

Generalist Multimodal Models

Yong Jae Lee

University of Wisconsin-Madison



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

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

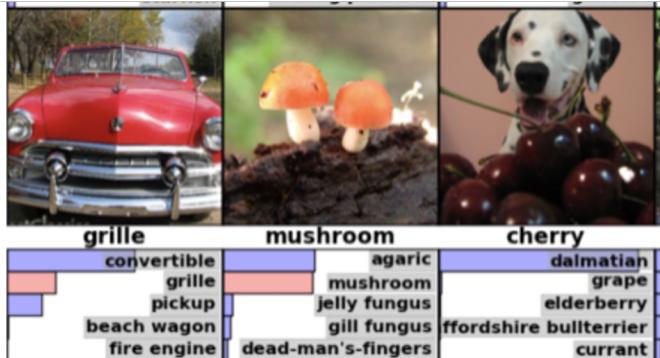
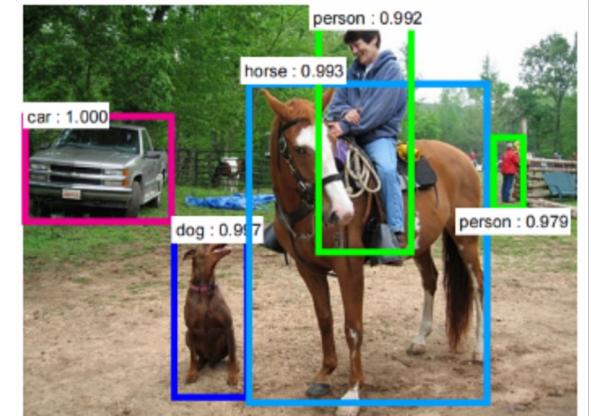


Image classification



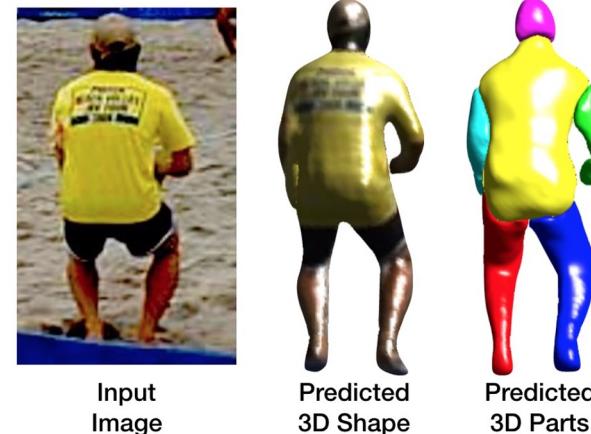
Semantic segmentation



Object detection



Pose recognition



3D prediction

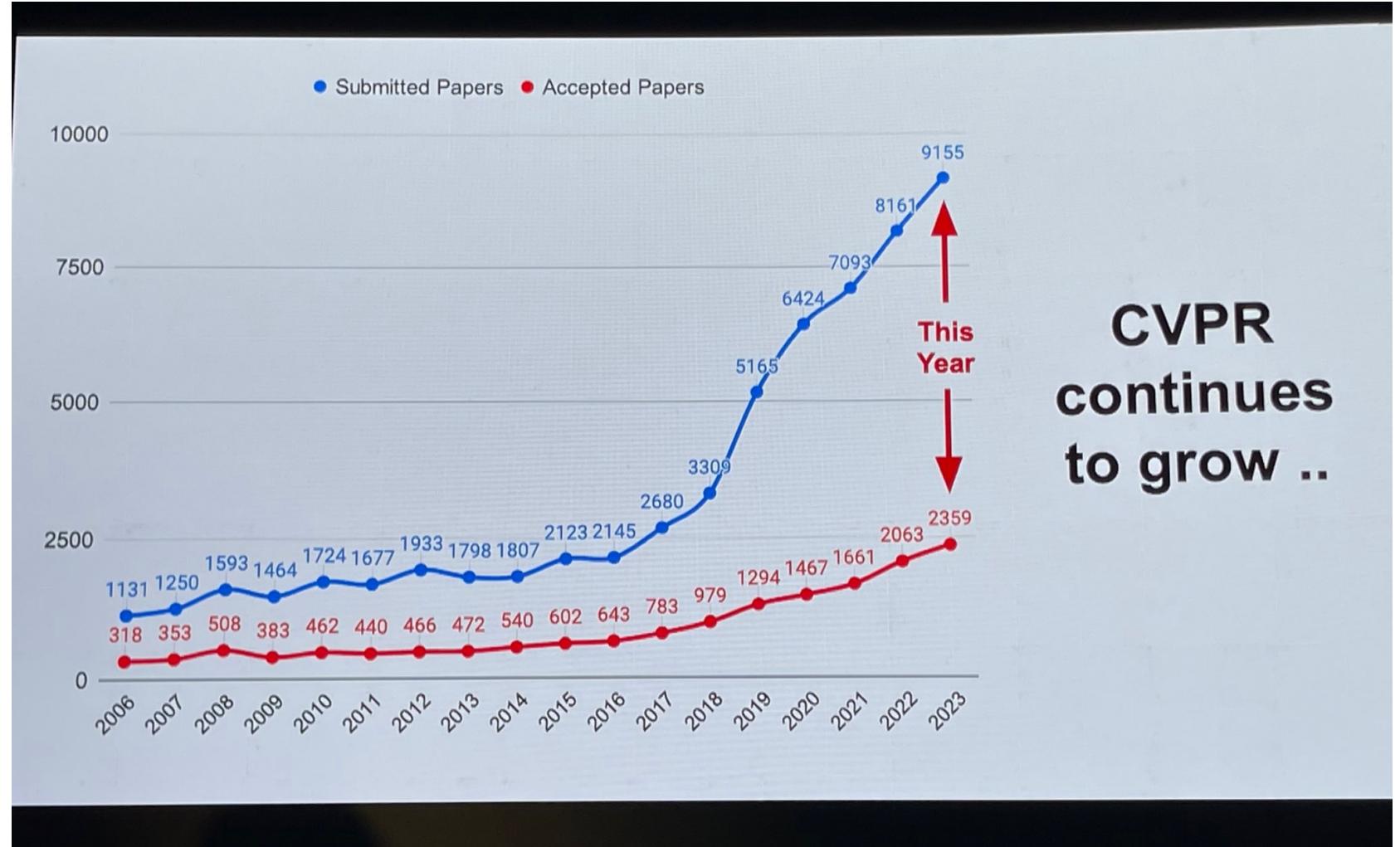


Surface normal prediction

... and many more

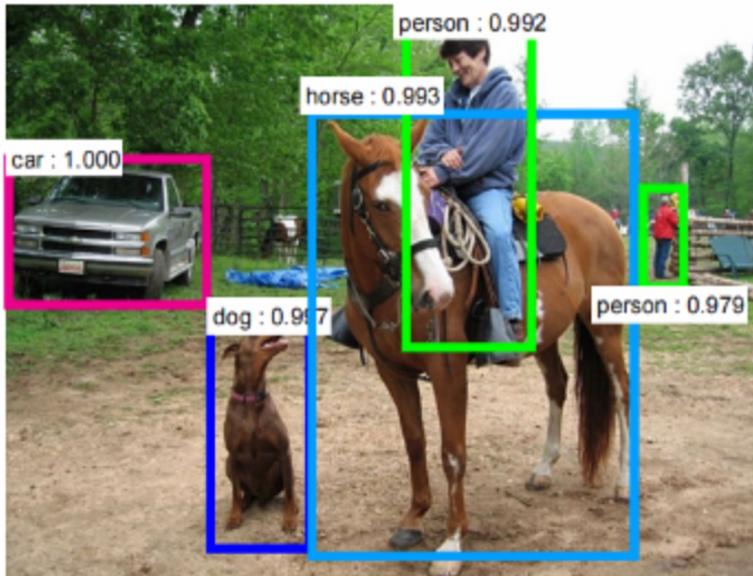
Explosion in ...

Students!
Startups!
Funding!
Hiring!



However, Prevailing Paradigm Thus Far:

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



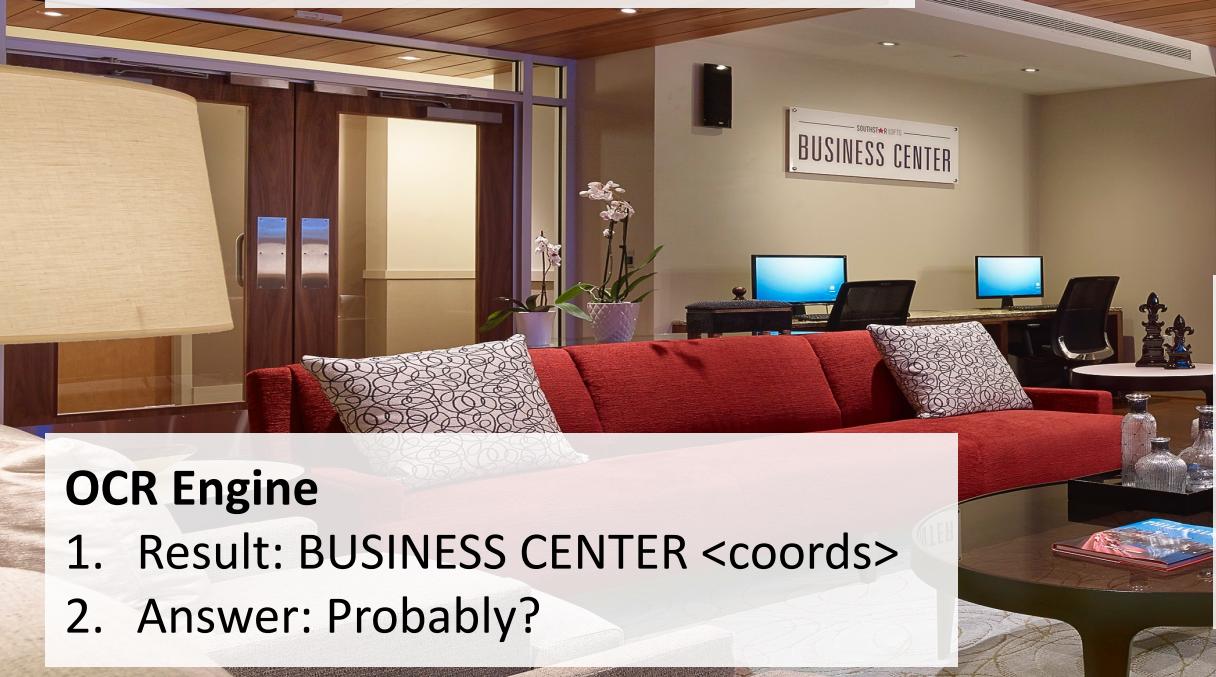
Object Detection *Only*



Pose Recognition *Only*

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?



User: Can I print my documents here?

