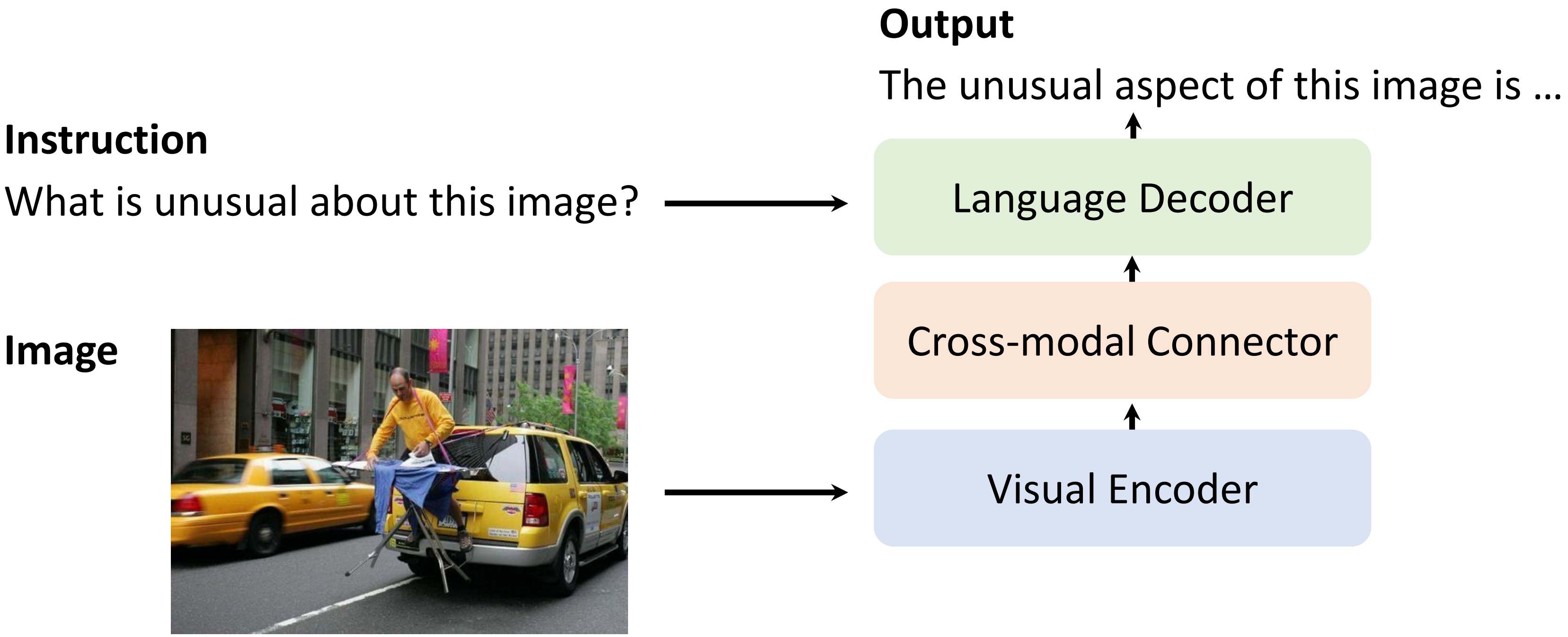
Instruction

What is unusual about this image?

Image

1. <u>Data</u> to tune the model for *instruction-following* capabilities
2. Strong <u>pretrained</u> vision and language models
3. <u>Connecting</u> vision and language

How to Build Generalist Multimodal Models?



Output

The unusual aspect of this image is …

Instruction

What is unusual about this image?

Language Decoder

Cross-modal Connector

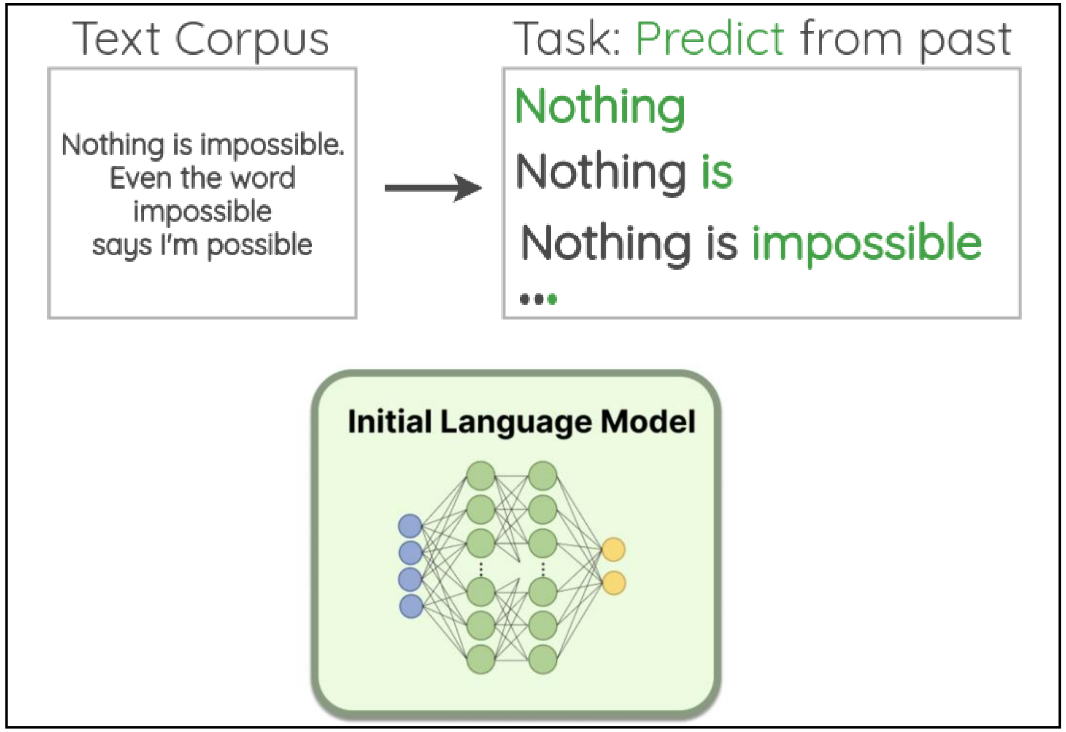
Image

Visual Encoder

1. <u>Data</u> to tune the model for *instruction-following* capabilities
2. Strong pretrained vision and language models
3. Connecting vision and language

Lessons from Training Large Language Models

- Step 1: Representation learning
  - Self-supervised pre-training



Lessons from Training Large Language Models

Prompt: *Explain the moon landing to a 6 year old in a few sentences.*

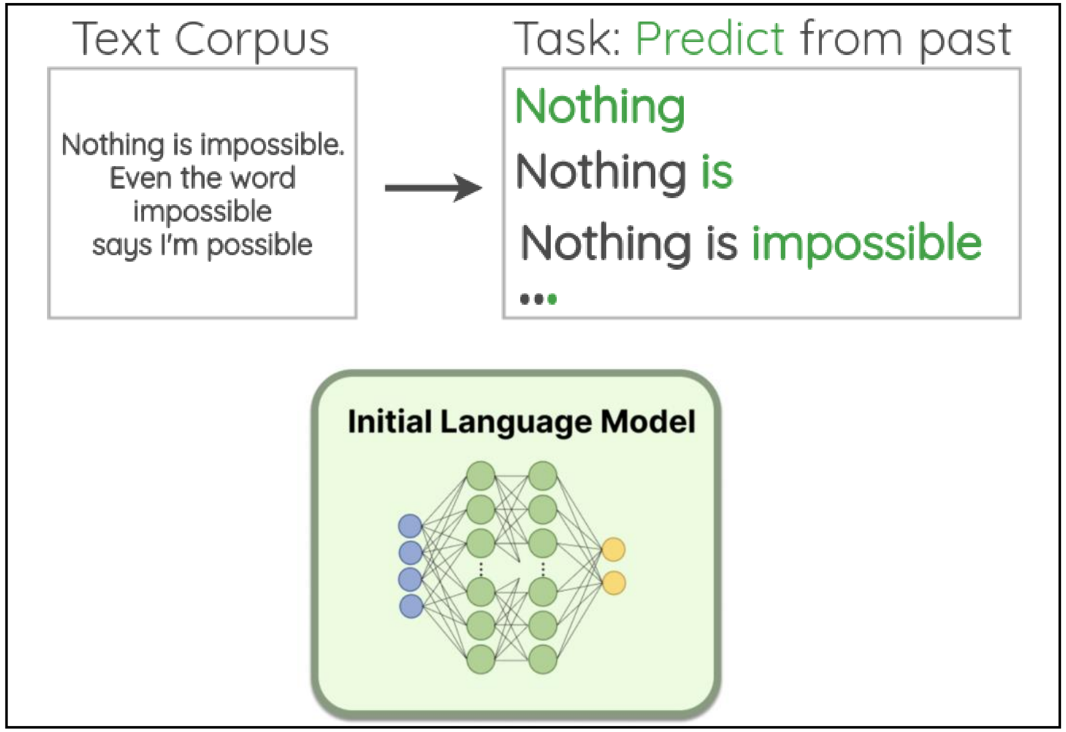
Completion: *Explain the theory of gravity to a 6 year old.*
Explain the big bang theory to a 6 year old.
Explain evolution to a 6 year old.

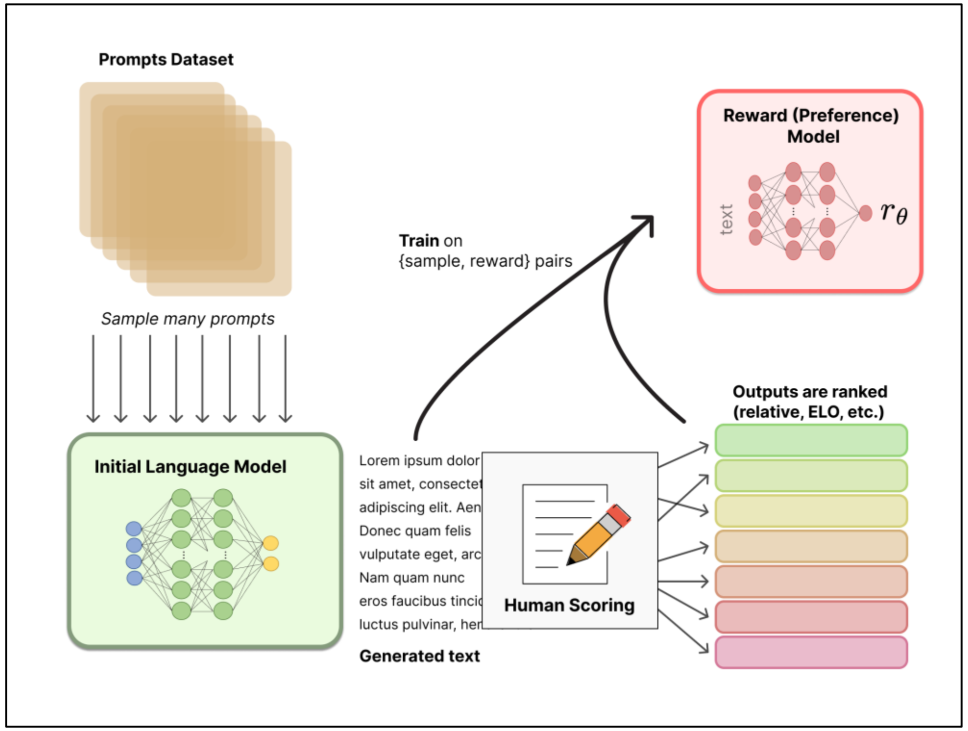
- Pre-training on bulk text does not naturally produce an LLM that understands user intent!
- *How do we get ChatGPT-like models?*

[Chris Manning, Simons Workshop on Large Language Models and Transformers, 2023.]

