

Building Steerable 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)

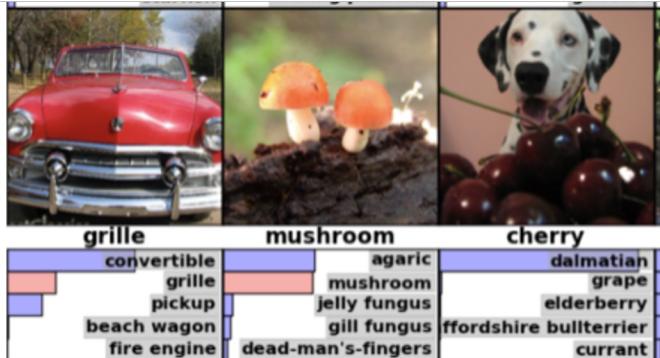
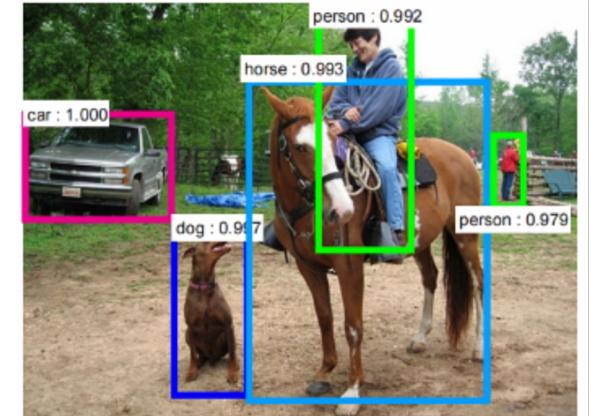