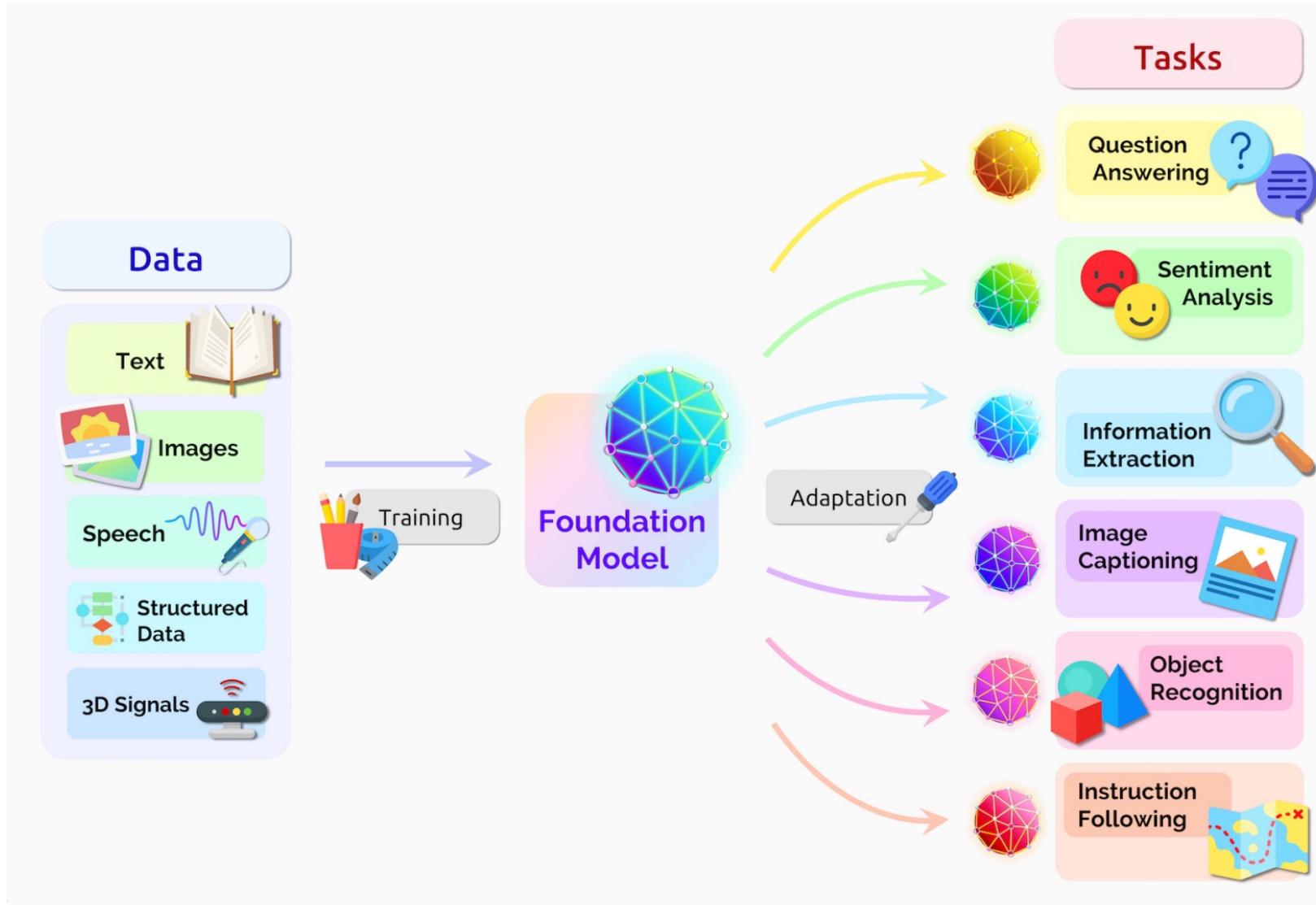
Final output to the user



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

Specialist models are insufficient
for open-ended generic tasks

Rise of “Generalist” Foundation Models (2020s)

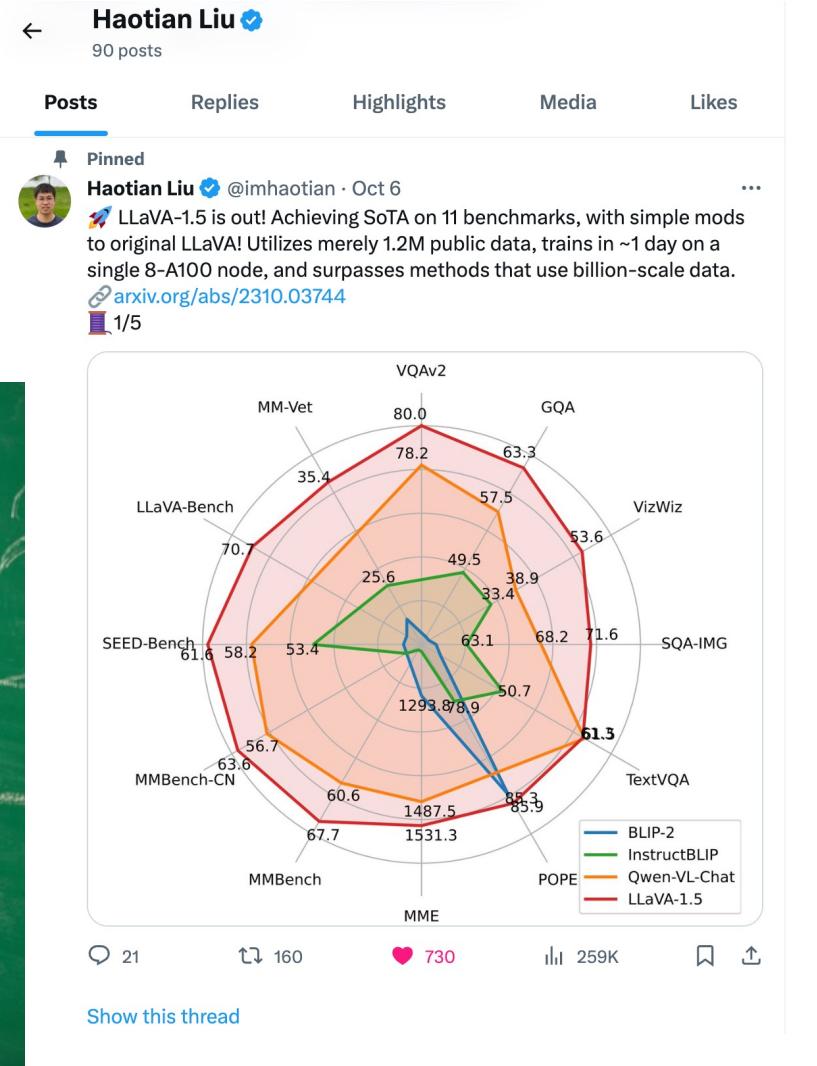


- **Single-model, many tasks**
- Large Language Models (e.g., GPT4)
- Vision Transformers
- Image-Text Models (e.g., CLIP)

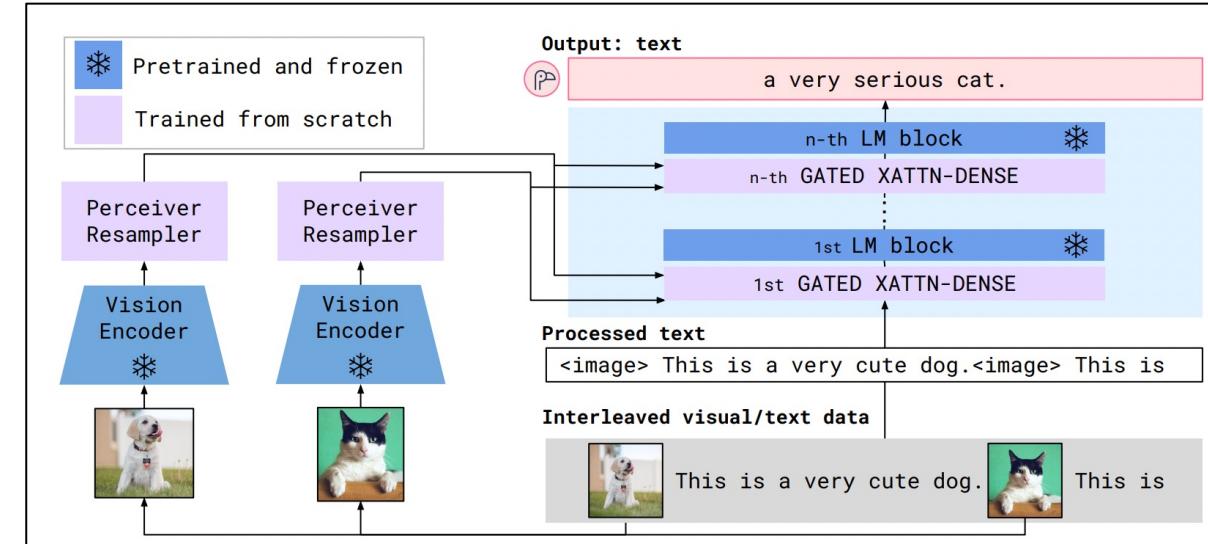
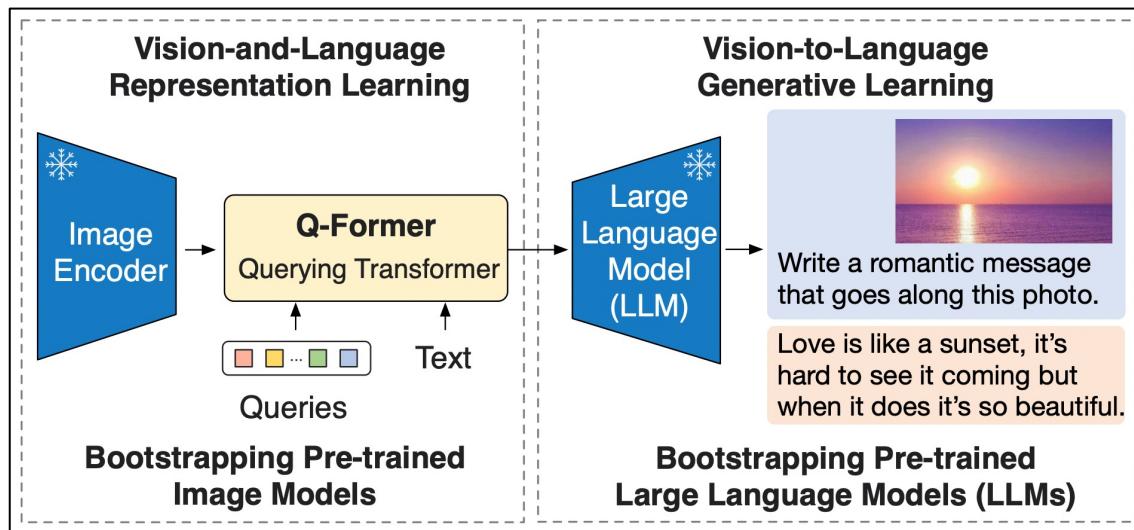
Today's talk: Large Generalist Multimodal Models

- *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



How to Build Generalist Multimodal Models?



- Prior methods (before early 2023) generally lacked 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:

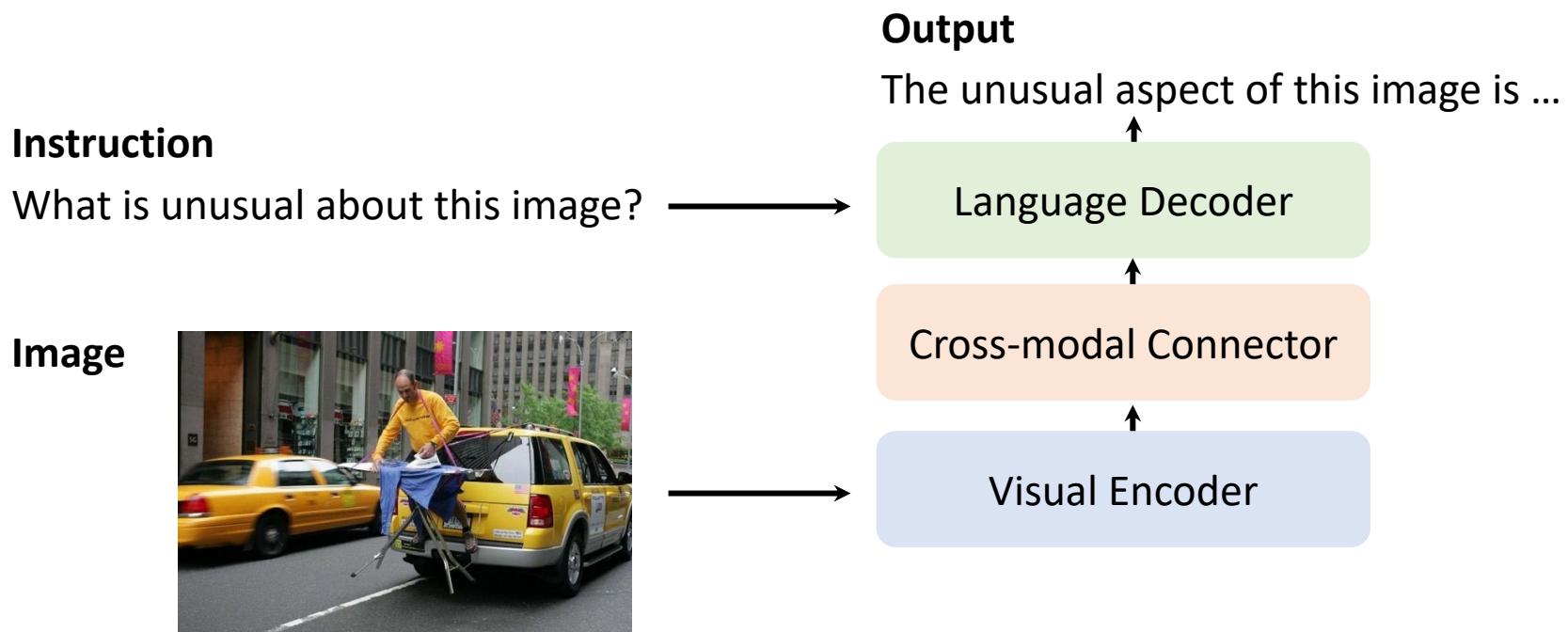
User What is unusual about this image?