Lessons from Training Large Language Models

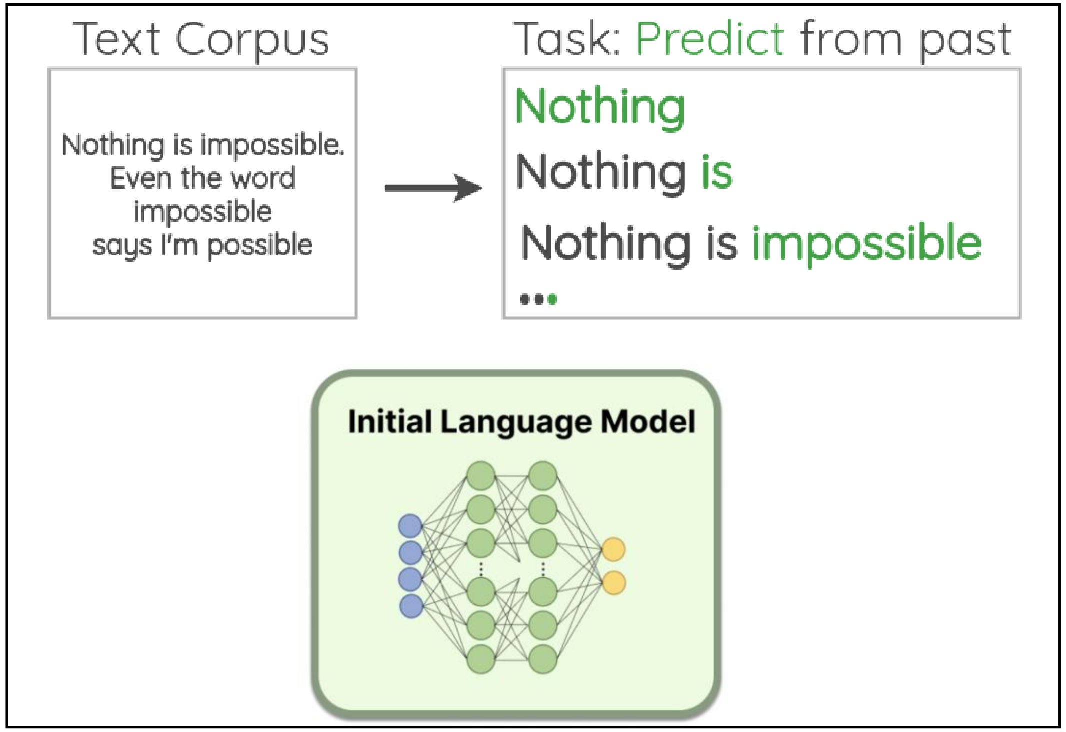
- Step 1: Representation learning
  - Self-supervised pre-training
- Step 2: Instruction Tuning
  - Supervised fine-tuning / RLHF



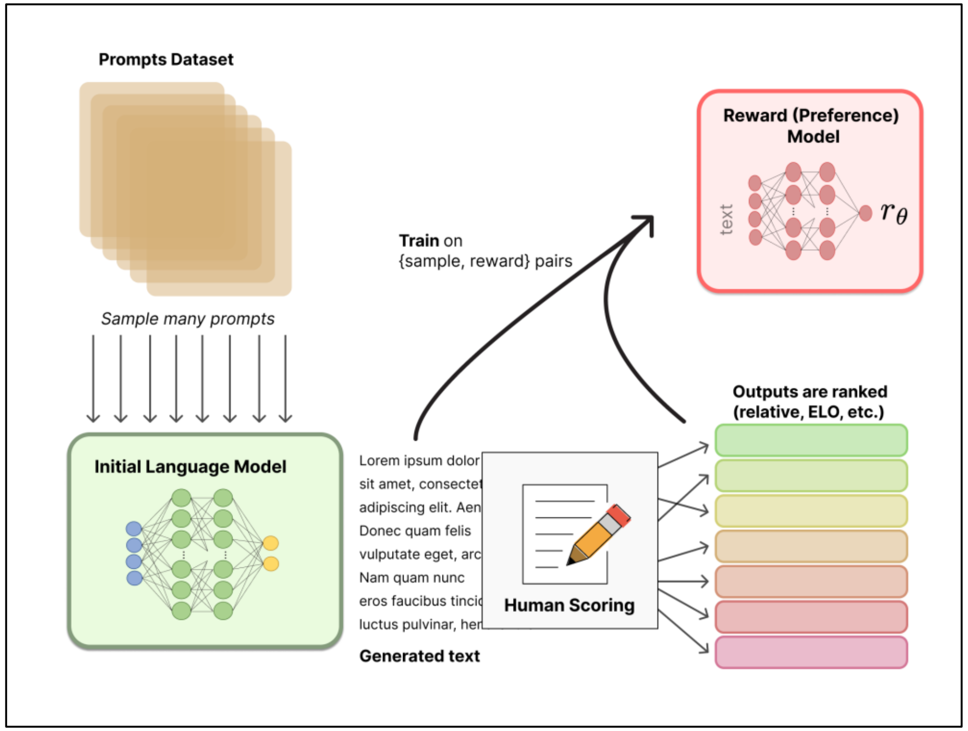

Prompts Dataset
Sample many prompts
Initial Language Model
Lorem ipsum dolor sit amet, consectet adipiscing elit. Aen Donec quam felis vulputate eget, arc Nam quam nunc eros faucibus tincid luctus pulvinar, her
Generated text
Human Scoring
Train on {sample, reward} pairs
Reward (Preference) Model
text
r_θ
Outputs are ranked (relative, ELO, etc.)

Lessons from Training Large Language Models

- Step 1: Representation learning
  - Self-supervised pre-training



- Step 2: Instruction Tuning
  - Supervised fine-tuning / RLHF


Prompts Dataset
Sample many prompts
Initial Language Model
Lorem ipsum dolor sit amet, consectet adipiscing elit. Aen Donec quam felis vulputate eget, arc Nam quam nunc eros faucibus tincid luctus pulvinar, her
Generated text
Human Scoring
Train on {sample, reward} pairs
Reward (Preference) Model
text
r_θ
Outputs are ranked (relative, ELO, etc.)

Lessons from Instruction Tuning in LLMs

Instruction	**Output**
Explain human's behavior. Behavior: cry.	There could be many reasons why a person might cry. They could be feeling sad, scared, angry, or frustrated…
Recommend a movie for me.	Certainly! Without specific preferences in mind, I'll recommend a popular and highly-rated movie: "The Shawshank Redemption." It's a classic drama film …
…	**…**

Lessons from Instruction Tuning in LLMs

- How to collect instruction tuning data?
  - Human: high-quality, hand-written by humans → high cost
  - Machine: strong LLM-based teacher like ChatGPT → *affordable* cost

Manual seed instruction-output pairs

Explain human's behavior.	<reference answer>
Recommend a movie for me.	<reference answer>

Please generate new instruction-output pairs that meet the following requirements: …

<new instruction 1>	<new output 1>

…

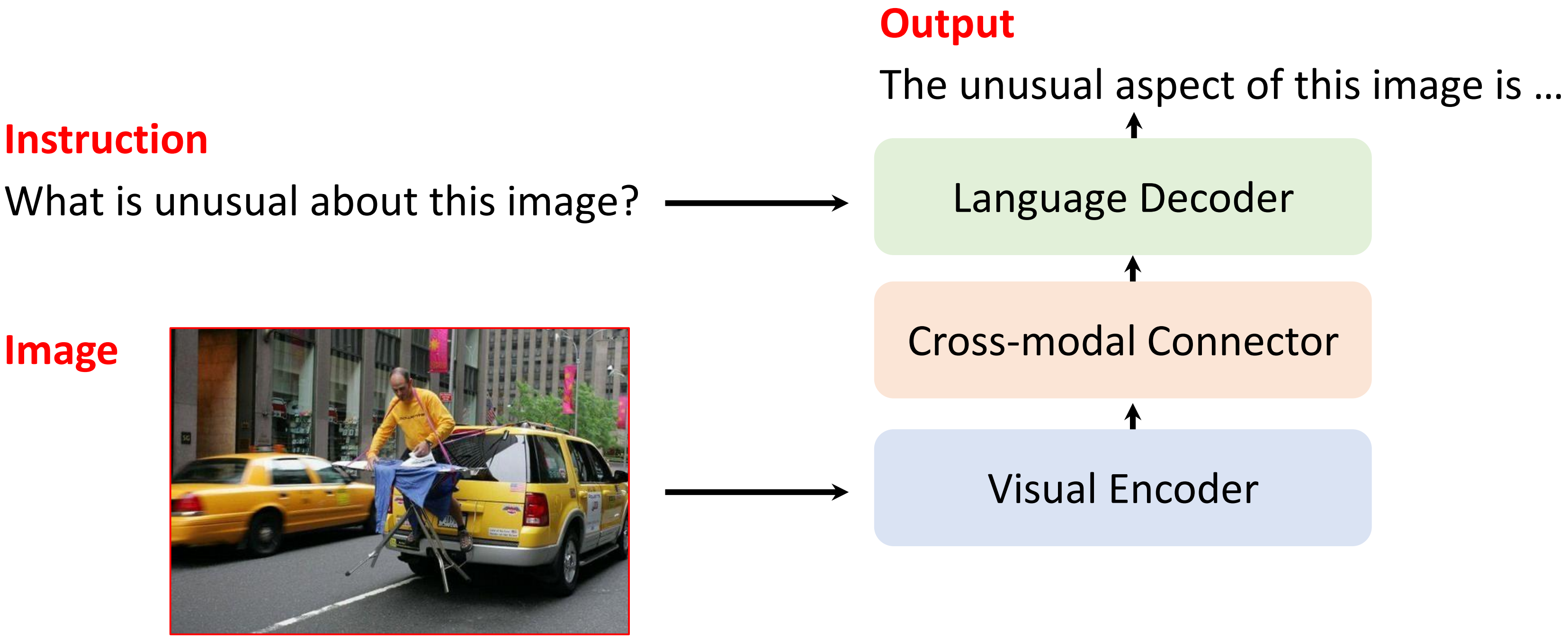
[Yizhong Wang. 2022. Self-instruct: Aligning language model with self generated instructions.]

Lessons from Instruction Tuning in LLMs

	LLaMA 	Alpaca 	Vicuna 	GPT4-Alpaca 	…	Tulu
Data Source		GPT-3.5	ShareGPT (Human & GPT)	GPT-4 (text-only)	…	Mixed Data
Instruction-following Data (#Turns)	None	52K	500K (~150K conversions)	52K	…	

All are text-only!

Instruction Tuning in Large Multimodal Models



Output

The unusual aspect of this image is …

Instruction

What is unusual about this image?

Image

- How to obtain Image-Instruction-Output triplet training data?
- Can we use a large language model for this?

Text-only GPT Assisted Visual Instruction Data Creation

Image

- *How do we get a text-only LLM to "see" an image?*

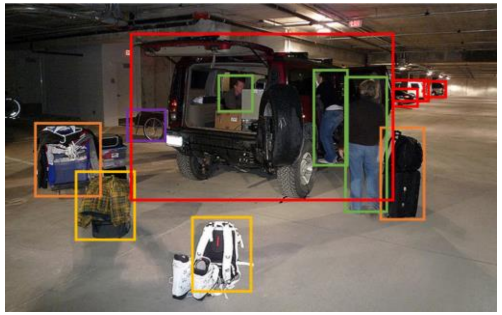
→ Let it read context information written in text

Context (caption)

A group of people standing outside of a black vehicle with various luggage.

MS-COCO
Each image associated with:
- 5 captions
- Object categories / bounding boxes

Context (layout)

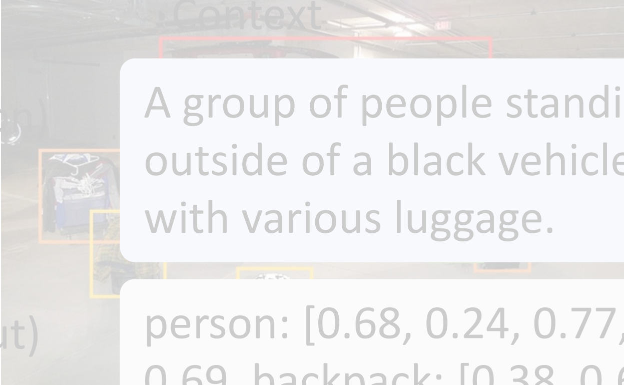


→

person: [0.68, 0.24, 0.77, 0.69], person: [0.63, 0.22, 0.68, 0.51], person: [0.44, 0.23, 0.48, 0.34], backpack: [0.38, 0.69, 0.48, 0.91], ….

[T.-Y. Lin et al. 2014. Microsoft coco: Common objects in context.]

Text-only GPT Assisted Visual Instruction Data Creation



Context

Context (caption)

A group of people standing outside of a black vehicle with various luggage.

Context (layout)

person: [0.68, 0.24, 0.77, 0.69, backpack: [0.38, 0.69, 0.48, 0.91] …

Instruction

What are the challenges these people might be facing?

Output

They may be having difficulty fitting all luggage into the back of the SUV. There are many bags, suitcases already in the back, while more…

Manual seed example(s) of context-instruction-output triplets.