Image classification



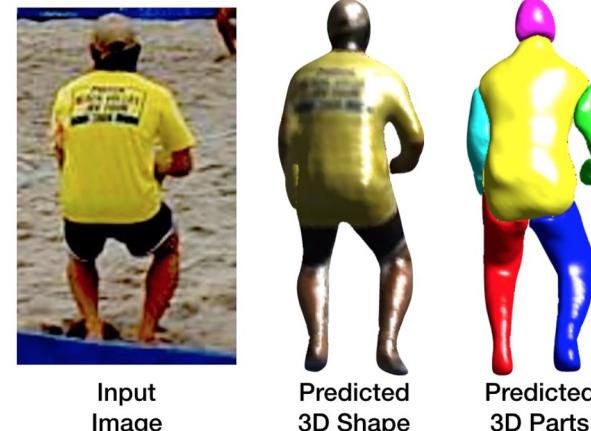
Semantic segmentation



Object detection



Pose recognition



3D prediction

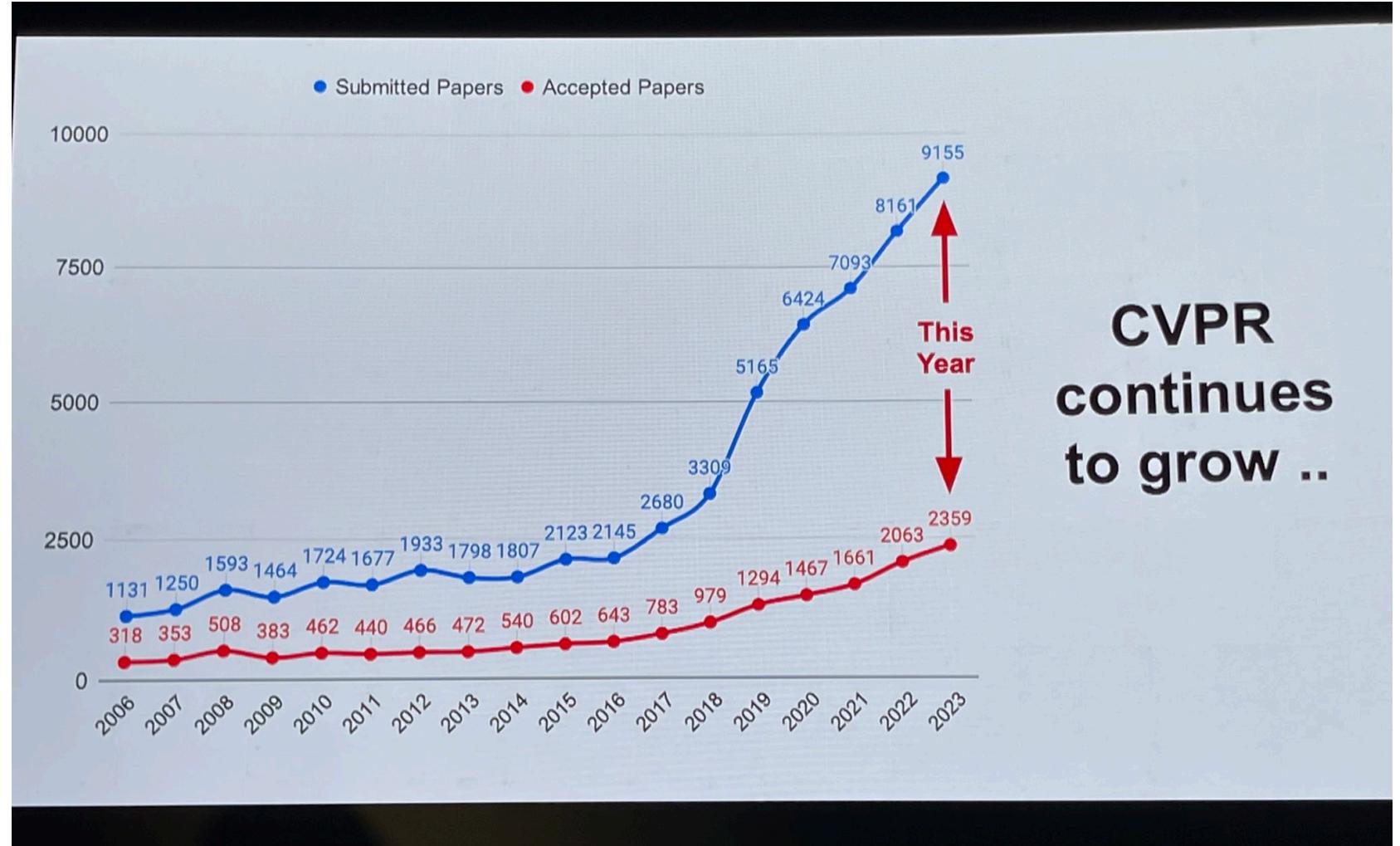


Surface normal prediction

... and many more

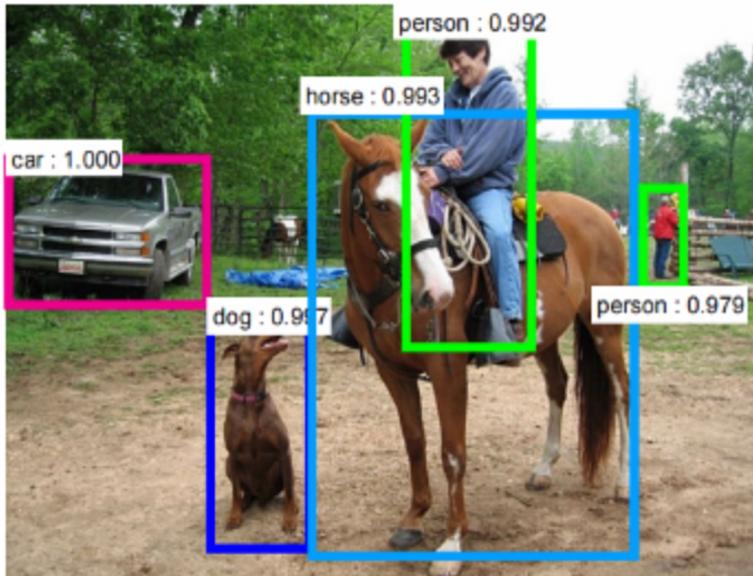
Explosion in ...

Students!
Startups!
Funding!
Hiring!



However, Prevailing Paradigm Thus Far:

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



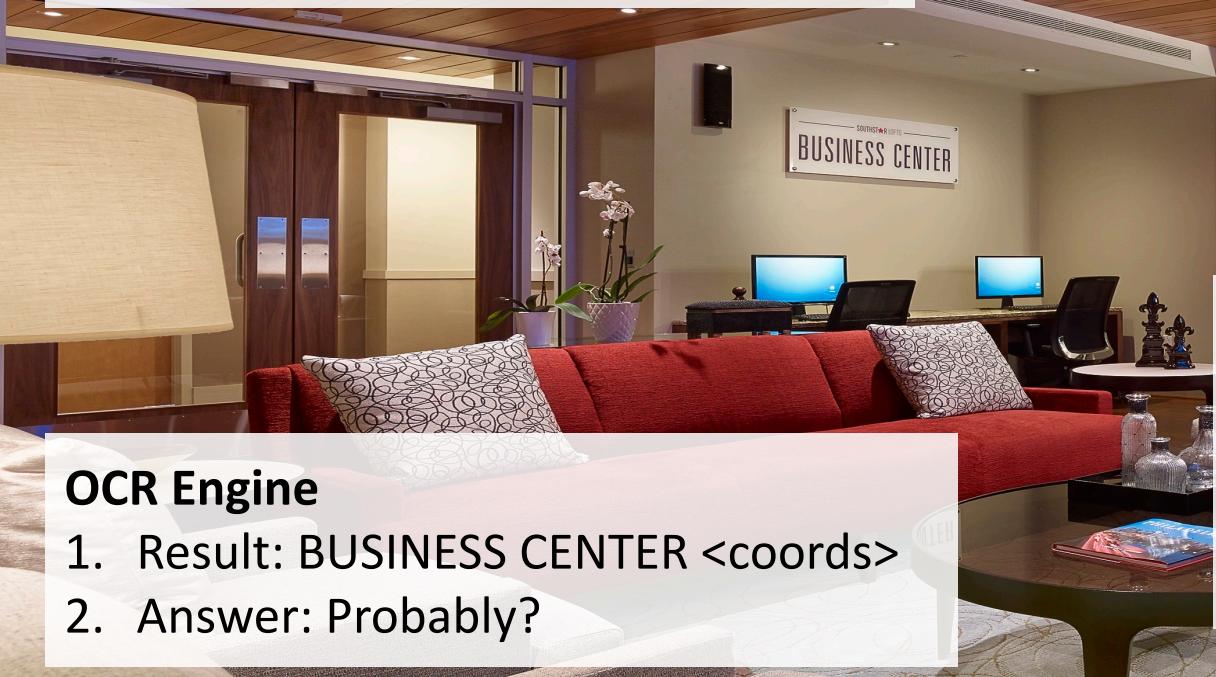
Object Detection *Only*



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?

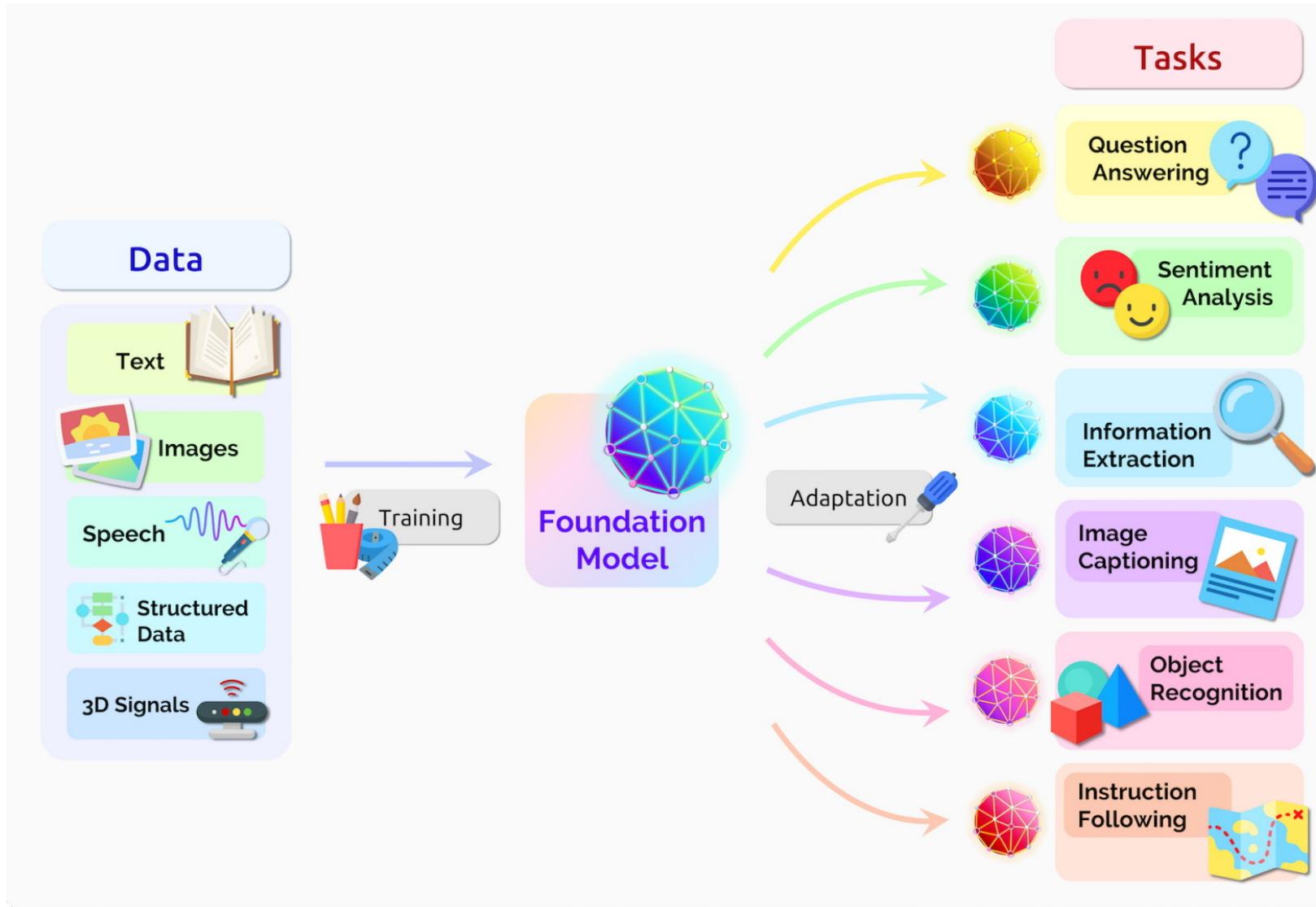
Final output to the user



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

Specialist models are insufficient
for open-ended generic tasks

Rise of “Generalist” Foundation Models (2020s)