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.



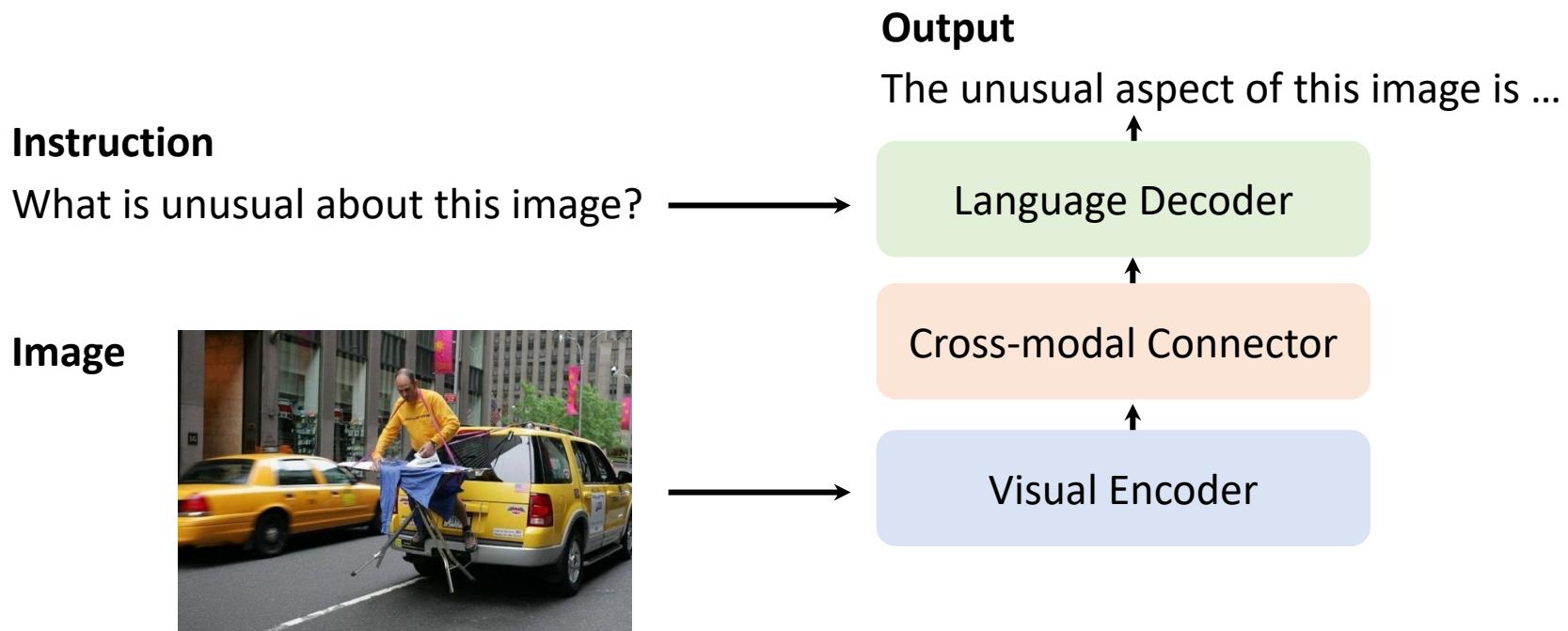
Source: <https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg>

How to Build Generalist Multimodal Models?



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

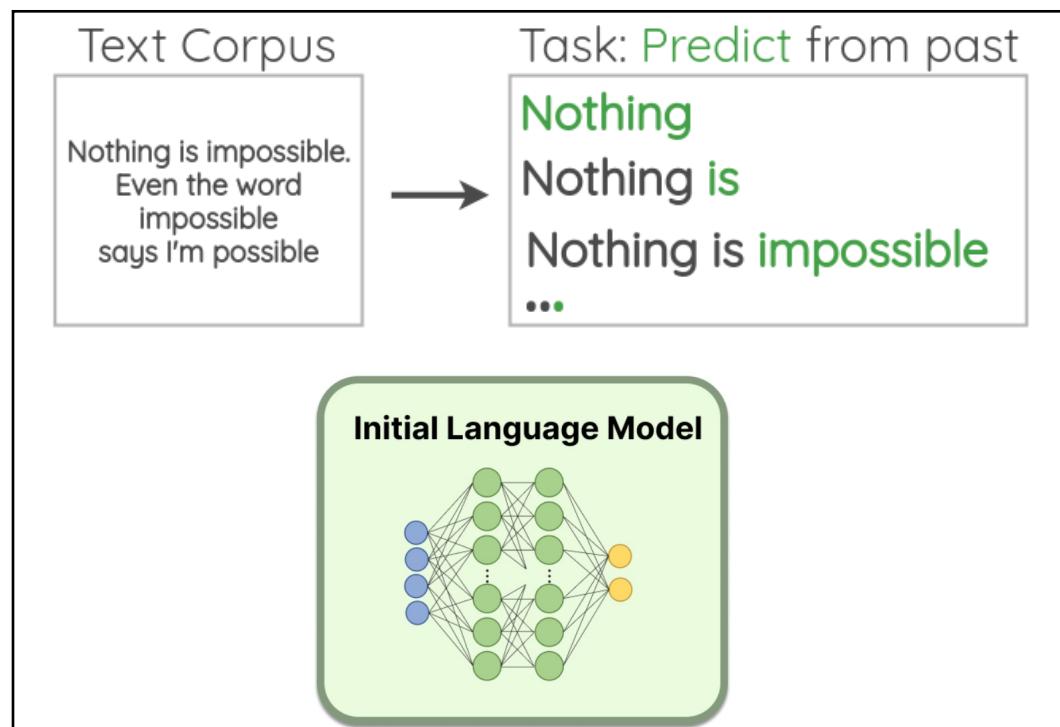
How to Build Generalist Multimodal Models?



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

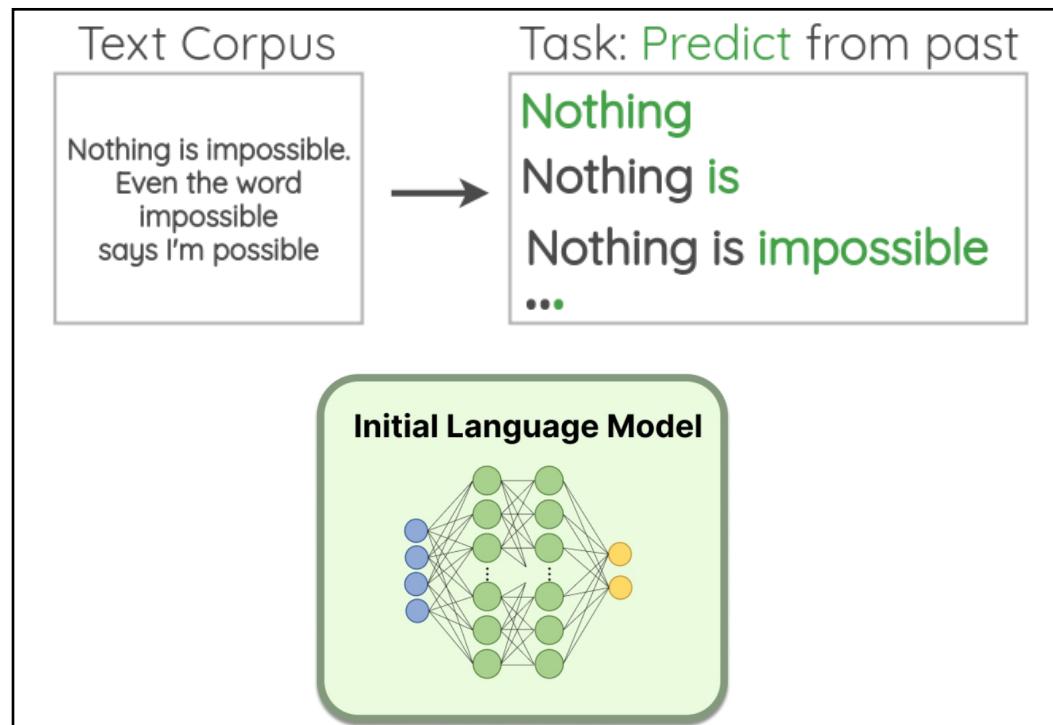
Lessons from Training Large Language Models

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

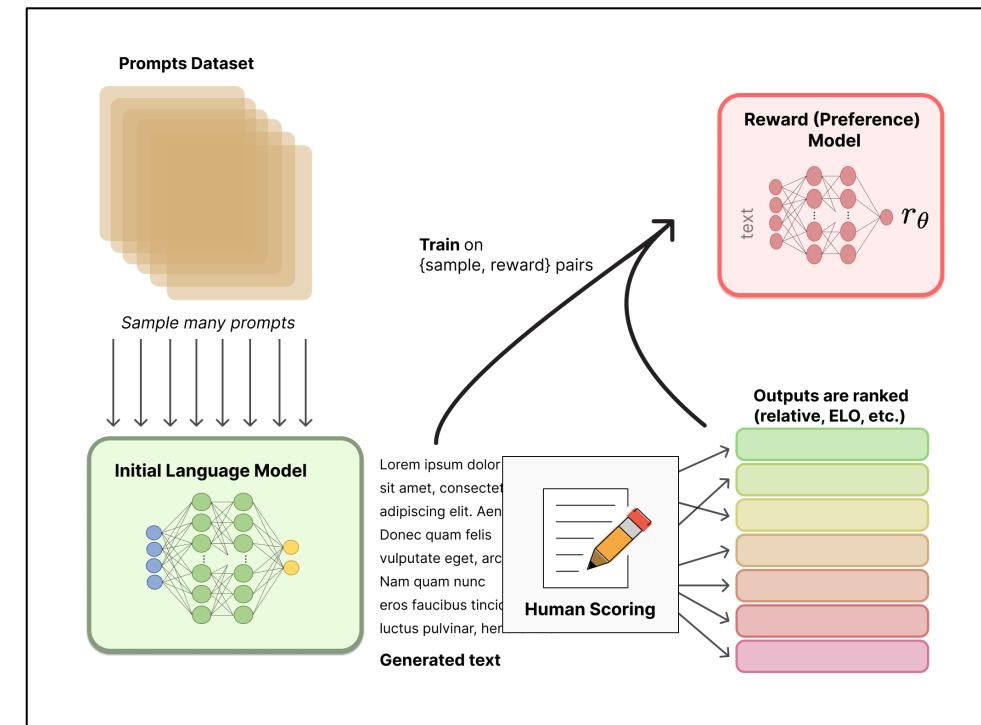


Lessons from Training Large Language Models

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



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



Lessons from Instruction Tuning in LLMs

Instruction

Explain human's behavior.
Behavior: cry.

Recommend a movie for me.

...

Output

There could be many reasons why a person might cry. They could be feeling sad, scared, angry, or frustrated...

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: ...