<new context (caption) 1>

<new context (layout) 1>

Please generate new Context-Instruction-Output triplets that meet the following requirements: …

Text-only ChatGPT

<new instruction 1>

<new output 1>

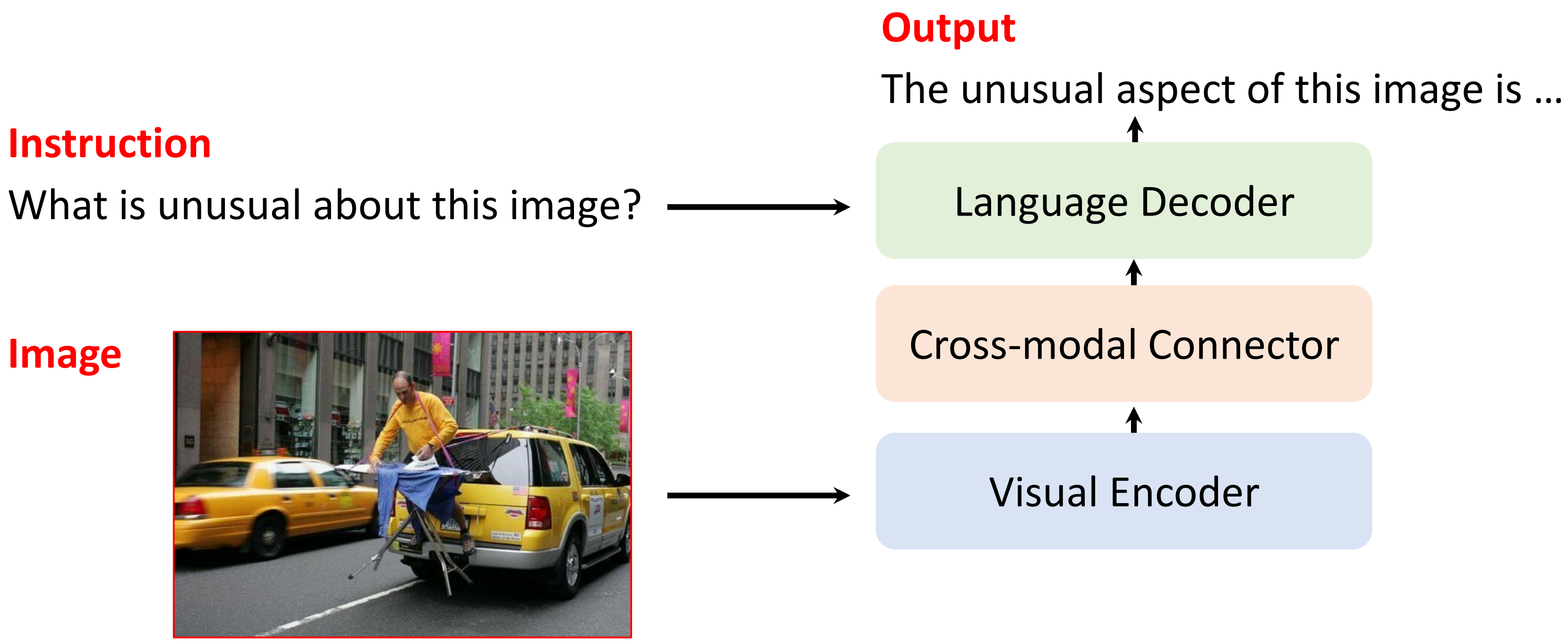
Visual Instruction-following Data:

Triplet (image, instruction, output)

What are the two people doing?

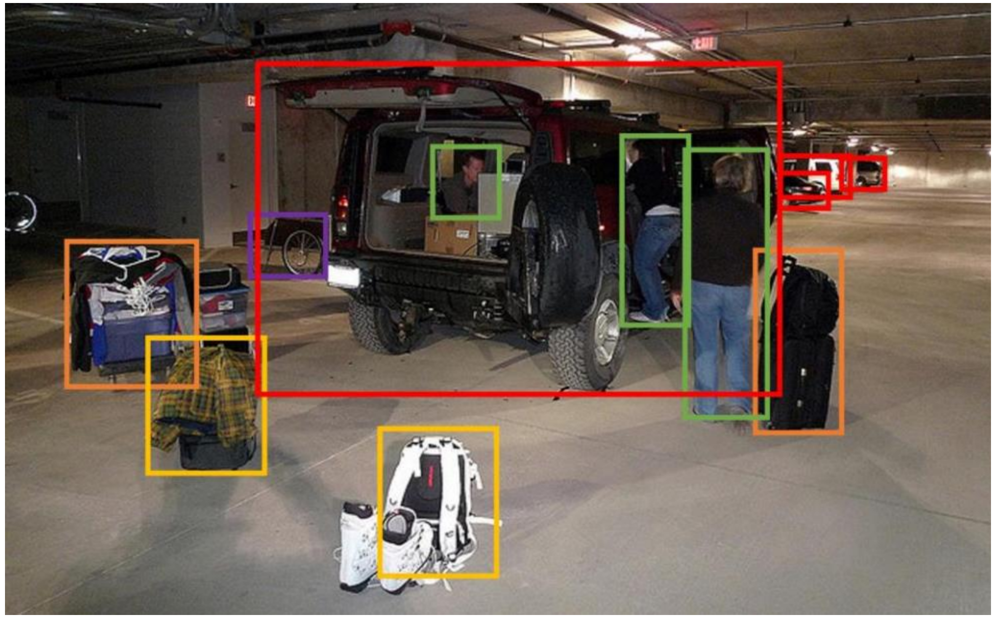
The two people are talking in front of a whiteboard about math …

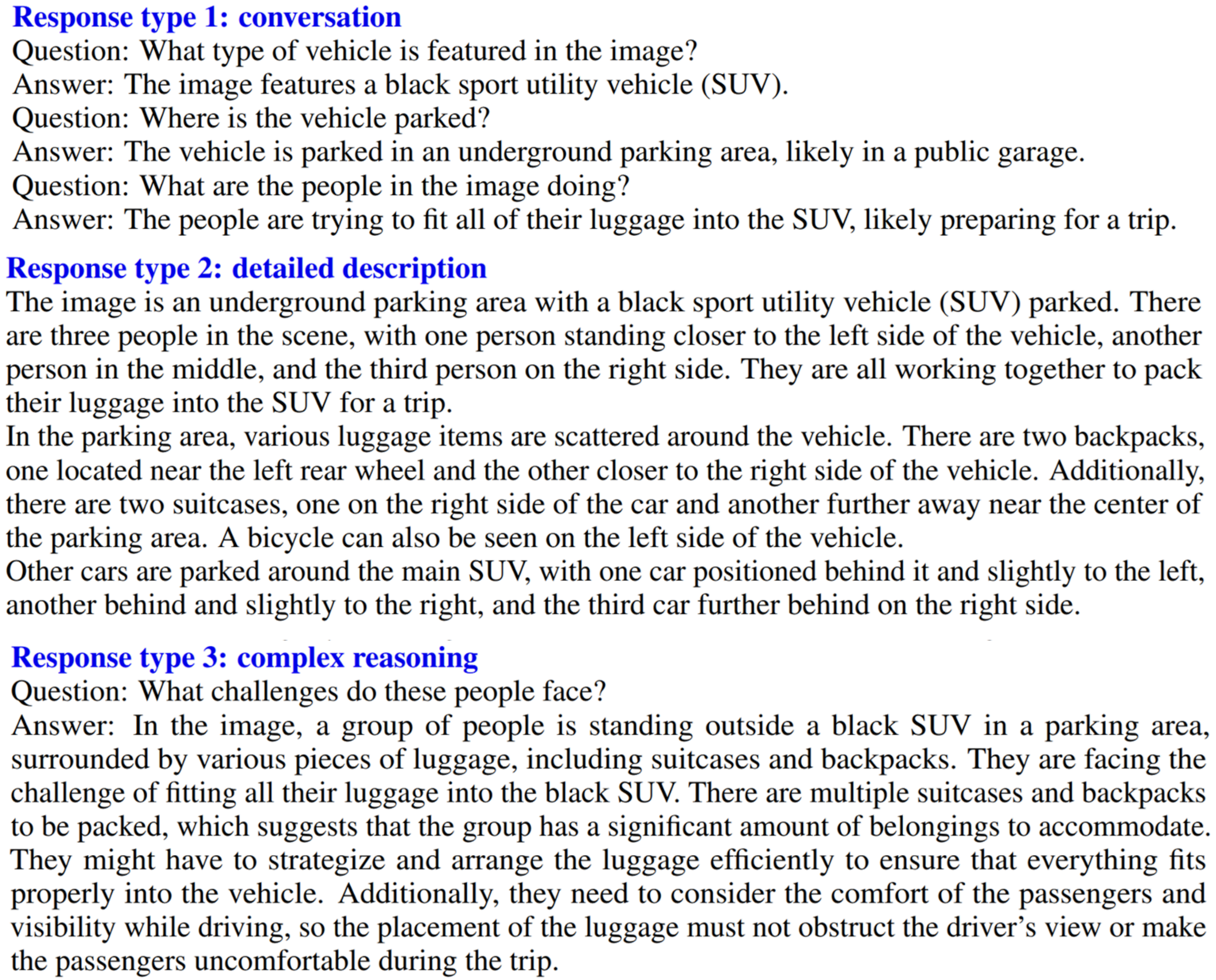
Instruction Tuning in Large Multimodal Models



- How to obtain Image-Instruction-Output triplet training data?
- Can we use a large language model for this?

Text-only GPT Assisted Visual Instruction Data Creation





Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV).
Question: Where is the vehicle parked?
Answer: The vehicle is parked in an underground parking area, likely in a public garage.
Question: What are the people in the image doing?
Answer: The people are trying to fit all of their luggage into the SUV, likely preparing for a trip.

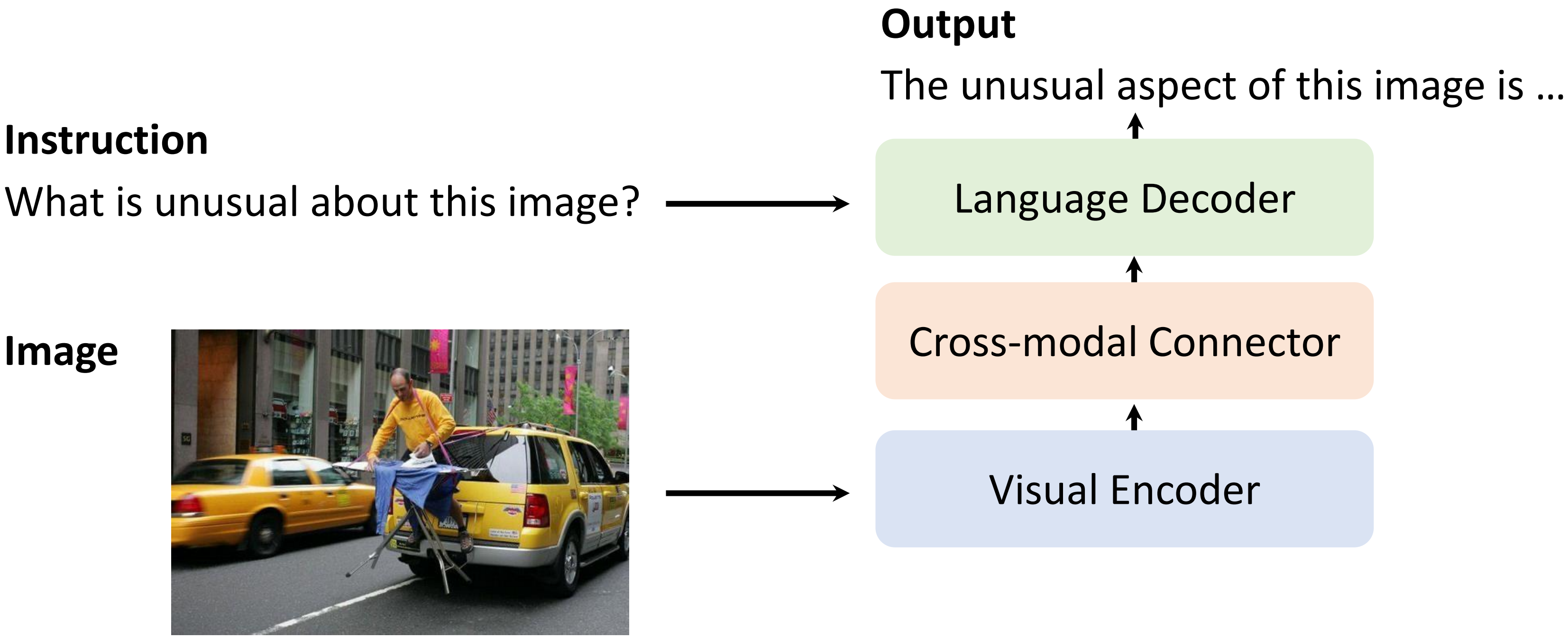
Response type 2: detailed description
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip.
In the parking area, various luggage items are scattered around the vehicle. There are two backpacks, one located near the left rear wheel and the other closer to the right side of the vehicle. Additionally, there are two suitcases, one on the right side of the car and another further away near the center of the parking area. A bicycle can also be seen on the left side of the vehicle.
Other cars are parked around the main SUV, with one car positioned behind it and slightly to the left, another behind and slightly to the right, and the third car further behind on the right side.