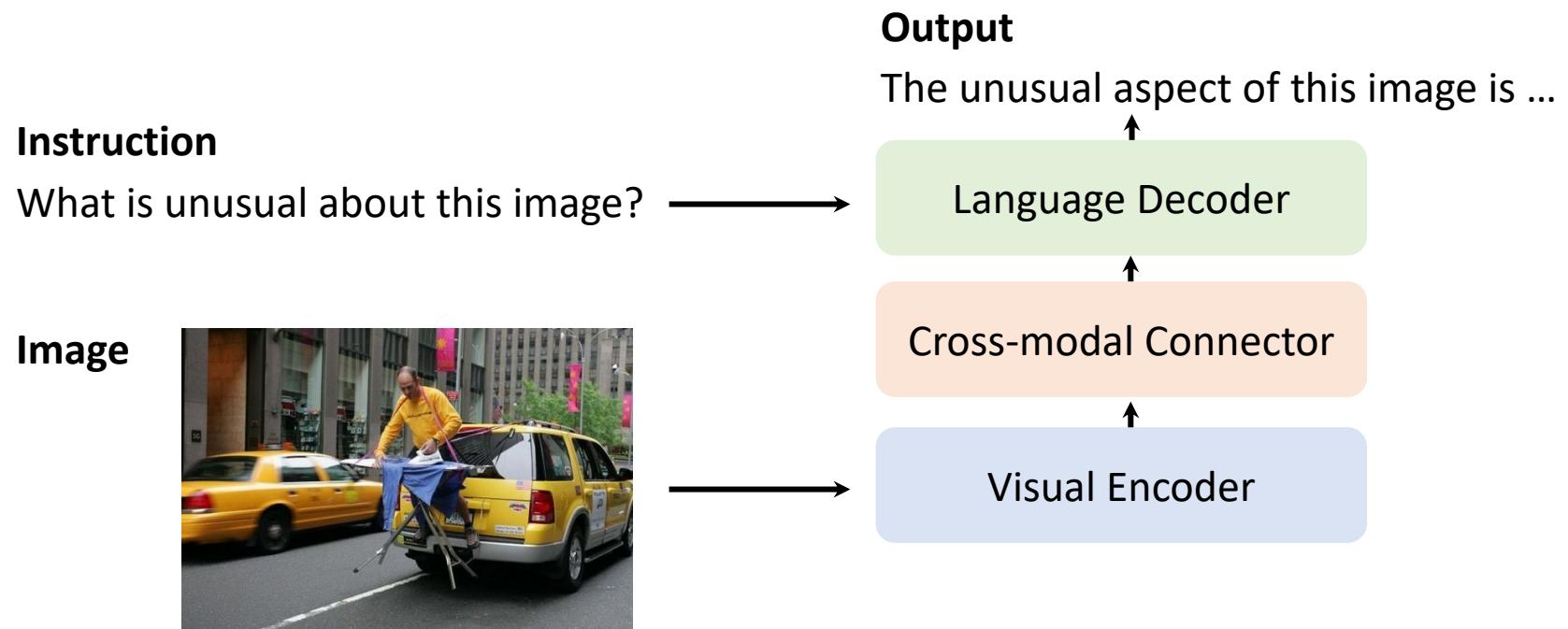


- **Single-model, many tasks**
- Large Language Models (e.g., GPT4)
- Vision Transformers
- Image-Text Models (e.g., CLIP)