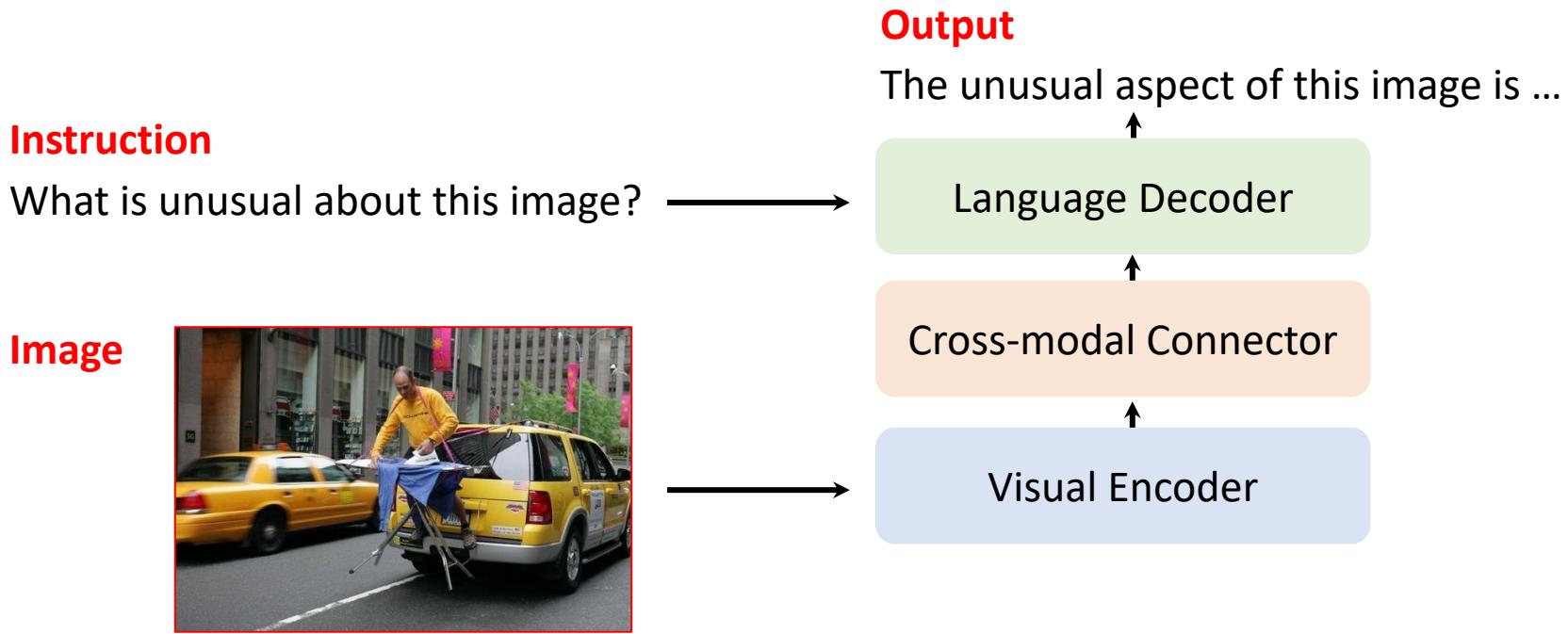
<new instruction 1>

<new output 1>

...

LLaMa → Alpaca, Vicuna, etc.

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

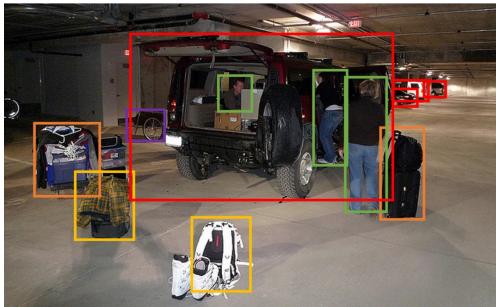
Image



Context (caption)

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

Context (layout)



- *How do we get a text-only LLM to “see” an image?*
- Let it read context information written in text

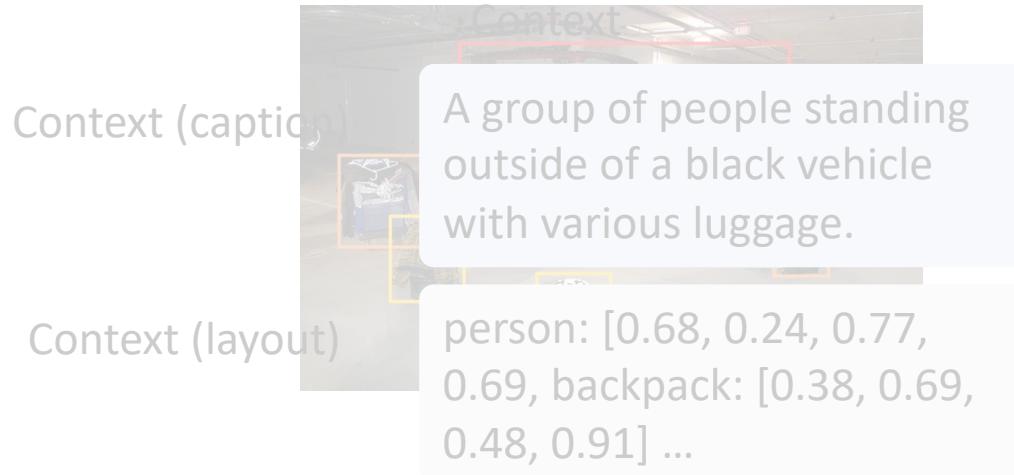
MS-COCO

Each image associated with:

- 5 captions
- Object categories / bounding boxes

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],
....

Text-only GPT Assisted Visual Instruction Data Creation



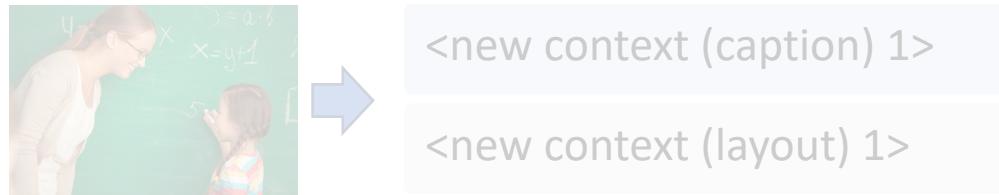
Instruction

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.



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

Text-only ChatGPT



<new instruction 1>

<new output 1>

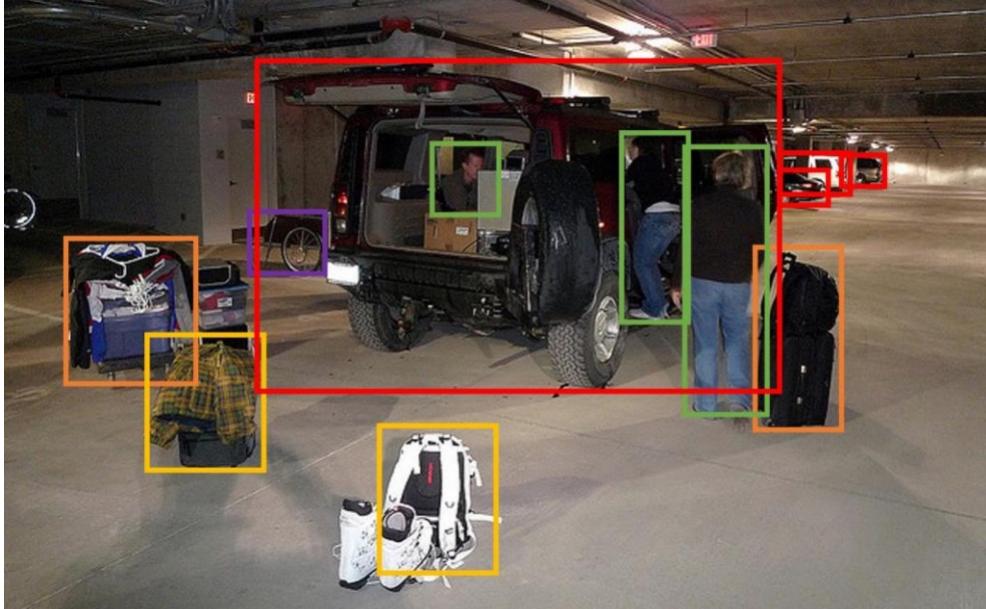
Visual Instruction-following Data:
Triplet (image, instruction, output)



What are the two people doing?

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

Text-only GPT Assisted Visual Instruction Data Creation



LLaVA-Instruct-158K

Conversation: 58K

Detailed description: 23K

Complex reasoning: 77K

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.

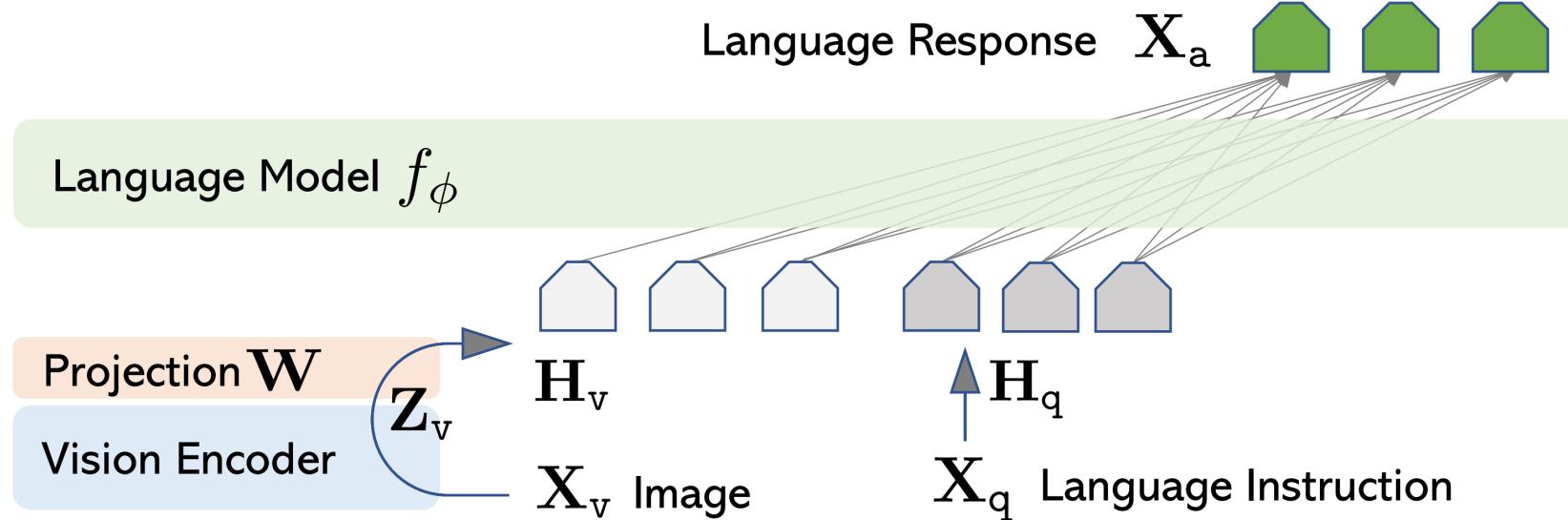
Other cars are parked around the main SUV, with one car positioned behind it and slightly to the left, another behind and slightly to the right, and the third car further behind on the right side.