Response type 3: complex reasoning
Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings to accommodate. They might have to strategize and arrange the luggage efficiently to ensure that everything fits properly into the vehicle. Additionally, they need to consider the comfort of the passengers and visibility while driving, so the placement of the luggage must not obstruct the driver's view or make the passengers uncomfortable during the trip.

LLaVA-Instruct-158K

Conversation: 58K
Detailed description: 23K
Complex reasoning: 77K

How to Build Generalist Multimodal Models?



Output

The unusual aspect of this image is …

Instruction

What is unusual about this image?

Image

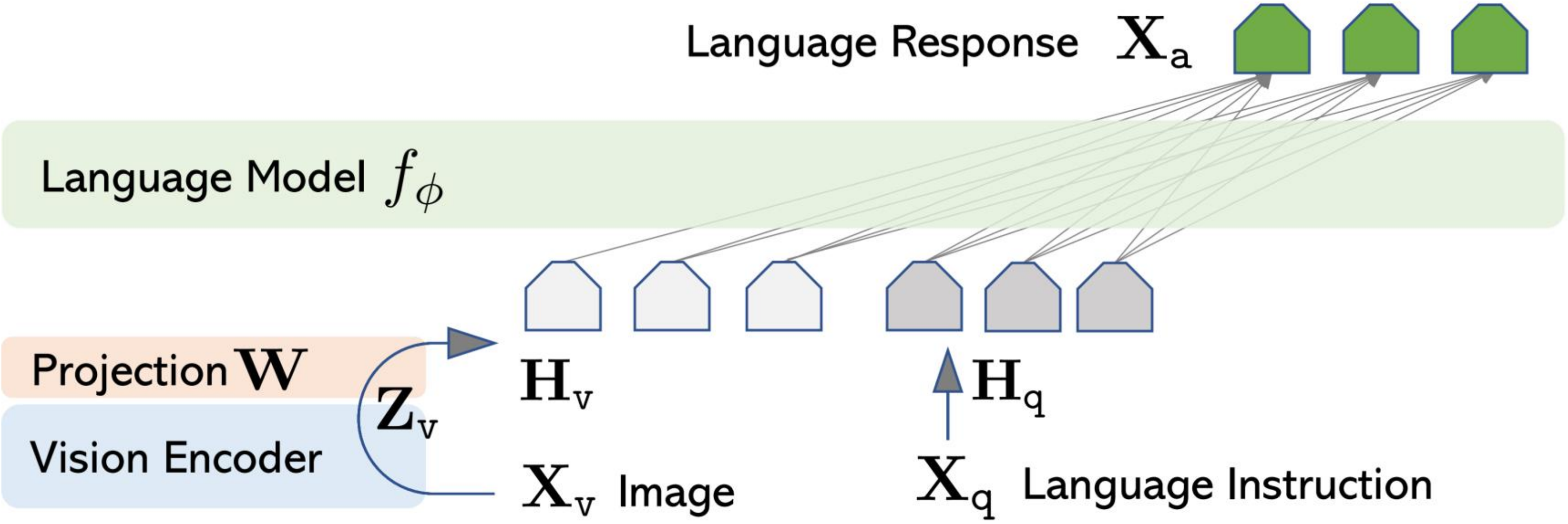
Language Decoder

Cross-modal Connector

Visual Encoder

1. Data to tune the model for instruction-following capabilities
2. Strong pretrained vision and language models
3. Connecting vision and language

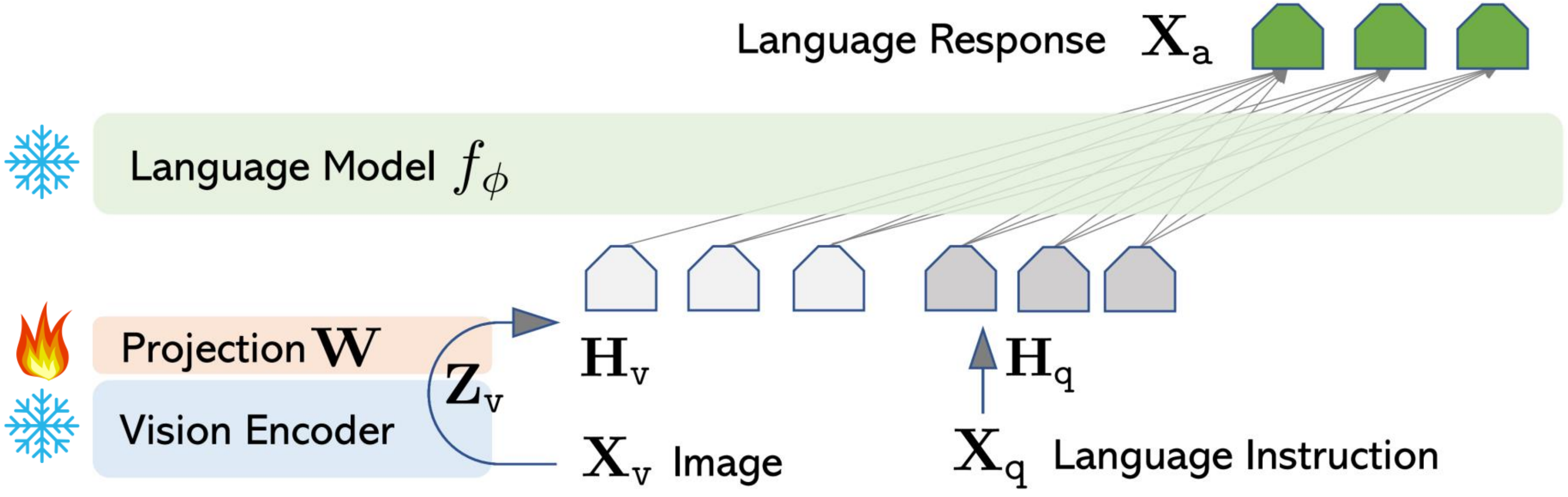
LLaVA: Large Language-and-Vision Assistant



Vision Encoder:	CLIP-ViT-L/14
Projection:	Linear layer (MLP in LLaVA-1.5)
Language Model:	Vicuna, LLaMA-2-Chat, MPT-Chat, etc.

[H. Liu et al. NeurIPS 2023. Visual Instruction Tuning. https://llava-vl.github.io]

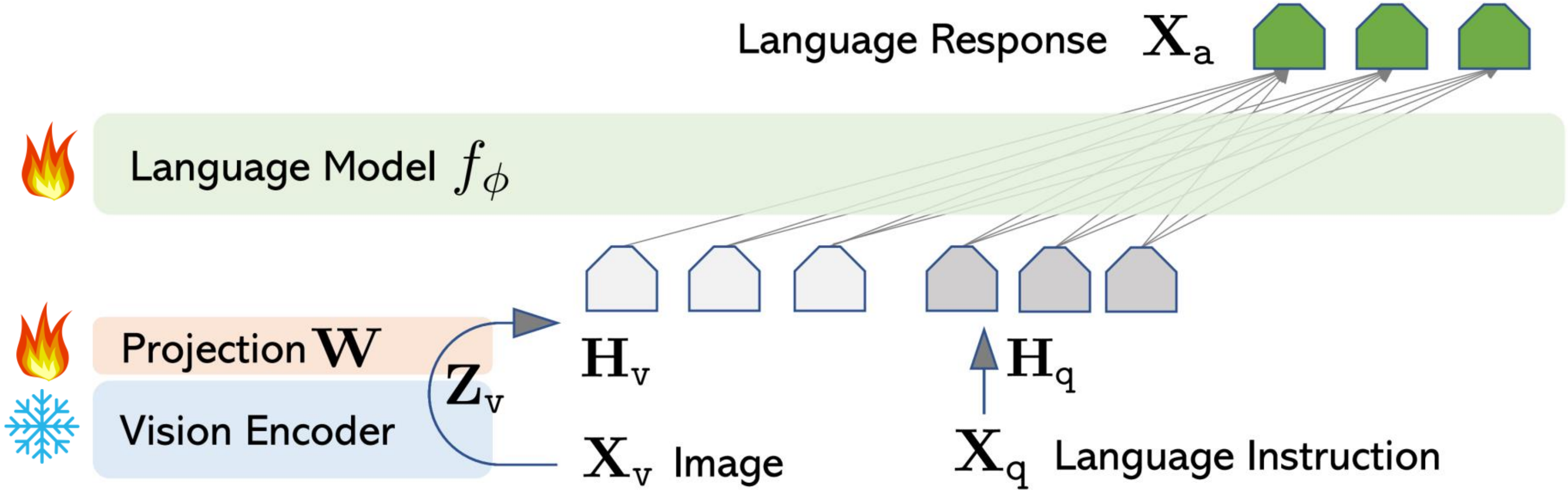
Stage 1: Pre-training for Feature Alignment



Data: Creative Captions (CC3M) data subset of 595K image-text pairs

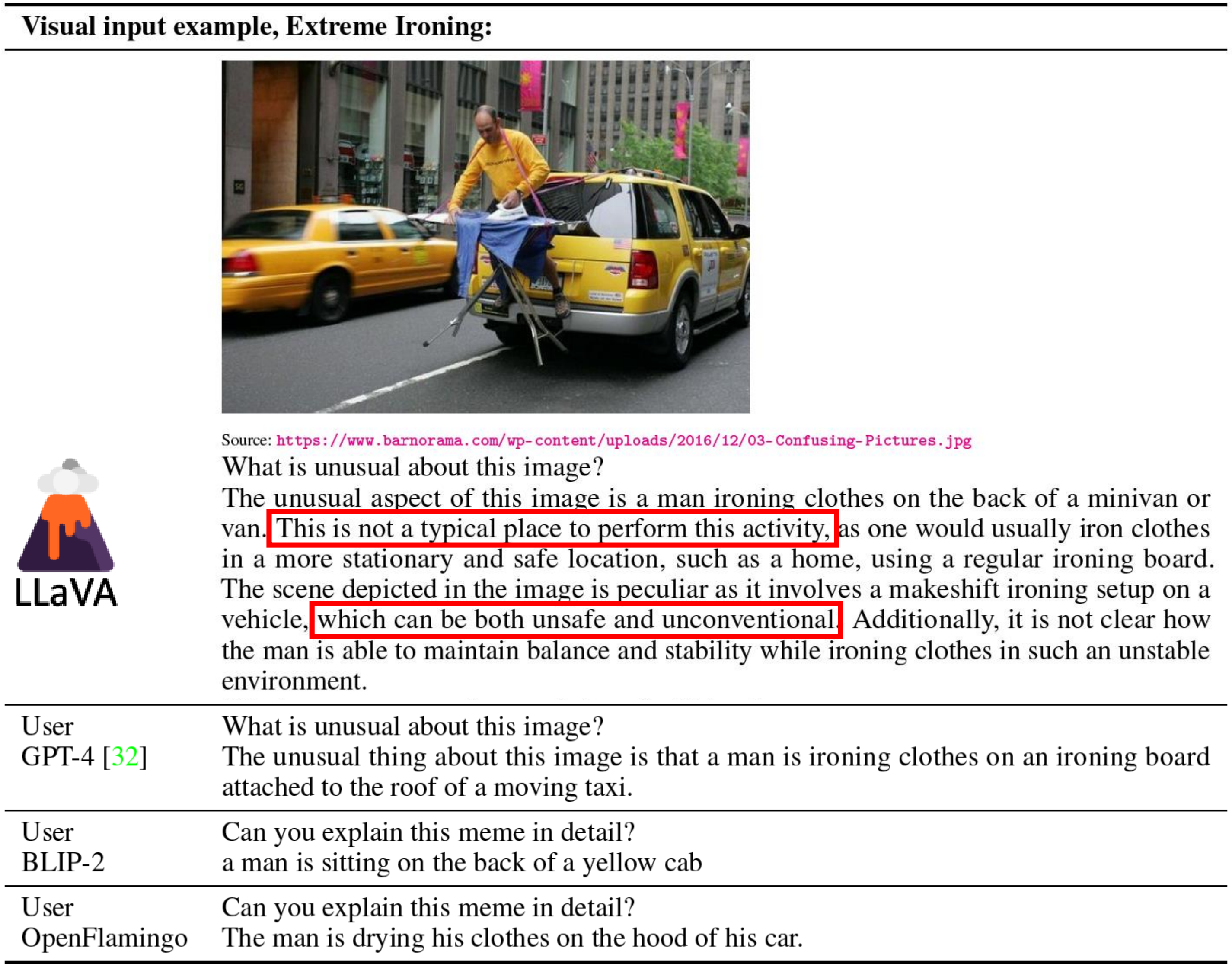
[H. Liu et al. NeurIPS 2023. Visual Instruction Tuning. https://llava-vl.github.io]

Stage 2: End-to-end Visual Instruction Tuning


Language Response $\mathbf{X}_{\mathrm{a}}$
Language Model f_{ϕ}
Projection $\mathbf{W}$
$\mathbf{Z}_{\mathrm{v}}$
$\mathbf{H}_{\mathrm{v}}$
$\mathbf{H}_{\mathrm{q}}$
Vision Encoder
$\mathbf{X}_{\mathrm{v}}$ Image
$\mathbf{X}_{\mathrm{q}}$ Language Instruction