Today's talk: Large Multimodal Generalist Models

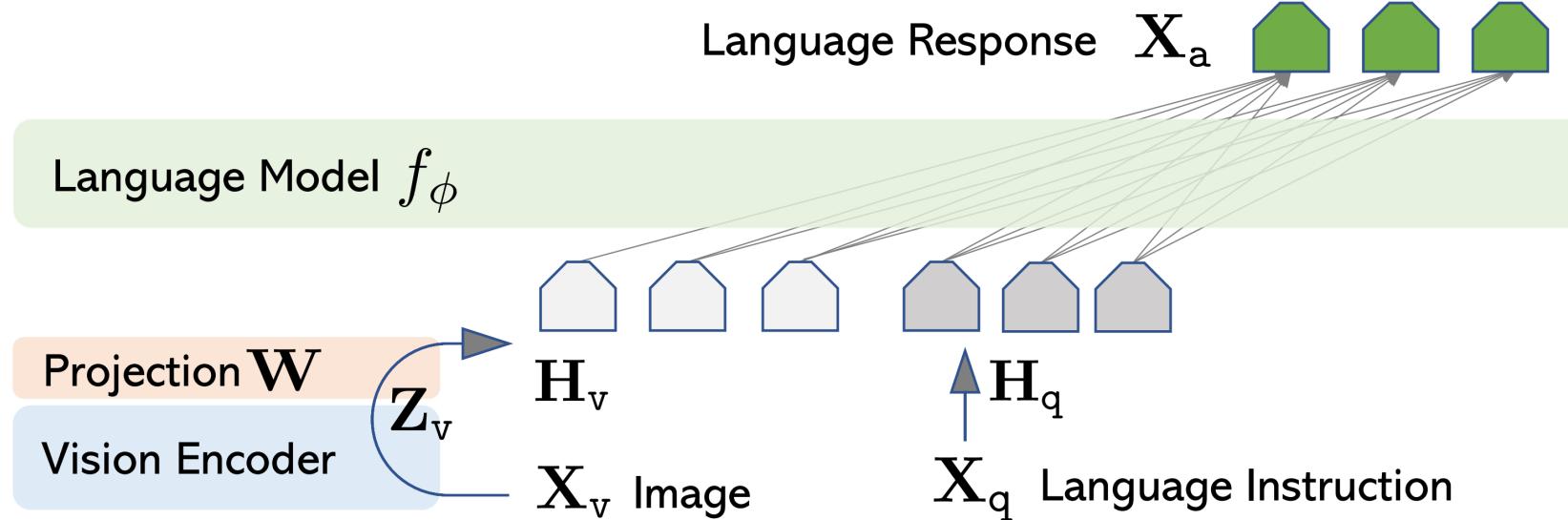
- *Generalist* vision-language models that understand visual data and communicate in natural language
- *Steerable* (“aligned”) models that produce desirable outputs for wide-concept knowledge
- Understand visual prompts; personalization; adaptive token lengths

LLaVA: Large Language-and-Vision Assistant



- Large multimodal model that (i) understands visual data, (ii) follows human instructions, (iii) communicates in natural 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.

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*



Emergent Multilingual Capability

Example: *French Quarter*



这个区域的名字叫什么？请简单介绍

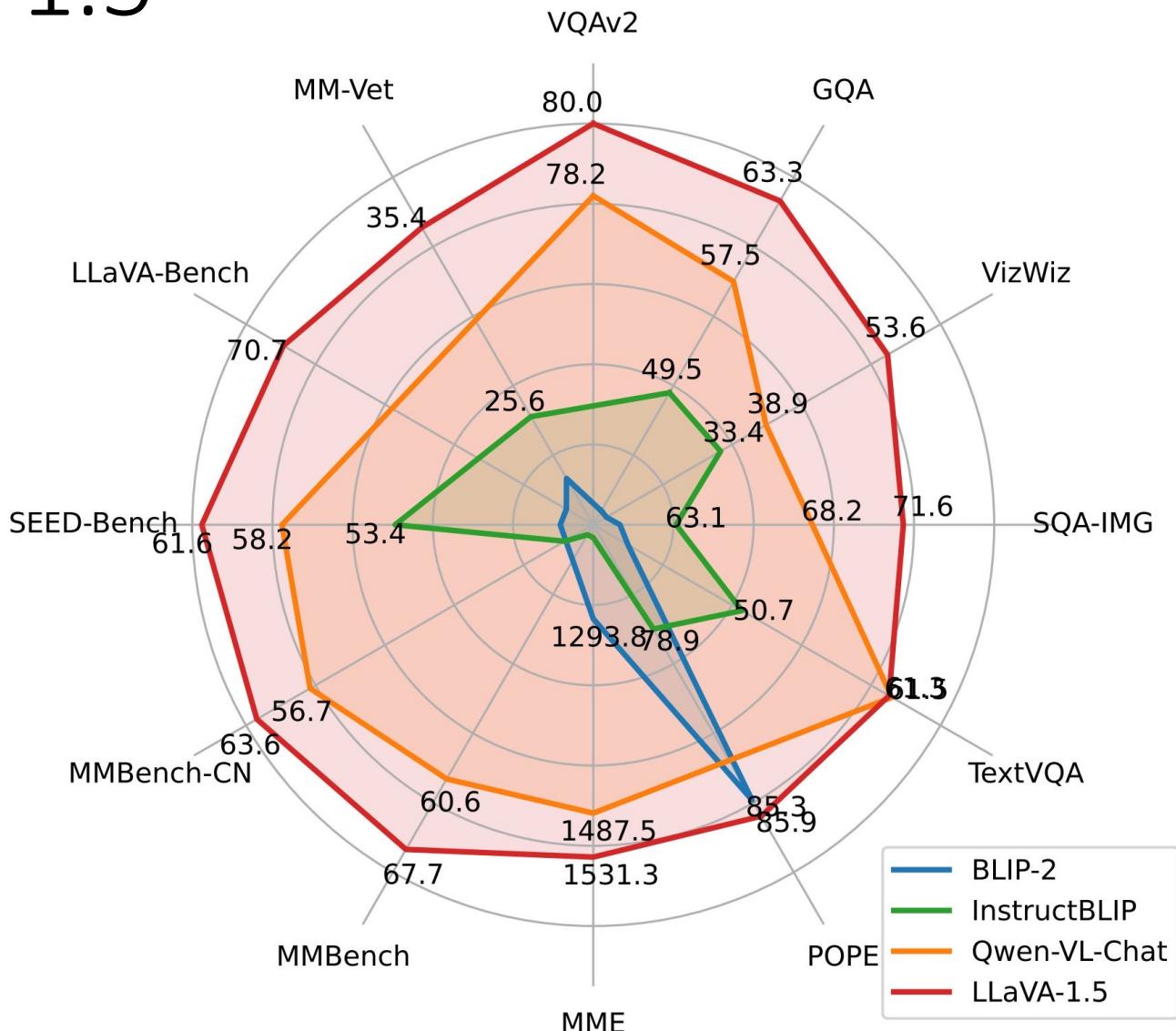
Translation: What is the name of this area? Please describe briefly.