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: Large Language-and-Vision Assistant

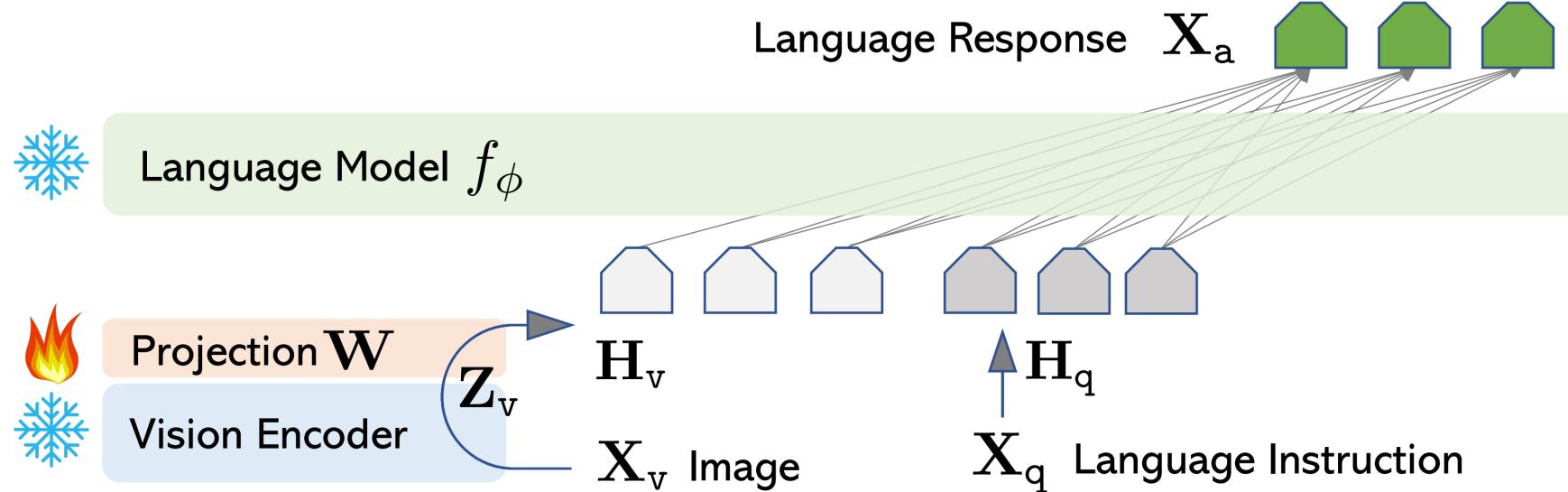


Vision Encoder: CLIP-ViT-L/14

Projection: Linear layer (MLP in LLaVA-1.5)

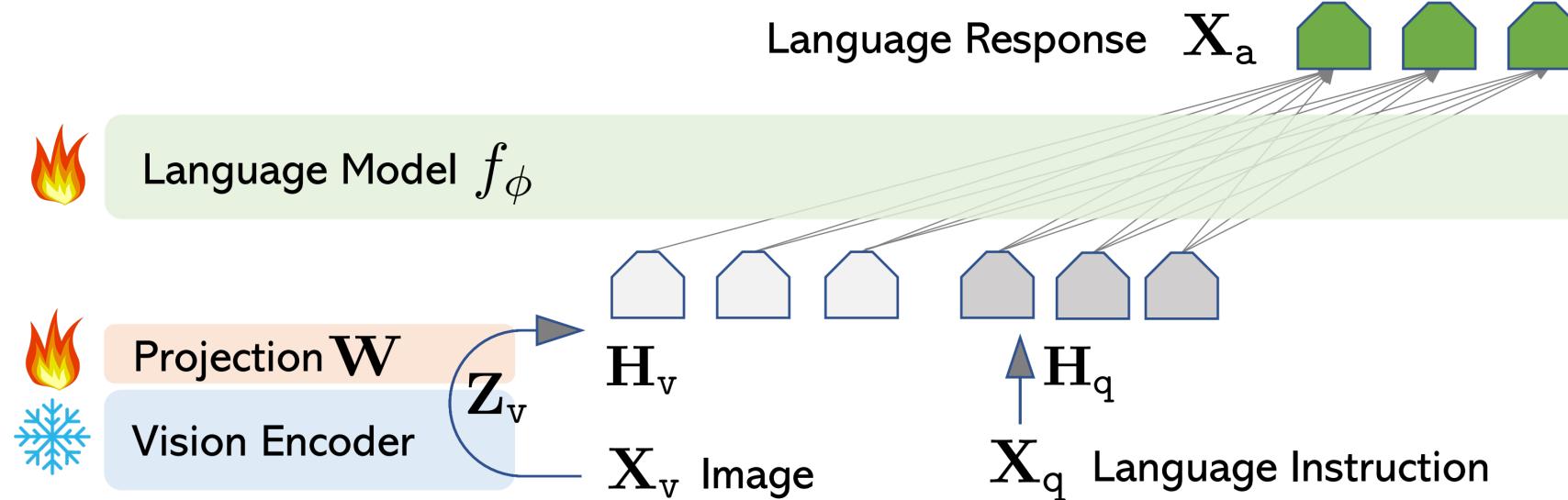
Language Model: Vicuna, LLaMA-2-Chat, MPT-Chat, etc.

Stage 1: Pre-training for Feature Alignment



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

Stage 2: End-to-end Visual Instruction Tuning



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

Visual input example, Extreme Ironing:



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



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*



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.

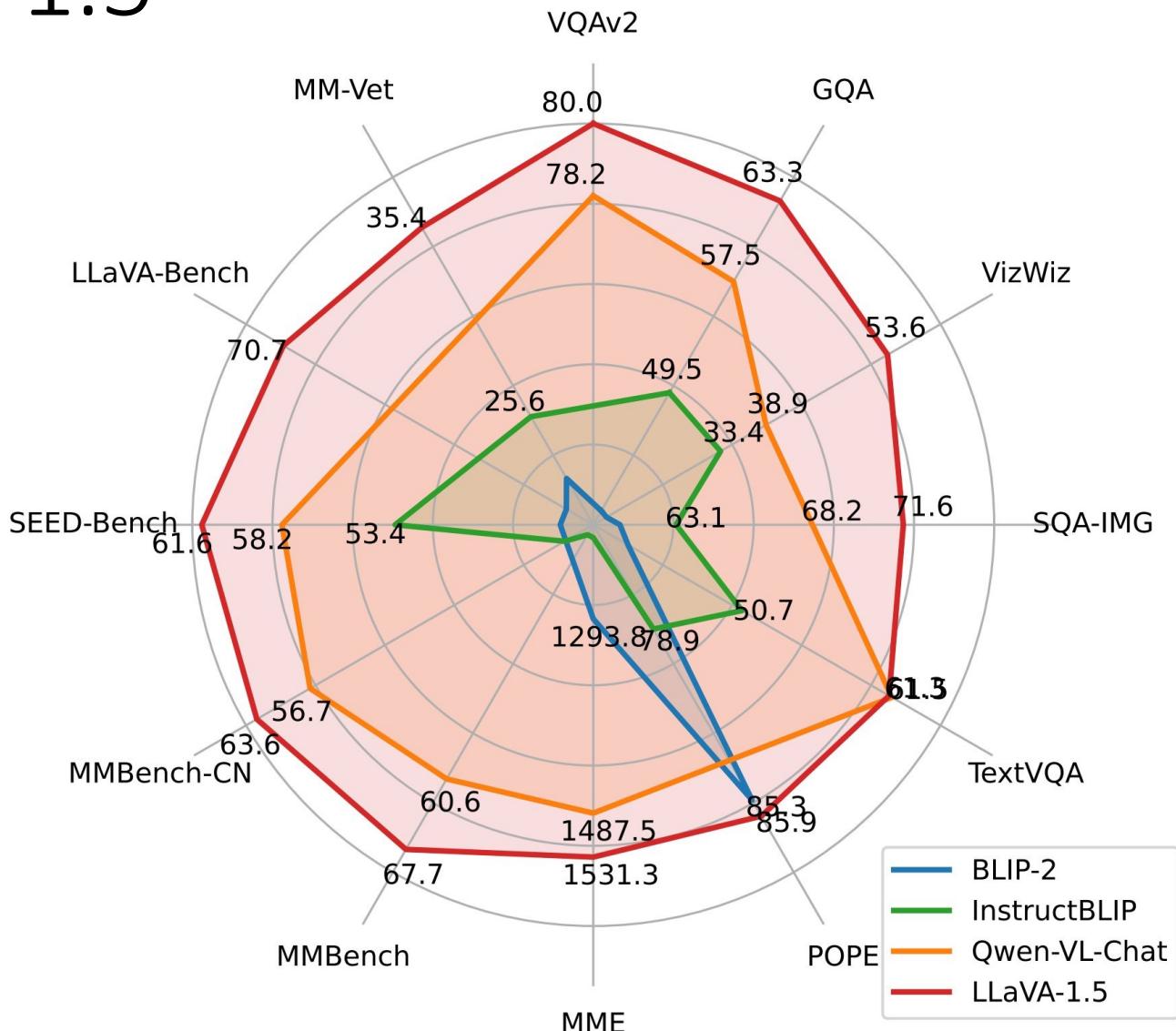
Strong Visual Reasoning Capability

Example: *Parodied Mona Lisa*



Extensions: LLaVA-1.5

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



Extensions: LLaVA-NeXT (1.6)

- Significantly outperforms LLaVA-1.5-13B
- Matched Gemini Pro (in Jan 2024) 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