Data: LLaVA-Instruct-158K for open-ended user-oriented visual instruction following tasks

[H. Liu et al. NeurIPS 2023. Visual Instruction Tuning. https://llava-vl.github.io]



Visual input example, Extreme Ironing:

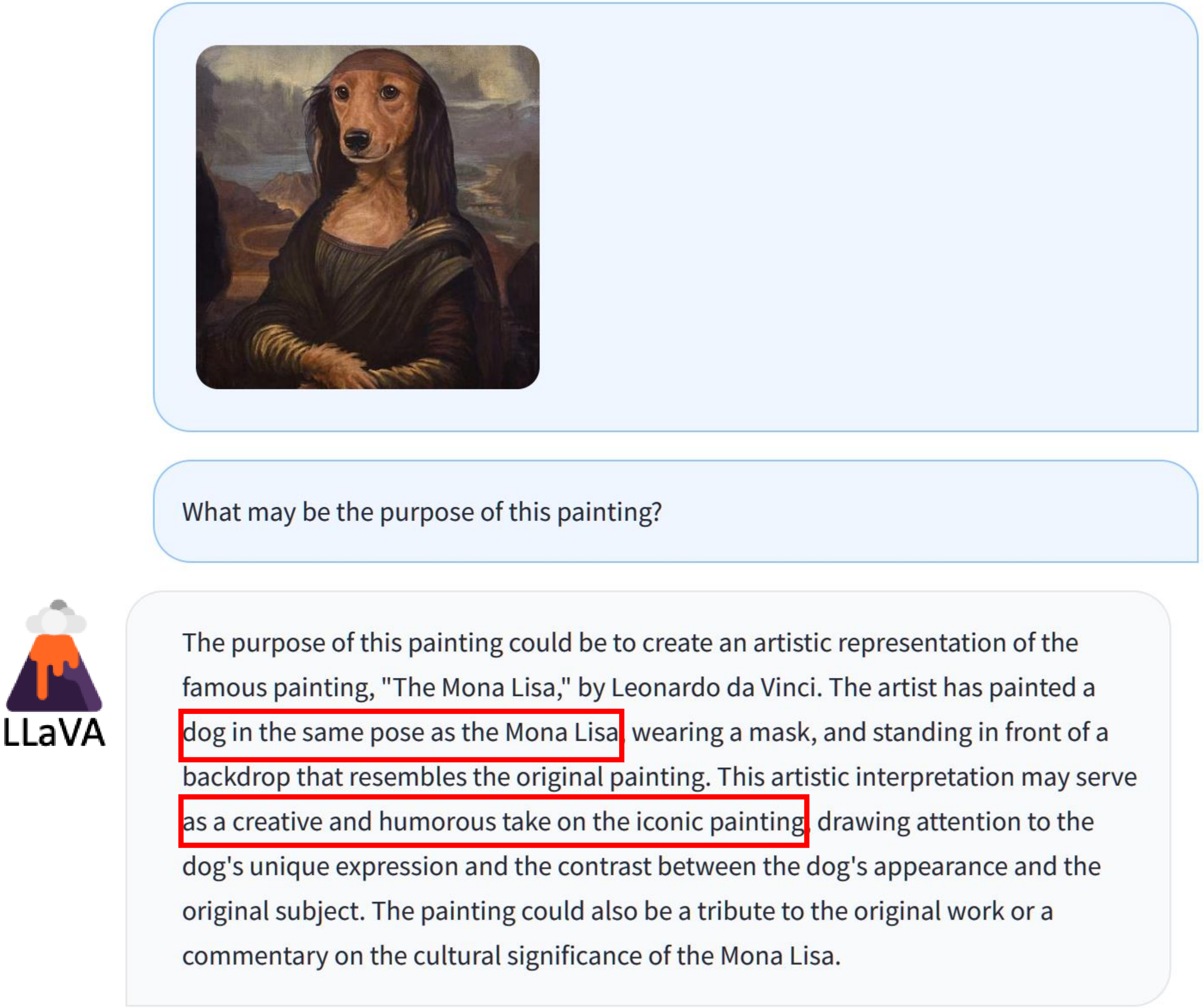
	Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg
LLaVA	What is unusual about this image? The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.
User GPT-4 [32]	What is unusual about this image? The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.
User BLIP-2	Can you explain this meme in detail? a man is sitting on the back of a yellow cab
User OpenFlamingo	Can you explain this meme in detail? The man is drying his clothes on the hood of his car.

Strong Visual Reasoning Capability

Example: *Extreme Ironing*

Strong Visual Reasoning Capability

Example: *Parodied Mona Lisa*

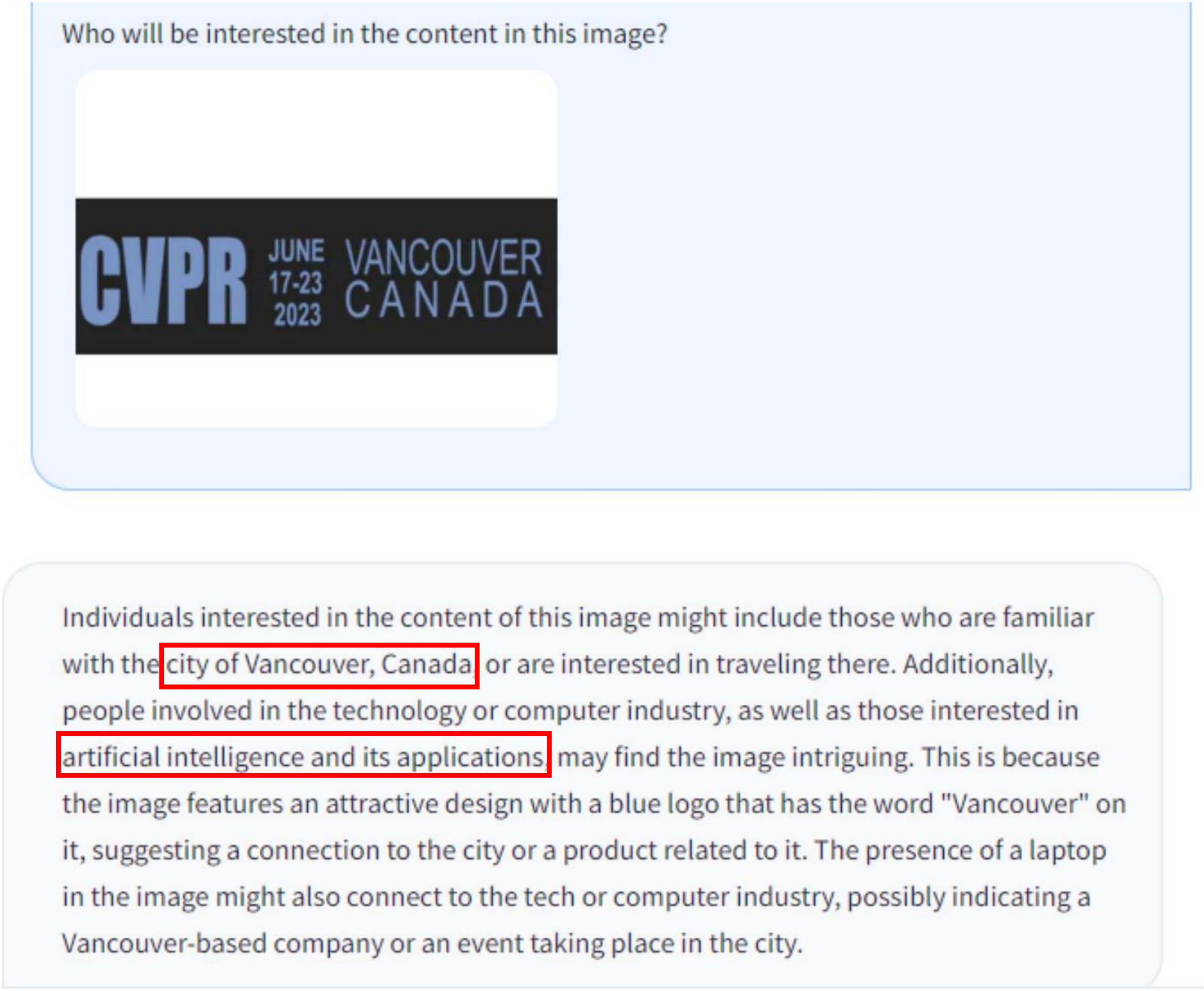


What may be the purpose of this painting?

The purpose of this painting could be to create an artistic representation of the famous painting, "The Mona Lisa," by Leonardo da Vinci. The artist has painted a dog in the same pose as the Mona Lisa, wearing a mask, and standing in front of a backdrop that resembles the original painting. This artistic interpretation may serve as a creative and humorous take on the iconic painting, drawing attention to the dog's unique expression and the contrast between the dog's appearance and the original subject. The painting could also be a tribute to the original work or a commentary on the cultural significance of the Mona Lisa.

Strong Emergent OCR Capability

Example: *CVPR & Vancouver*



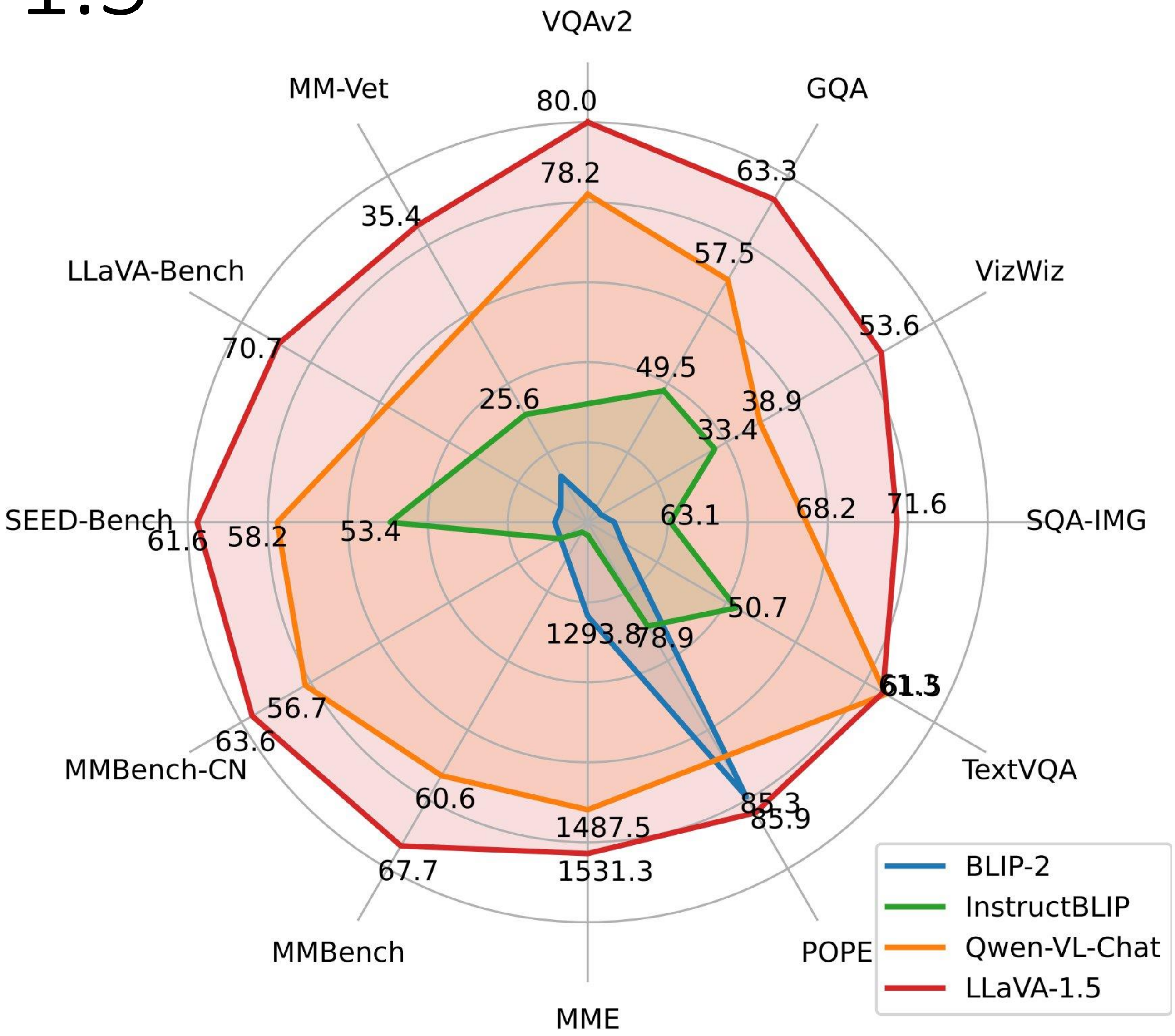
Who will be interested in the content in this image?