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。



Extensions: LLaVA-1.5

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

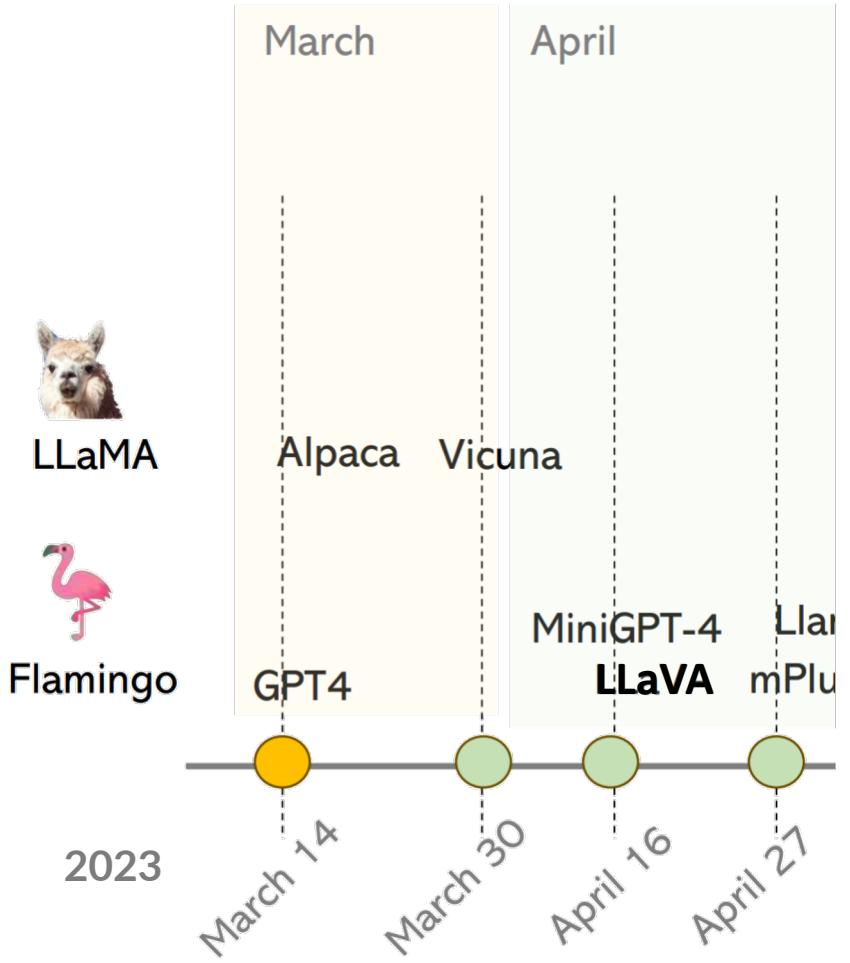


Extensions: LLaVA-NeXT (1.6)