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*



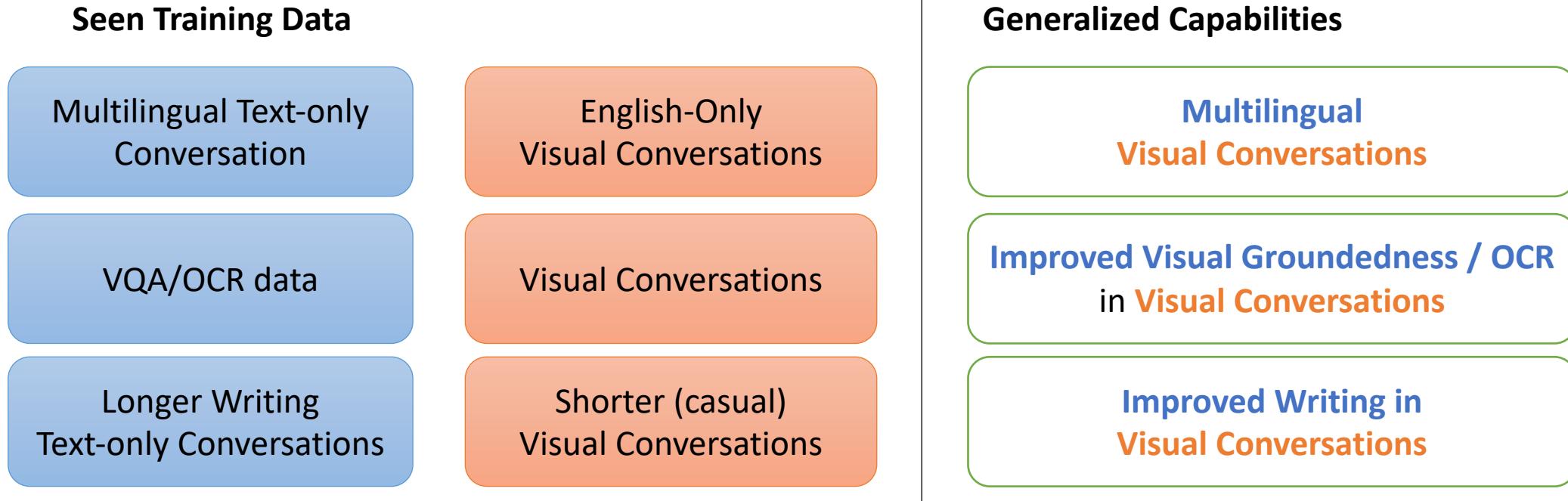
这个区域的名字叫什么？请简单介绍

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

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。

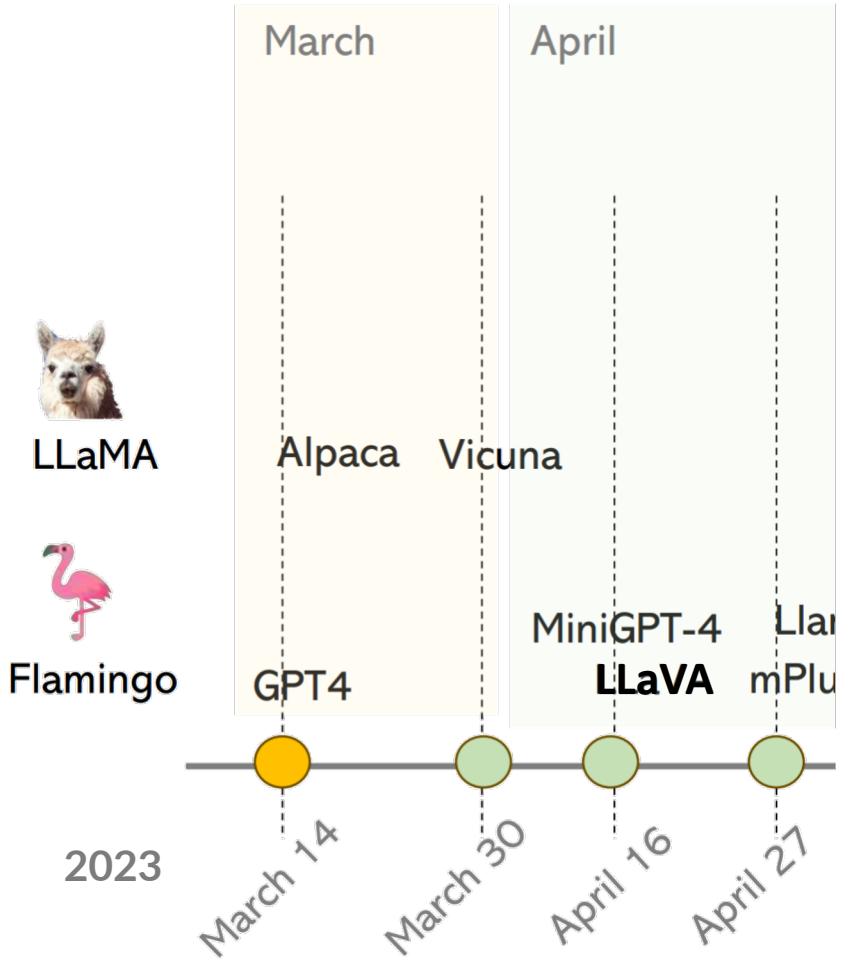


Combinatorial Task Generalization

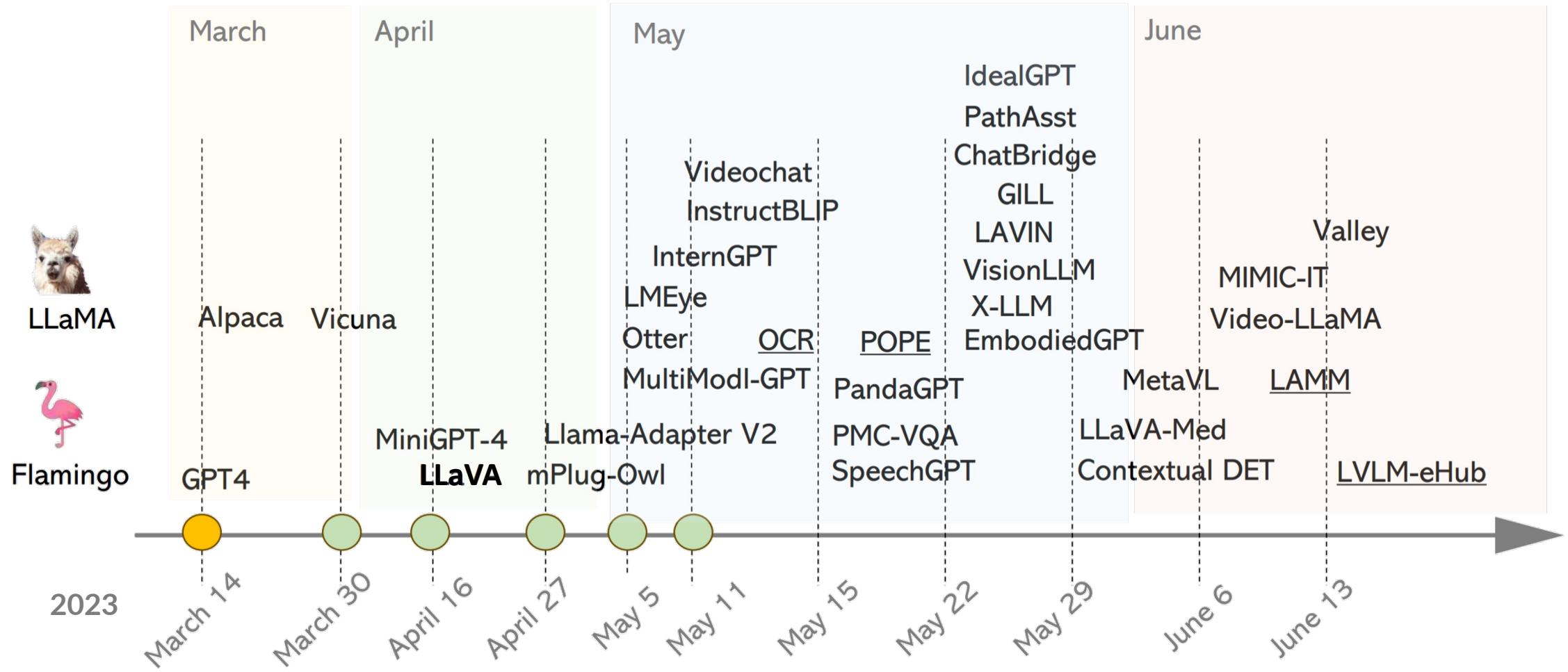


Do not need to create all combinations of data in training; let LMMs generalize!

Community Efforts on LMMS



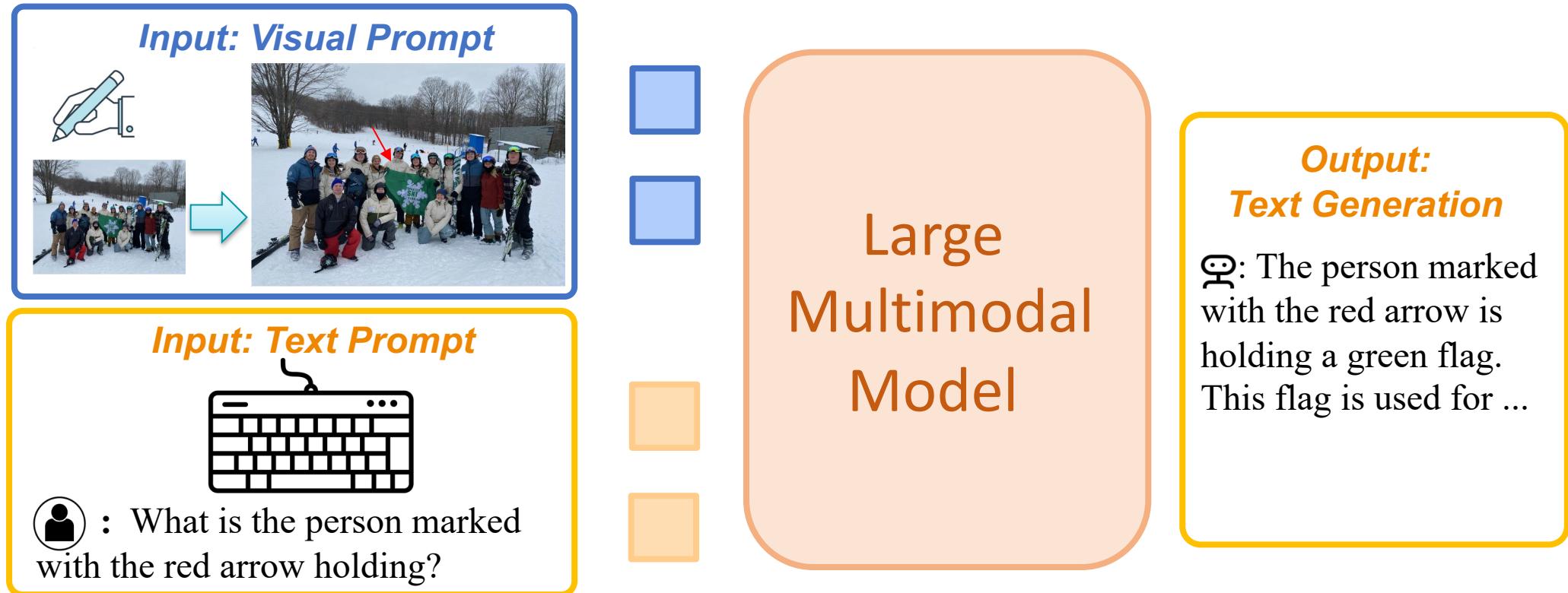
Community Efforts on LMMS



LLaVA that Understands Visual Prompts?

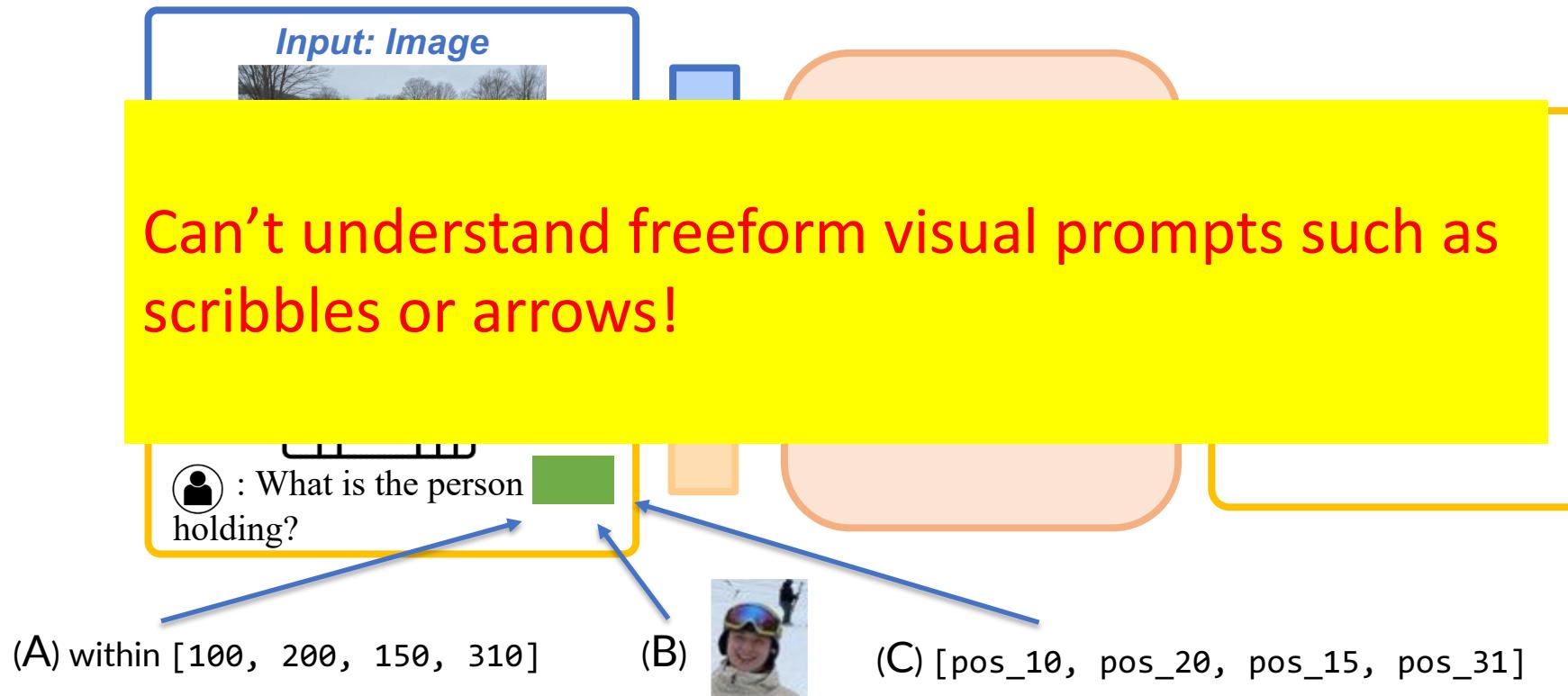
- Existing Large Multimodal Models (LMMs) do a good job in *whole* image understanding
- Can we make them also understand *regions / visual prompts* (e.g., arrows, scribbles)?