LLaVA

Individuals interested in the content of this image might include those who are familiar with the city of Vancouver, Canada or are interested in traveling there. Additionally, people involved in the technology or computer industry, as well as those interested in artificial intelligence and its applications may find the image intriguing. This is because the image features an attractive design with a blue logo that has the word "Vancouver" on it, suggesting a connection to the city or a product related to it. The presence of a laptop in the image might also connect to the tech or computer industry, possibly indicating a Vancouver-based company or an event taking place in the city.

Extensions: LLaVA-1.5

- Stronger performance on visual understanding benchmarks
- Better OCR, Yes/No answering, etc., due to scaling up data, model, image resolution



"Improved Baselines with Visual Instruction Tuning (LLaVA-1.5)" Haotian Liu, Chunyuan Li, Yuheng Li, Yong Jae Lee. CVPR 2024

Extensions: LLaVA-NeXT (1.6)

- Significantly outperforms LLaVA-1.5-13B
- Matches Gemini Pro on selected benchmarks

Model	**MMMU (val)**	**MMMU (test)**	**MathVista**	**MMBench-EN**	**MMBench-CN**	**MM-Vet**
GPT-4V	56.8	55.7	49.9	75.8	73.9	67.6
Gemini Ultra	59.4	-	53	-	-	-
Gemini Pro	47.9	-	45.2	73.6	74.3	64.3
LLaVA-1.5-13B	36.4	33.6	27.6	67.8	63.3	36.3
LLaVA-1.6-34B	**51.1**	**45.3**	**46.5**	**79.3**	**79.0**	**57.4**

“LLaVA-NeXT: Improved reasoning, OCR, and world knowledge” Haotian Liu et al. January 2024 (blog)

Combinatorial Task Generalization

Seen Training Data

Multilingual Text-only Conversation

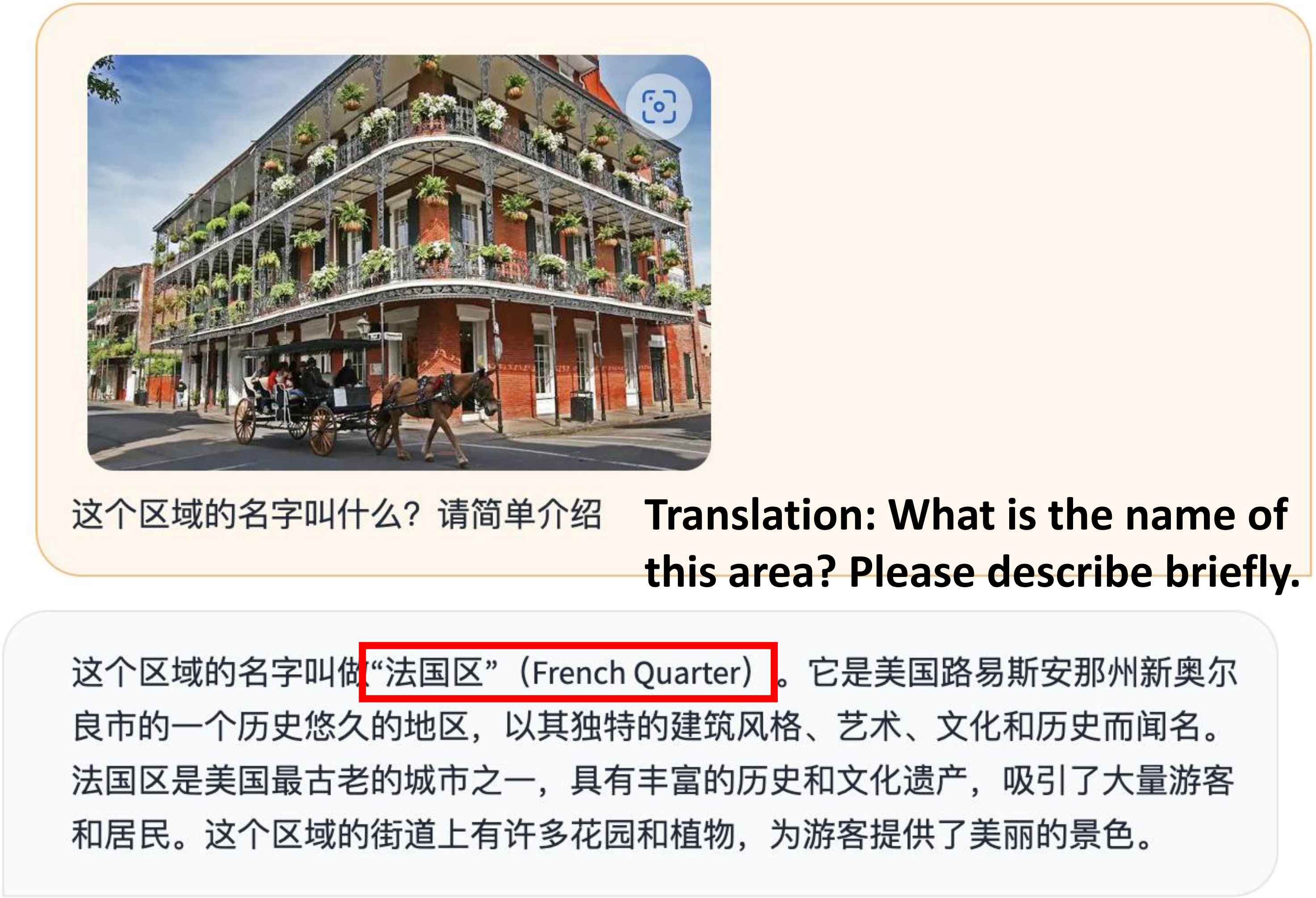
English-Only Visual Conversations

Generalized Capabilities

Multilingual Visual Conversations

Emergent Multilingual Capability

Example: *French Quarter*



这个区域的名字叫什么？请简单介绍

Translation: What is the name of this area? Please describe briefly.

LLaVA

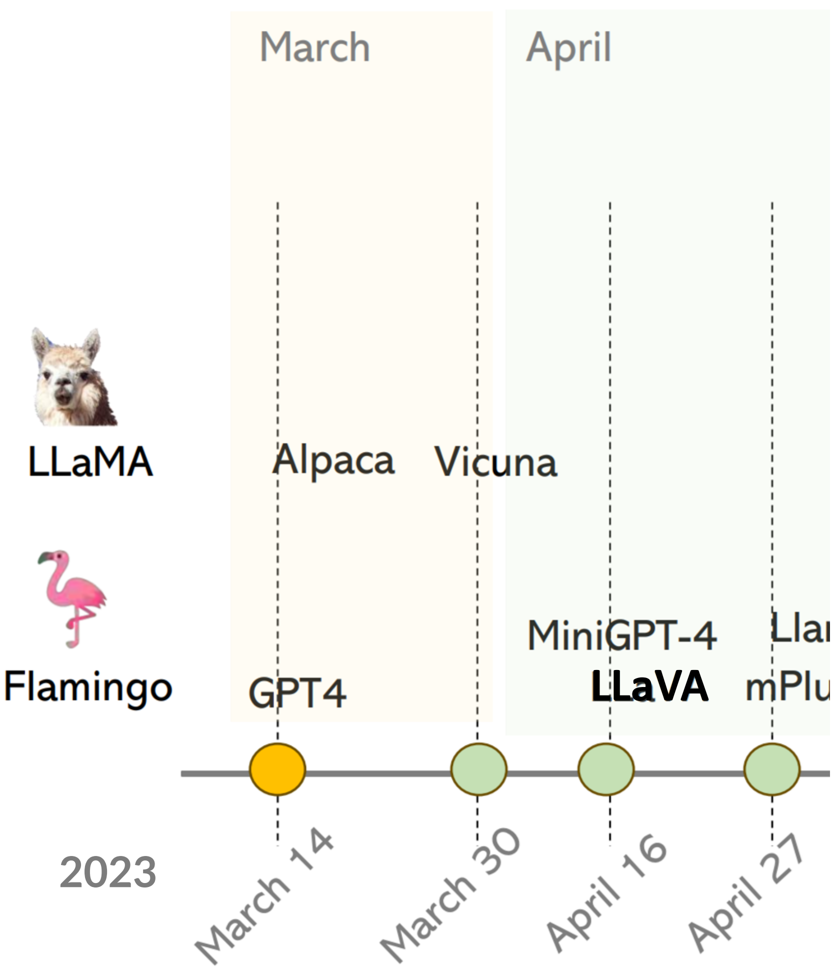
这个区域的名字叫做“法国区”（French Quarter）。它是美国路易斯安那州新奥尔良市的一个历史悠久的地区，以其独特的建筑风格、艺术、文化和历史而闻名。法国区是美国最古老的城市之一，具有丰富的历史和文化遗产，吸引了大量游客和居民。这个区域的街道上有许多花园和植物，为游客提供了美丽的景色。

Combinatorial Task Generalization

Seen Training Data		**Generalized Capabilities**
Multilingual Text-only Conversation	English-Only Visual Conversations	**Multilingual Visual Conversations**
VQA/OCR data	Visual Conversations	**Improved Visual Groundedness / OCR** in **Visual Conversations**
Longer Writing Text-only Conversations	Shorter (casual) Visual Conversations	**Improved Writing in Visual Conversations**

Do not need to create all combinations of data in training; let LMMs generalize!