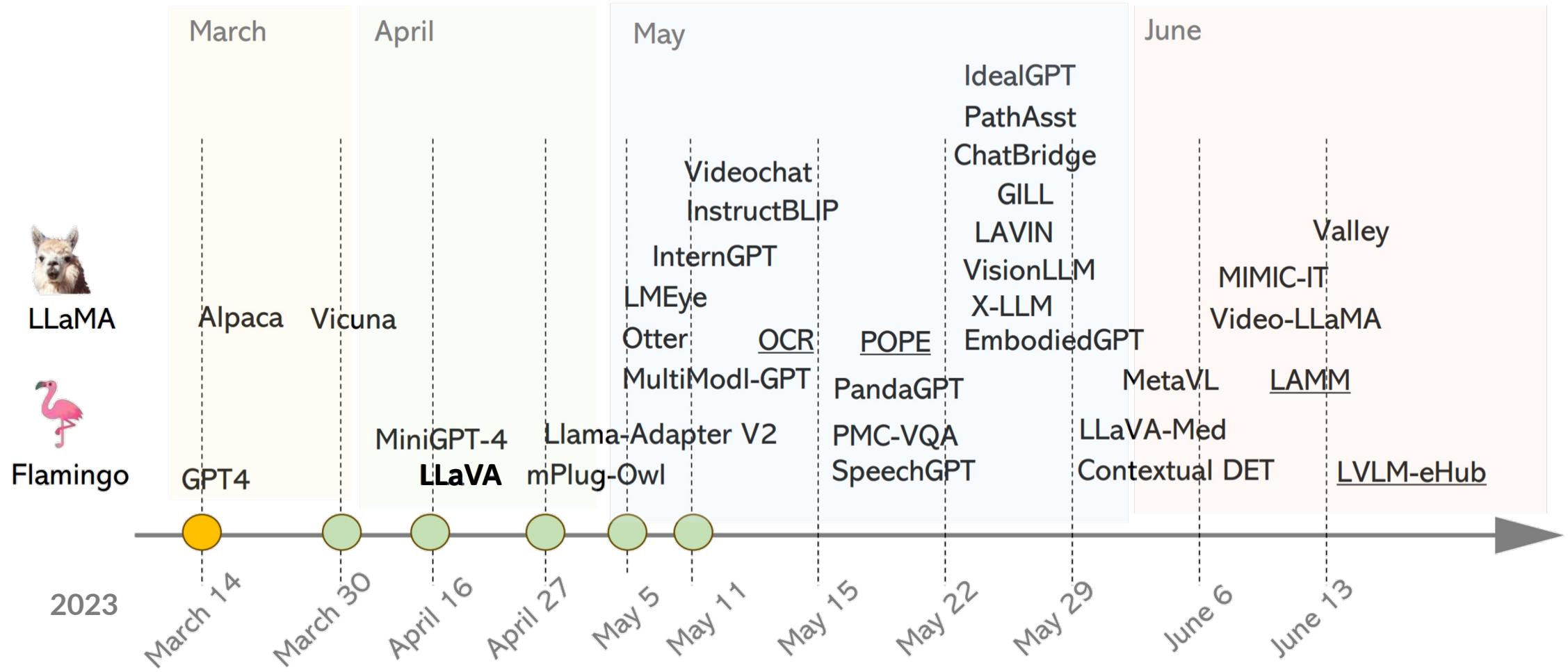
- Significantly outperforms LLaVA-1.5-13B
- Matched Gemini Pro (in Jan 2024) on selected benchmarks