ViP-LLaVA: Understanding Freeform Visual Prompts

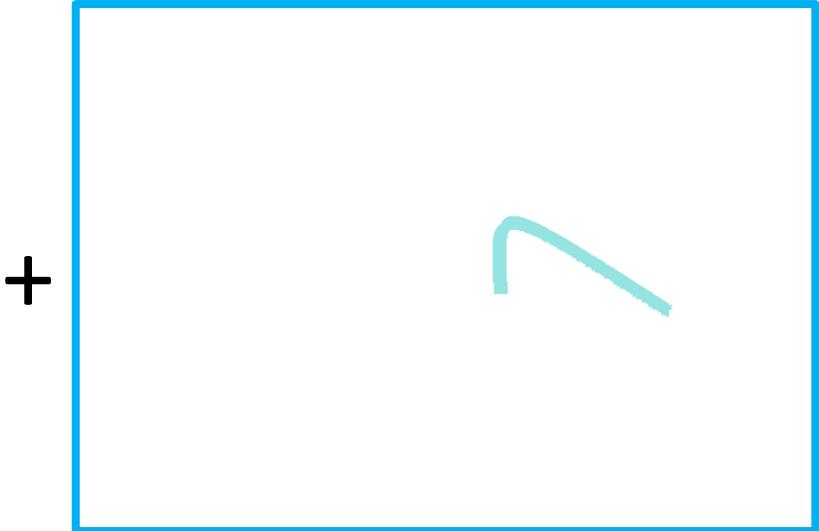


Prior/Concurrent Work

- Text to denote box coordinates e.g. Shikra, MiniGPT-v2, Ferret
- Region of Interest feature e.g. GPT4ROI
- Learned vocabularies or positional embedding e.g. GPT4ROI, Kosmos-2



ViP-LLaVA: Understanding Freeform Visual Prompts



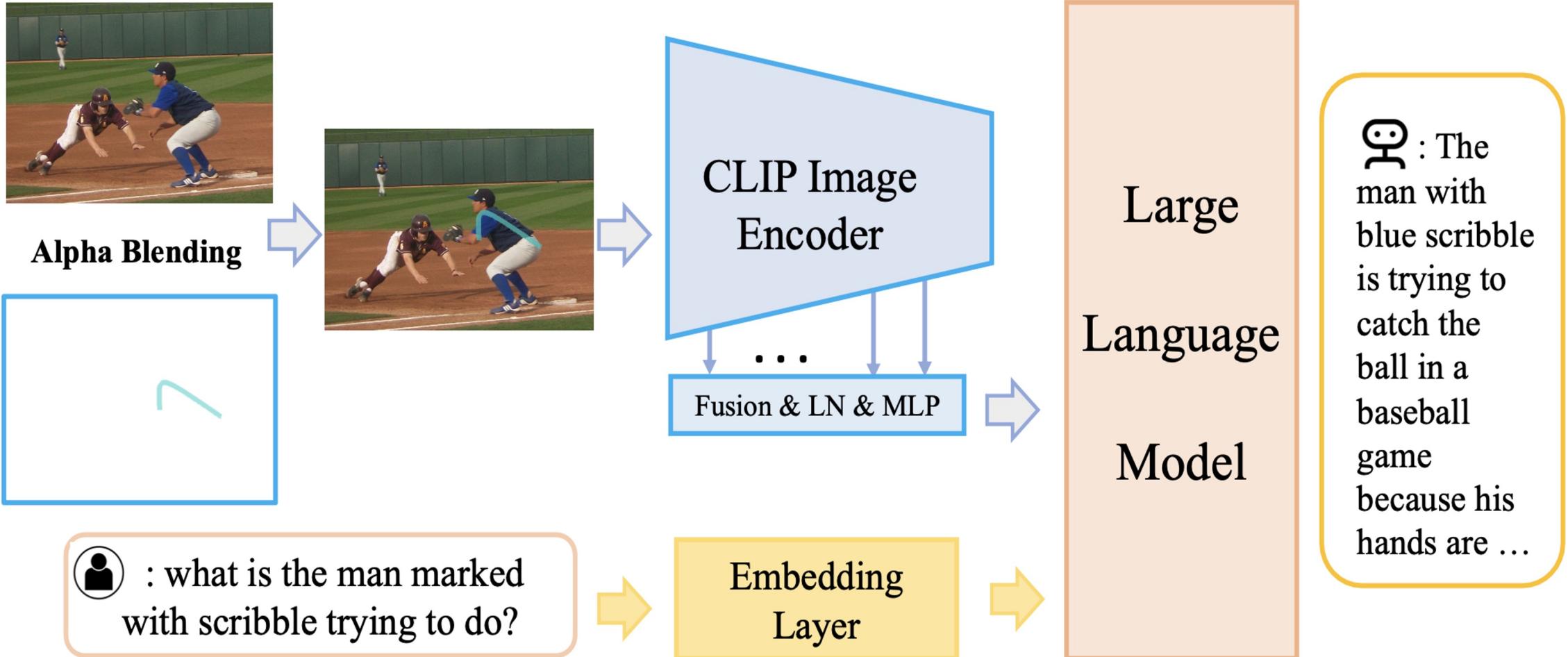
Separate Images

Overlay

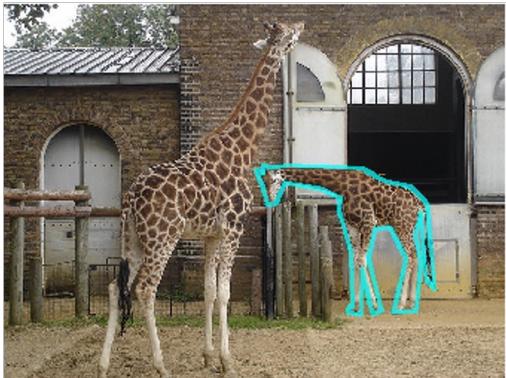
- Super simple!
- A portion of CLIP pretraining data has visual prompt annotations

[Shtedritski et al. "What does CLIP know about a red circle? Visual prompt engineering for VLMs." ICCV 2023]

ViP-LLaVA: Architecture



Which Visual Prompts Can ViP-LLaVA Support?



Mask contour



Ellipse



Bounding box



Triangle



Scribble



Point



Arrow

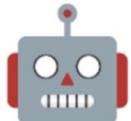


Mask

- Leverage region level visual recognition and reasoning datasets
(RefCOCOg, Visual7W, PointQA-LookTwice, VCR) - 520k images



Among the regions with red, blue, and green circles, which two share the same content?



The two regions with the same content are the red and blue circles. Both circles contain a train.

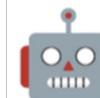
Describe the instance pointed to by the red arrow.



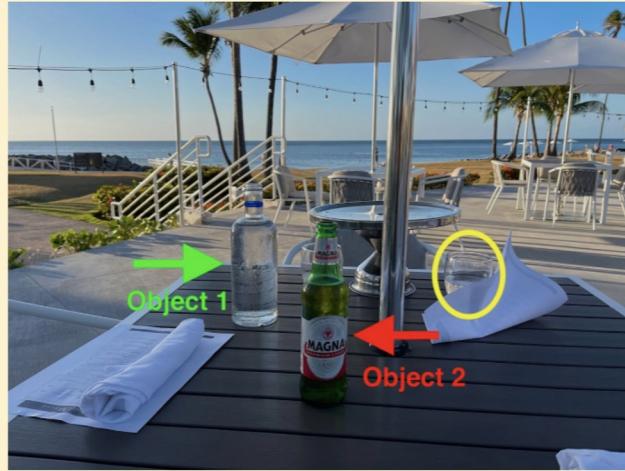
Describe the instance pointed to by the red arrow.



A person in a blue jacket and black pants is standing on skis in the snow.



A woman in a white jacket and white pants is standing on skis in the snow.



What is contained in the circled glass?



The glass within the ellipse appears to contain a clear liquid, which is likely water.



Describe what is Object 1 and object 2 .



Object 1 is a bottle of water, and Object 2 is a bottle of beer.



What is in the circled glass? The stuff within the circle should belong to Object 1 or 2 ?



The stuff within the circle is the liquid from Object 1, which is water.

Quantitative Evaluation

**Table 1: Object recognition
in Visual7W**

Method	Accuracy (%)
12in1	83.35
GPT4ROI-7B	81.83
GPT4ROI-13B	84.82
Shikra-13B	85.33
Ours-13B	87.91

**Table 2: Object counting
in PointQA-LookTwice**

Method	Accuracy (%)
Point and ask	60.20
LLaVA-1.5-7B	56.19 [†]
LLaVA-1.5-13B	57.93 [†]
Shikra-13B	70.30
Ours-13B	71.77

**Table 3: Visual Reasoning
in VCR**

Model	$Q \rightarrow AR$ (%)
ViLBERT	54.0
Unicoder-VL	54.5
VLBERT-L	58.9
GPT4RoI-7B	78.6
Ours-7B	78.93

**Table 4: Region Captioning
in RefCOCOg**

Model	METEOR	CIDEr
GRIT	15.2	71.6
Kosmos-2	14.1	62.3
GLaMM	16.2	105.0
Ours-7B	16.2	105.9

- SoTA results on recognition, counting, captioning, and commonsense reasoning tasks

Visual Prompt Understanding Benchmark



Q: Which one has the lowest price, the fruit within the red rectangle, within the blue rectangle, or within the green rectangle? Respond with the name of the fruit.

GT: Orange.

Required capabilities: Recognition, OCR, Math

- 303 questions:
Recognition, OCR
(Optical Character
Recognition),
Knowledge, Math,
Object relationship
reasoning, Language
generation

Yo'LLaVA: Your Personalized LMM



Can you see if <thao> is in this photo?

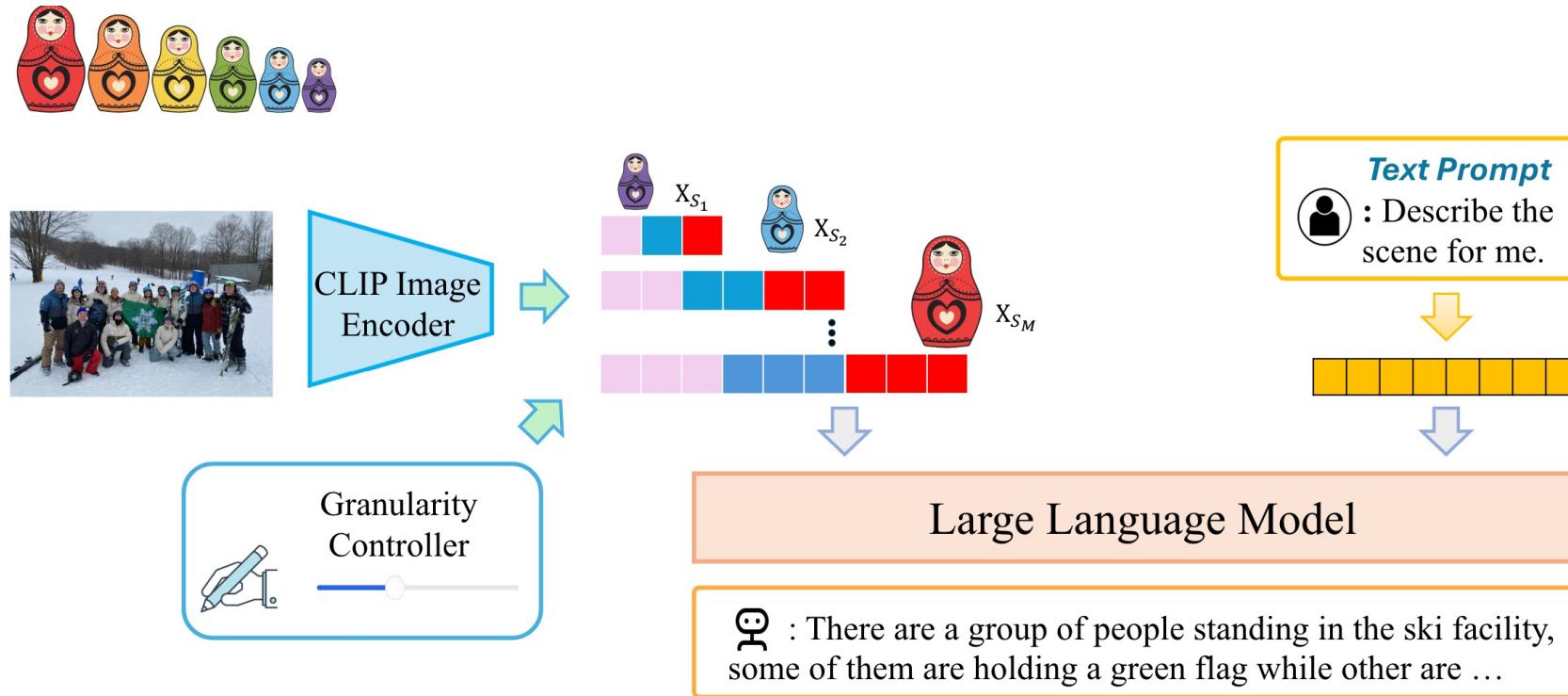
<thao> is in the photo, and she is standing next to another person. They both appear to be having a good time, possibly at a party or event. <thao> is smiling and making a peace sign with her hand, which suggests a positive and fun atmosphere.