Community Efforts on LMMs



March

April

LLaMA

Alpaca

Vicuna

Flamingo

GPT4

MiniGPT-4

LLaVA

Lla

mPlu

2023

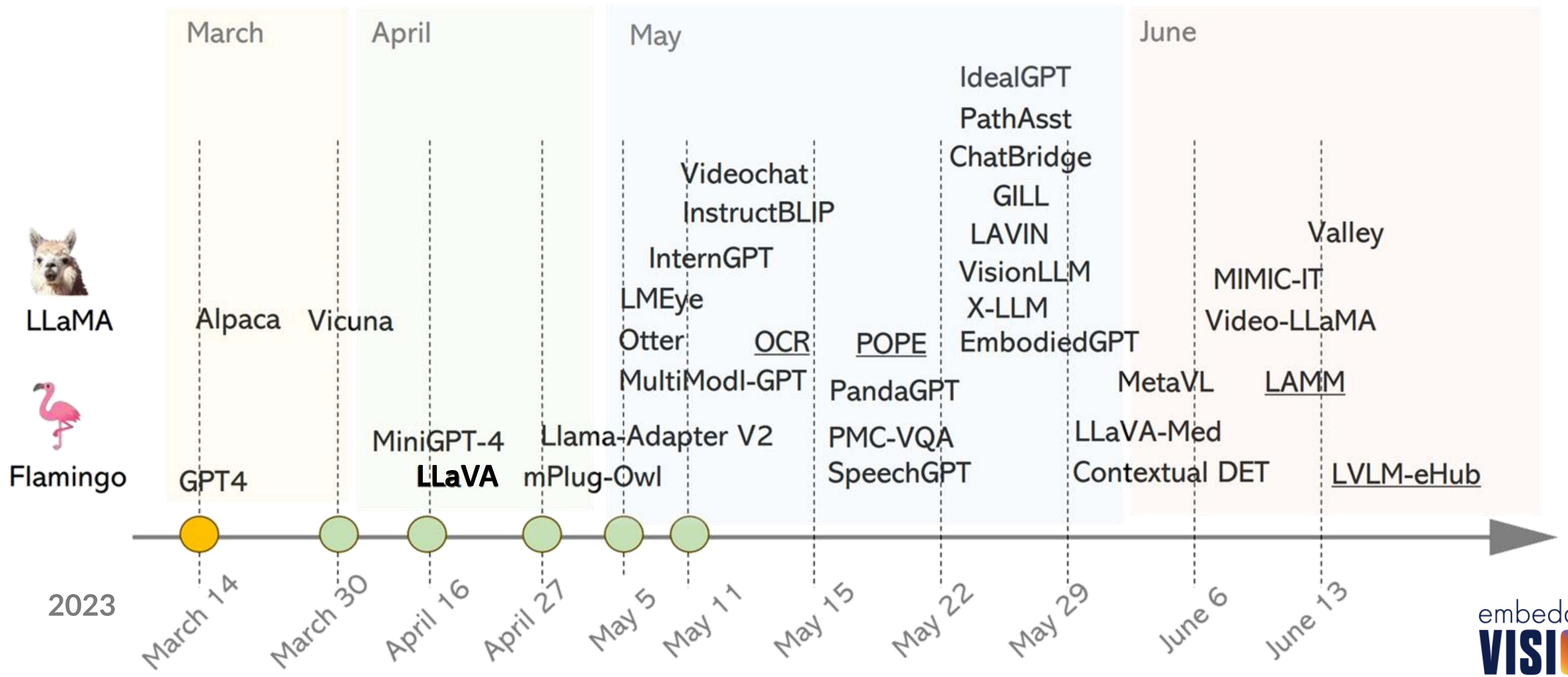
March 14

March 30

April 16

April 27

Community Efforts on LMMs



March

April

May

June

LLaMA

Flamingo

2023

March 14: Alpaca, GPT4

March 30: Vicuna

April 16: MiniGPT-4, **LLaVA**

April 27: Llama-Adapter V2, mPlug-Owl

May 5: LMEye, Otter, MultiModl-GPT, InternGPT

May 11: Videochat, InstructBLIP, OCR

May 15: POPE, PandaGPT, PMC-VQA, SpeechGPT

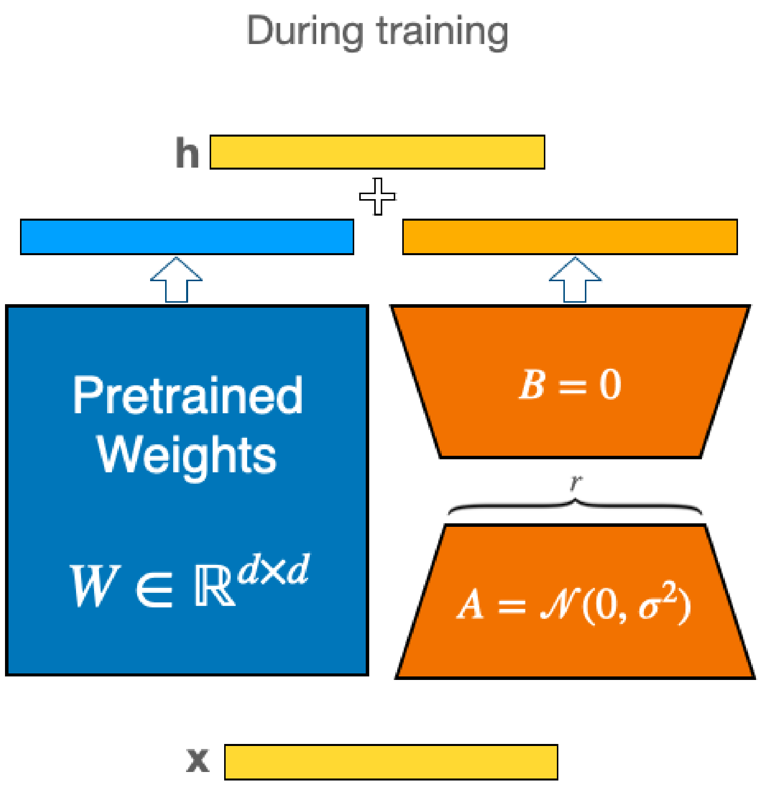
May 22: IdealGPT, PathAsst, ChatBridge, GILL, LAVIN, VisionLLM, X-LLM, EmbodiedGPT

May 29: MetaVL, LLaVA-Med, Contextual DET

June 6: MIMIC-IT, Video-LLaMA, LAMM

June 13: Valley, LVLM-eHub

How to Train (Fine-tune) Large Models Efficiently?


During training
h
Pretrained Weights
$W \in \mathbb{R}^{d \times d}$
$B = 0$
r
$A = \mathcal{N}(0, \sigma^2)$
x

- Parameter Efficient Fine-Tuning (e.g., Low-Rank Adaptation, Hu et al. 2021)
- LLaVA can be fine-tuned with LoRA

Image Source: https://huggingface.co/docs/peft/main/en/conceptual_guides/lora

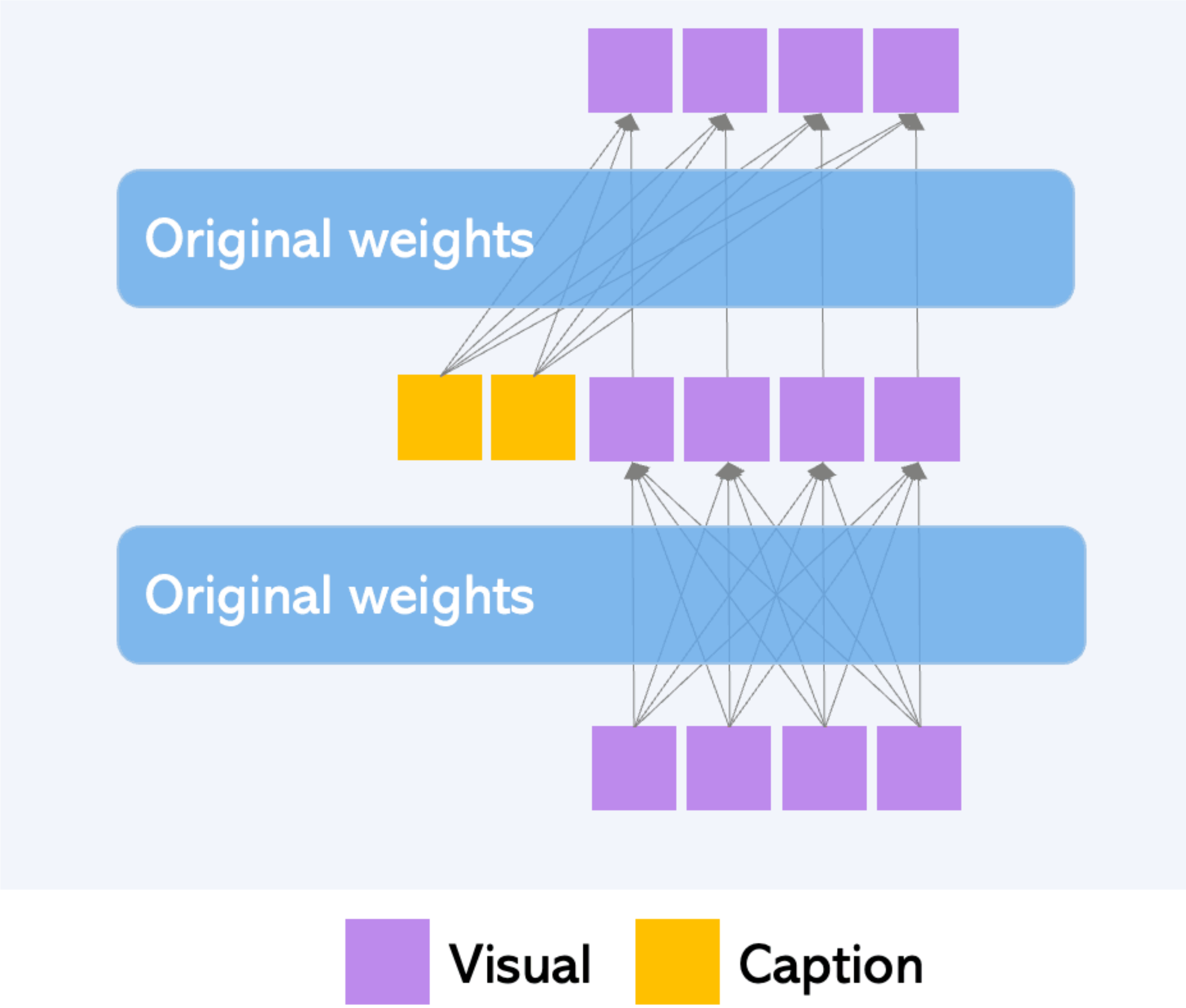
GLIGEN: Grounded Language-Image Generation

Text prompt: “A hen is hatching a huge egg”

Stable-Diffusion

- Efficiently converts a text-to-image (T2I) model to grounded generation model

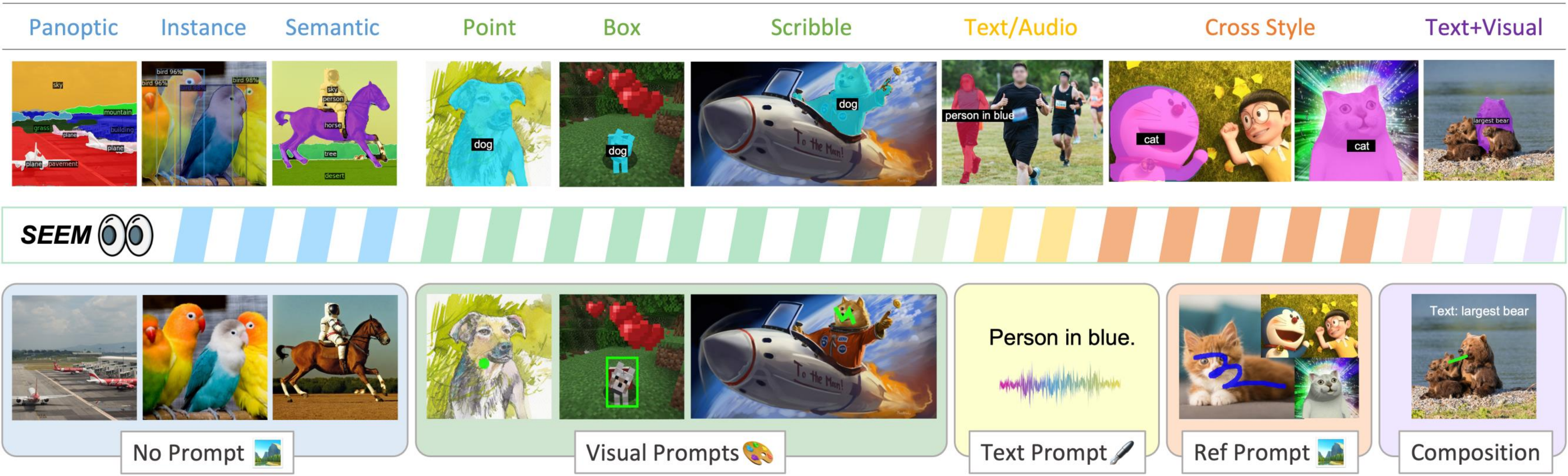
[GLIGEN: Open-Set Grounded Text-to-Image Generation, Yuheng Li et al., CVPR 2023.]



T2I Generation (e.g., Stable Diffusion)

Caption: “a hen is hatching a huge egg”

Segment Everything Everywhere All at Once

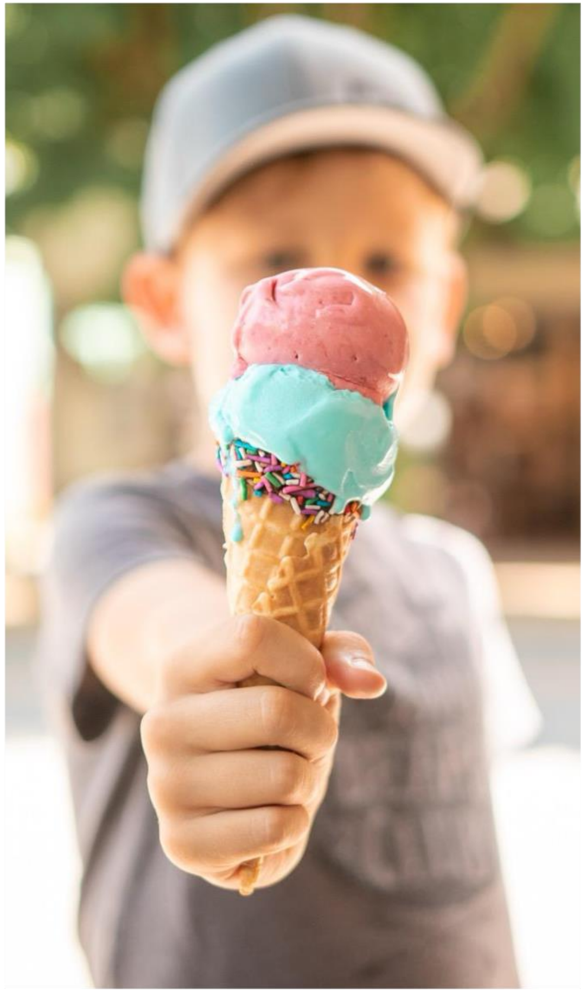


- Generalist segmentation model that can be prompted with text and visual inputs

[Segment Everything Everywhere All At Once, Zou et al., NeurIPS 2023.]

Looking Forward: Is Visual Understanding Solved?

Not quite …



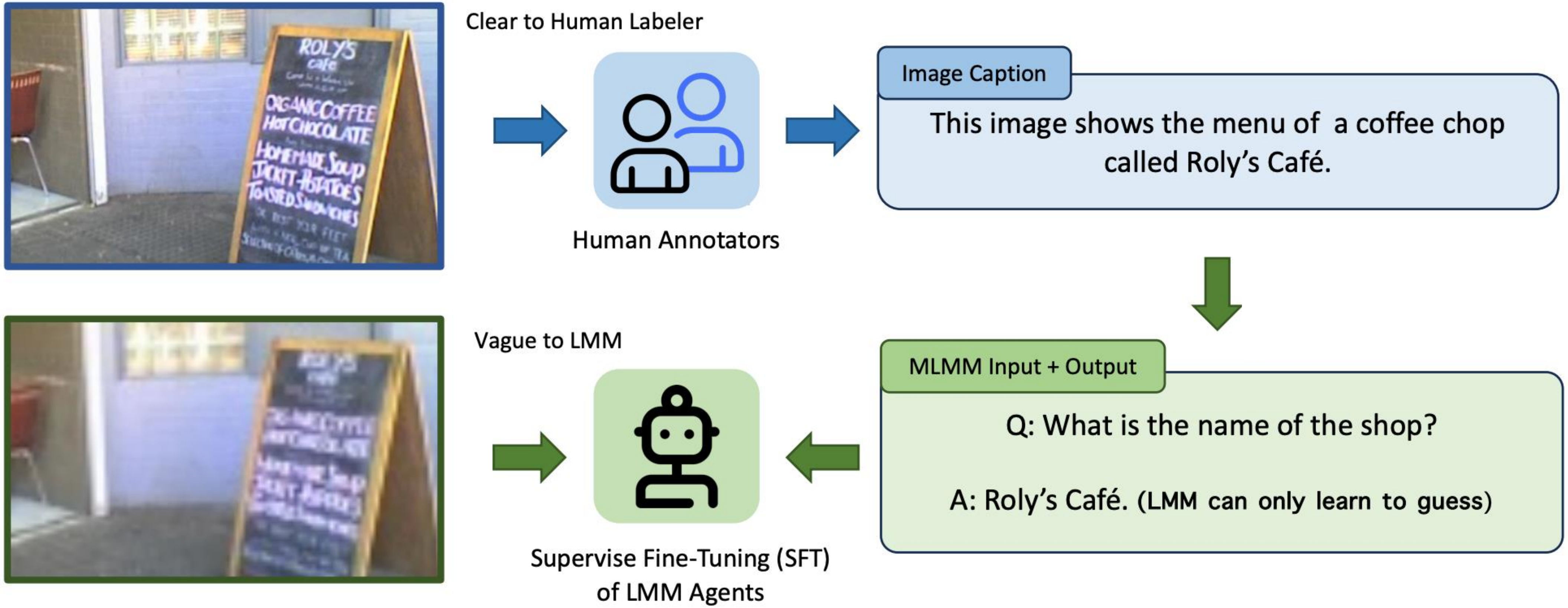
👤 : ***Which caption correspond to this image?***
0: a pink scoop on top and a blue scoop on bottom
1: a pink scoop on bottom and a blue scoop on top

: The image shows an ice cream cone with a blue scoop of ice cream on top and what appears to be a pink scoop on the bottom. Therefore, Caption 1 is correct.

Looking Forward: Limitations of Current Models

- Capabilities
  - Hallucinations
  - Alignment without forgetting
  - Video understanding
  - Smaller performant models
  - …
- Understanding
  - Origination of emergent behaviors like OCR
  - How does the performance of LLMs affect the capability of the LMMs
  - Impact of instruction tuning on knowledge
  - …

Looking Forward: Hallucinations in LMMs



Clear to Human Labeler

Human Annotators

Image Caption

This image shows the menu of a coffee chop called Roly's Café.

Vague to LMM

Supervise Fine-Tuning (SFT) of LMM Agents

MLMM Input + Output

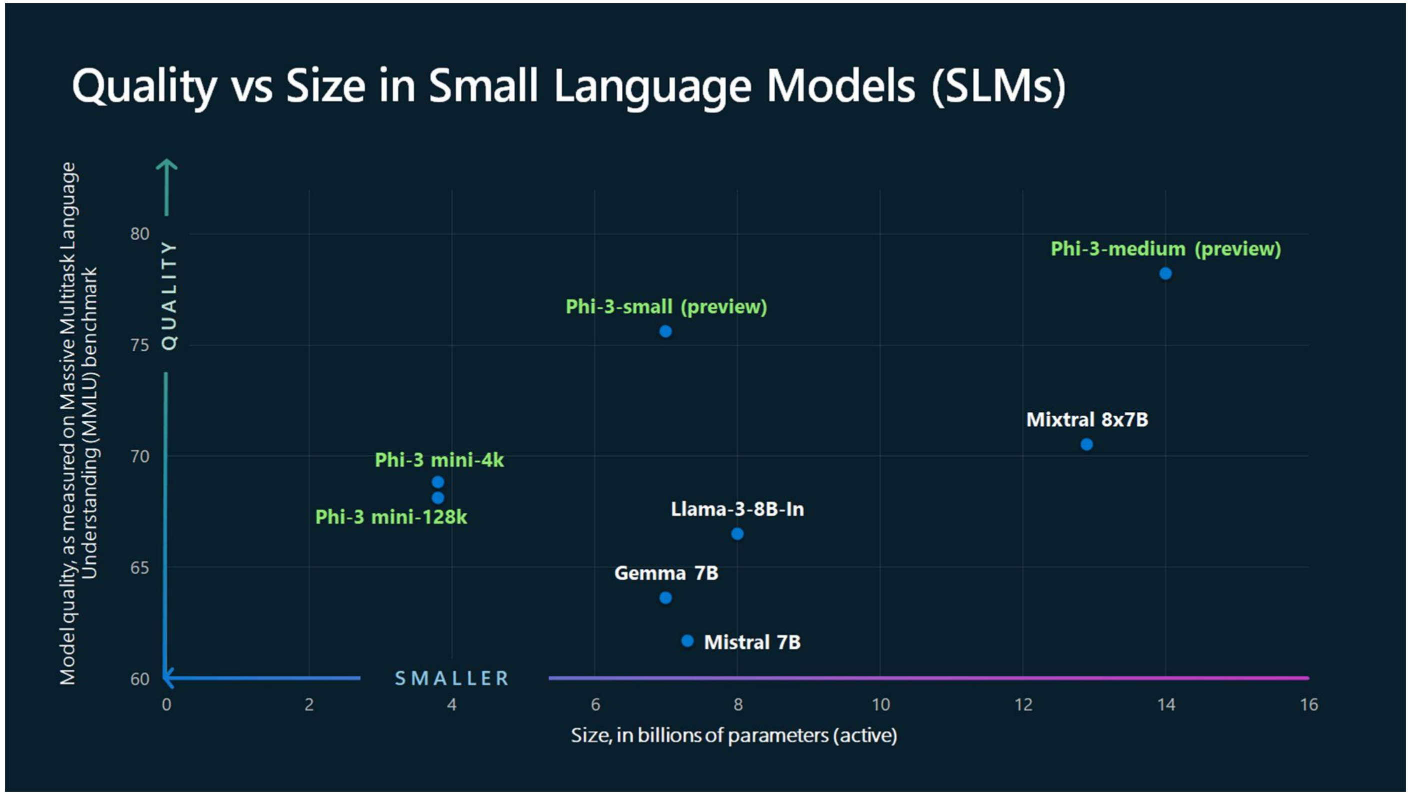
Q: What is the name of the shop?

A: Roly's Café. (LMM can only learn to guess)

- When a task is beyond a model's capabilities, SFT encourages it to hallucinate

Image Source: Aligning Large Multimodal Models with Factually Augmented RLHF, Sun et al. 2023

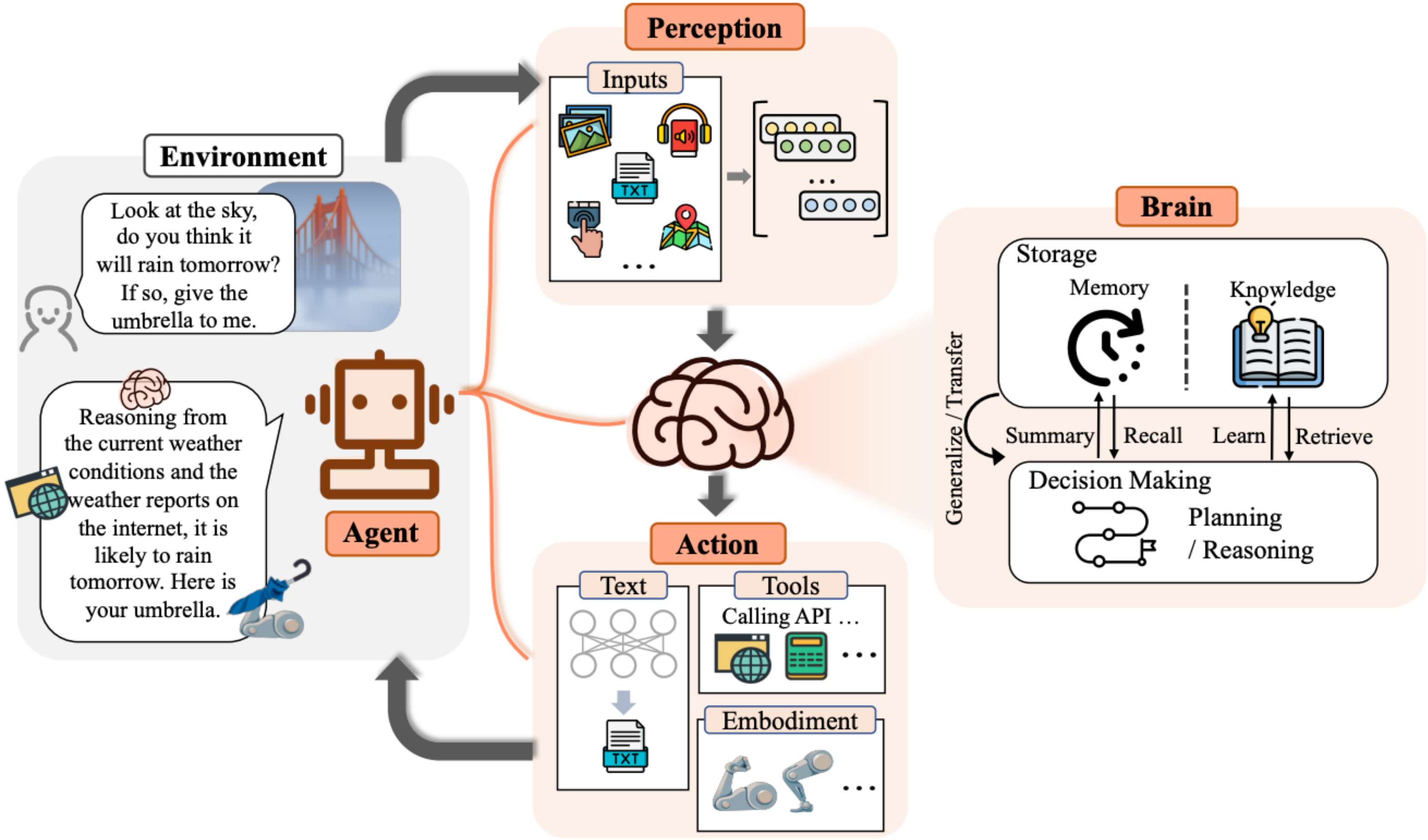
Looking Forward: Smaller Models



- (Small models + high quality data) ≈ (larger models + lower quality data)
- LLaVA w/ Phi-3 LLM for multimodal shows similar trends

Image Source: Microsoft

Looking Forward: Multimodal AI Agents


Environment
Look at the sky, do you think it will rain tomorrow? If so, give the umbrella to me.
Reasoning from the current weather conditions and the weather reports on the internet, it is likely to rain tomorrow. Here is your umbrella.
Agent
Perception
Inputs
TXT
...
Brain
Storage
Memory
Knowledge
Summary
Recall
Learn
Retrieve
Generalize / Transfer
Decision Making
Planning
/ Reasoning
Action
Text
TXT
Tools
Calling API ...
...
Embodiment
...

- AI Agents that can self reflect, use tools, plan, and collaborate with other agents

Image Source: The Rise and Potential of Large Language Model Based Agents: A Survey, Xi et al. 2023

Conclusions

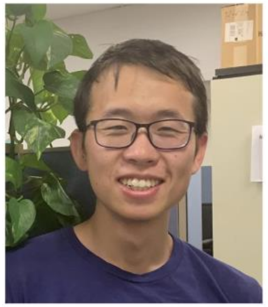
- From *specialist* to *generalist* multi-modal models
- Controllable ("aligned") image understanding for open-world concepts
- Build upon pre-trained foundation models, design semi-automatic data collection methods
- Code, models, online demo available:

  https://llava-vl.github.io/, https://gligen.github.io/, https://github.com/UX-Decoder

Thank you

- Haotian Liu, Yuheng Li, Utkarsh Ojha, Mu Cai, Xueyan Zou, Chunyuan Li, Jianwei Yang, Jianfeng Gao

Haotian Liu


Yuheng Li

Utkarsh Ojha

Mu Cai

Xueyan Zou

Questions and Answers

Yong Jae Lee
University of Wisconsin-Madison and GivernyAI

Text your questions to +1 408-400-2702