I'm **Yo'LLaVA**
Your Personalized
LLaVA!

Matryosha Multimodal Models

- Adaptively control visual token length (i.e. information granularity) in LLaVA



Inspired by:
Matryosha Representation Learning

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



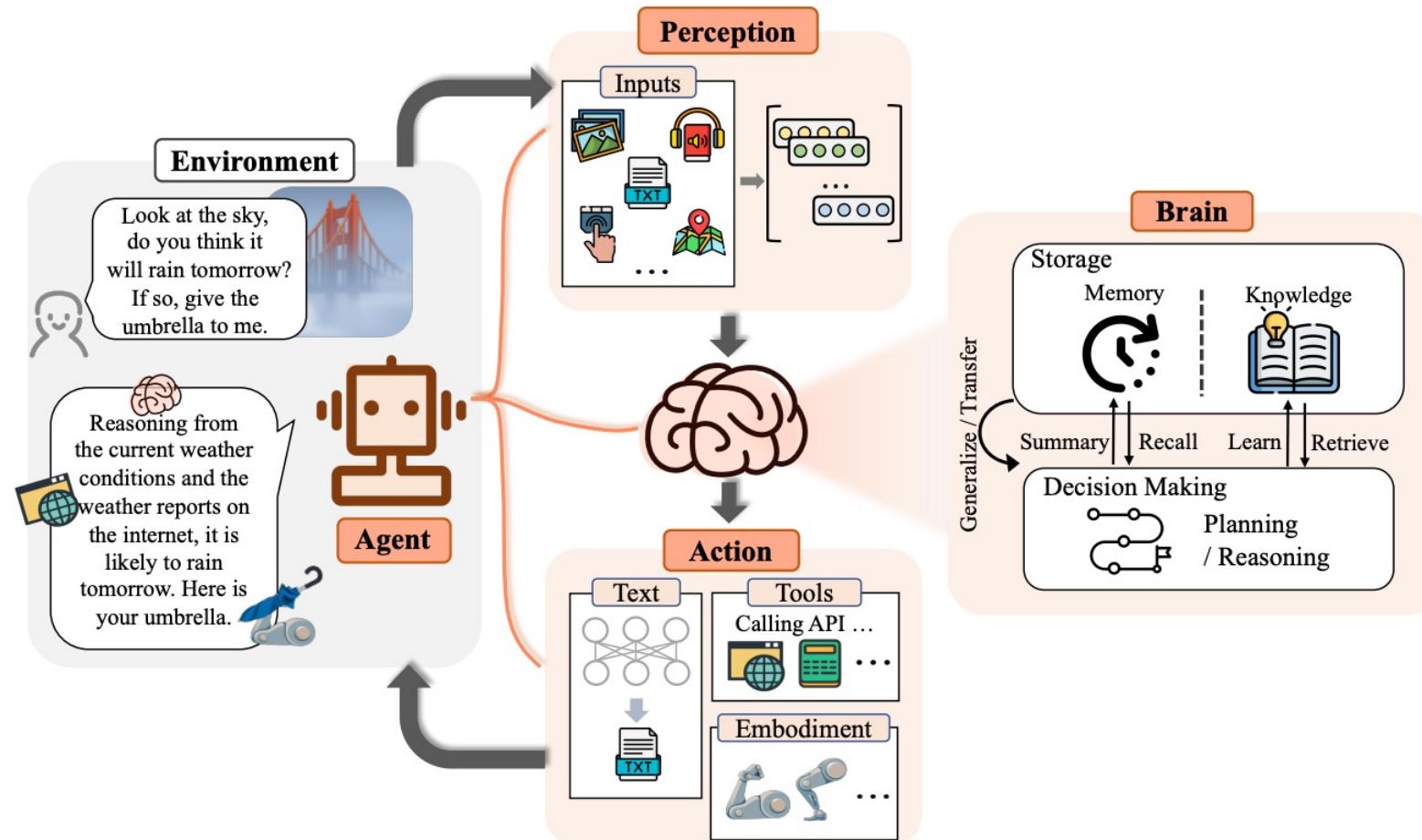
: The image shows an ice cream cone with a blue scoop of ice cream on top and what appears to be a pink scoop on the bottom. Therefore, Caption 1 is correct.

Check out: “CounterCurate”,
J. Zhang*, M. Cai*, et al. ACL
Findings 2024.

Looking Forward: Limitations of Current Models

- Capabilities
 - Hallucinations
 - Video understanding
 - Smaller performant models
 - Multimodal agents
 - ...
- 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: Multimodal AI Agents



- 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

Visual Prompting for Robot Learning

Google DeepMind

2024-2-13

PIVOT: Iterative Visual Prompting Elicits Actionable Knowledge for VLMs

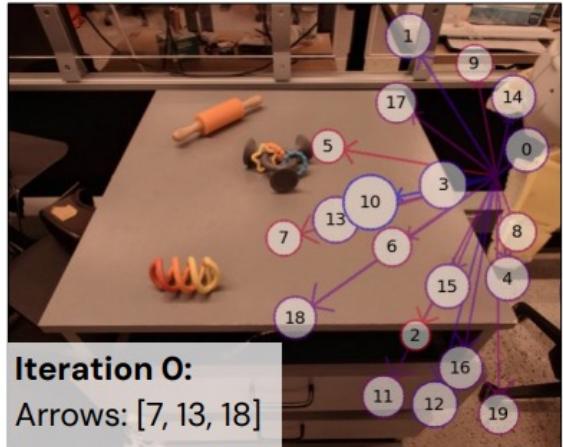
Soroush Nasiriany^{*,†,1,3}, Fei Xia^{*,1}, Wenhao Yu^{*,1}, Ted Xiao^{*,1}, Jacky Liang¹, Ishita Dasgupta¹, Annie Xie²,
Danny Driess¹, Ayzaan Wahid¹, Zhuo Xu¹, Quan Vuong¹, Tingnan Zhang¹, Tsang-Wei Edward Lee¹,
Kuang-Huei Lee¹, Peng Xu¹, Sean Kirmani¹, Yuke Zhu³, Andy Zeng¹, Karol Hausman¹, Nicolas Heess¹,
Chelsea Finn¹, Sergey Levine¹, Brian Ichter^{*,1}

¹Google DeepMind, ²Stanford University, ³The University of Texas at Austin

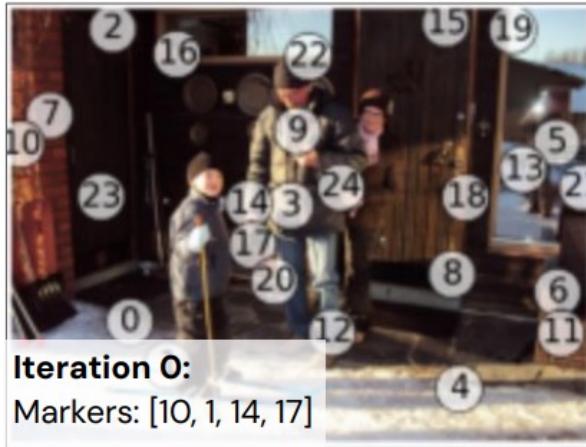
Correspond to: {soroushn, xiafei, magicmelon, tedxiao, ichter}@google.com

Website: pivot-prompt.github.io and HuggingFace: <https://huggingface.co/spaces/pivot-prompt/pivot-prompt-demo>

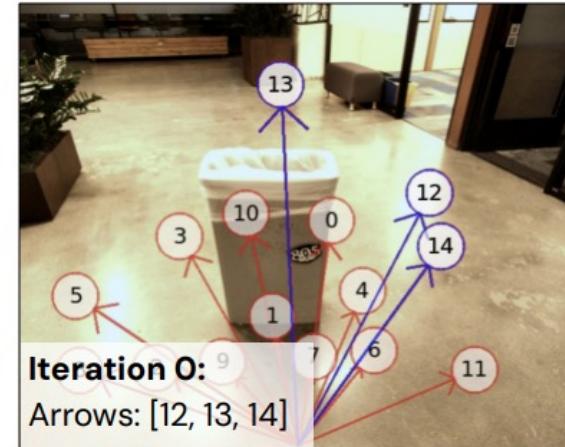
Task: What actions should the robot take to pick up the DNA chew toy?



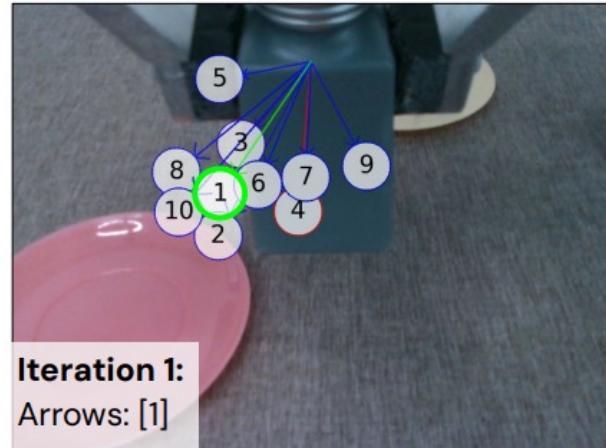
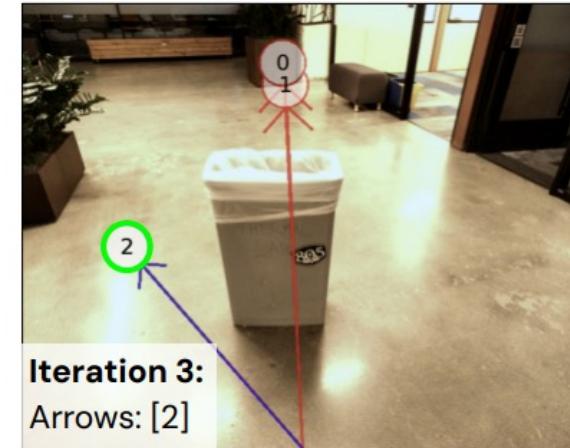
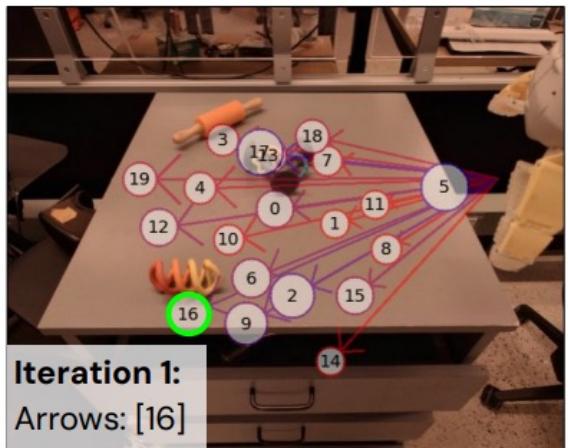
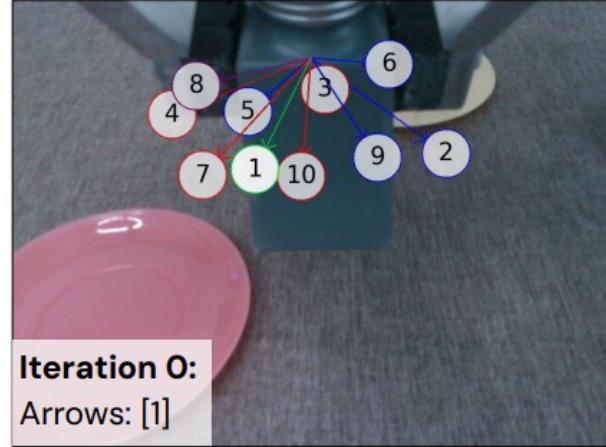
Task: What numbers overlay the "L kid"?



Task: What actions should the robot take to go to wooden bench without hitting the obstacle?



Task: What actions should the robot take to put the pepper shaker on the pink plate?



Where to throw this away?



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://vip-llava.github.io/>

Thank you

- Haotian Liu, Yuheng Li, Utkarsh Ojha, Mu Cai, Thao Nguyen, Chunyuan Li, Jianwei Yang, Jianfeng Gao , Siva Mustikovela, Greg Meyer, Yuning Chai, Dennis Park



Haotian Liu



Yuheng Li



Utkarsh Ojha



Mu Cai



Thao Nguyen