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

Community Efforts on LMMS



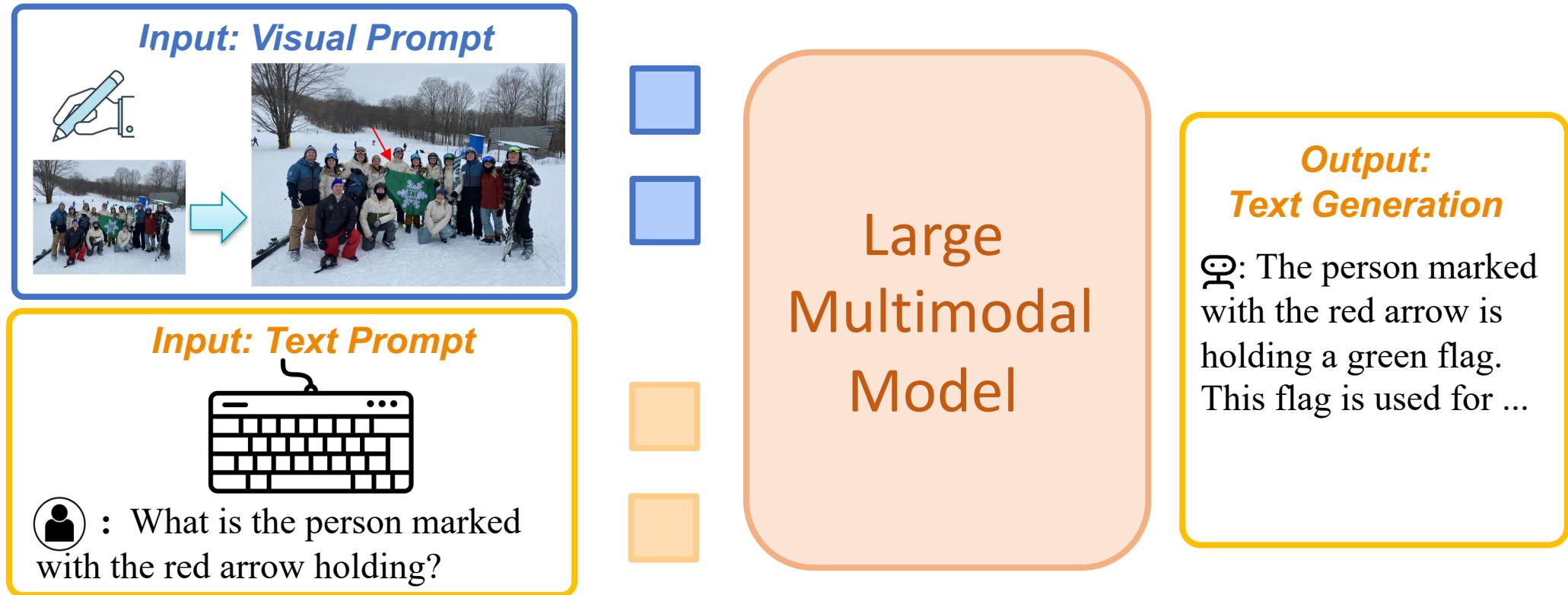
Community Efforts on LMMS



LLaVA that Understands Visual Prompts?

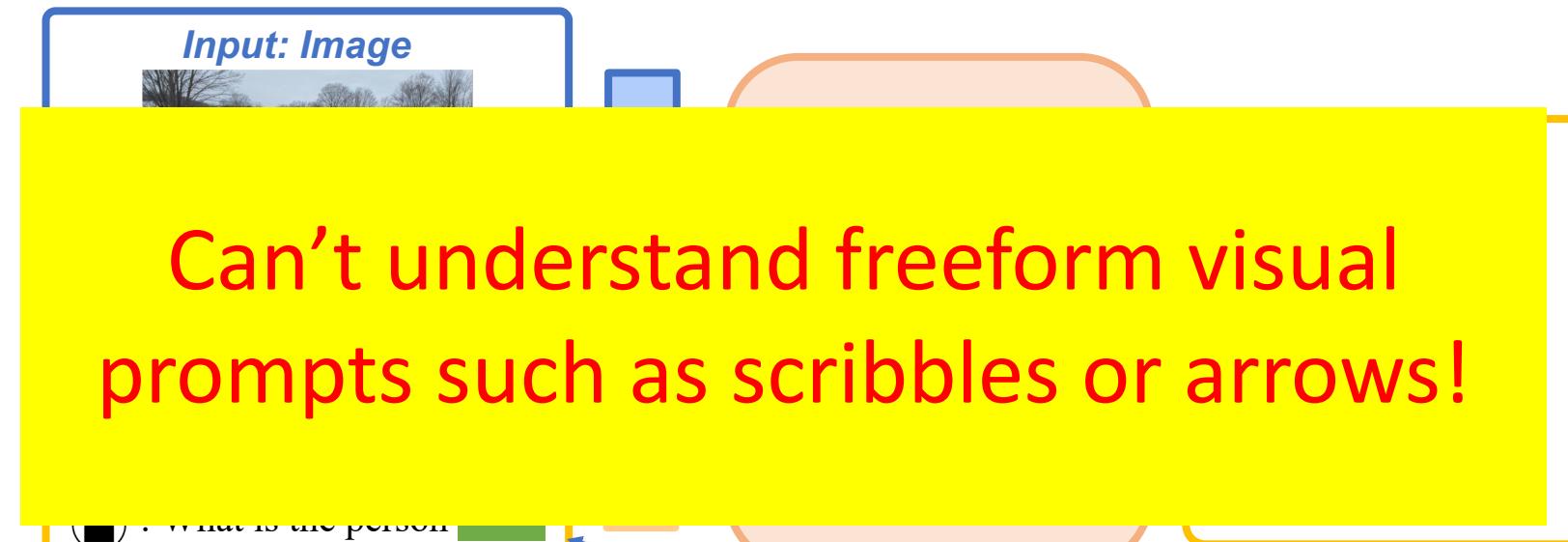
- Existing Large Multimodal Models (LMMs) do a good job in *whole* image understanding
- Can we make them also understand *regions / visual prompts* (e.g., arrows, scribbles)?
- Describing complex visual concepts with text is hard

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



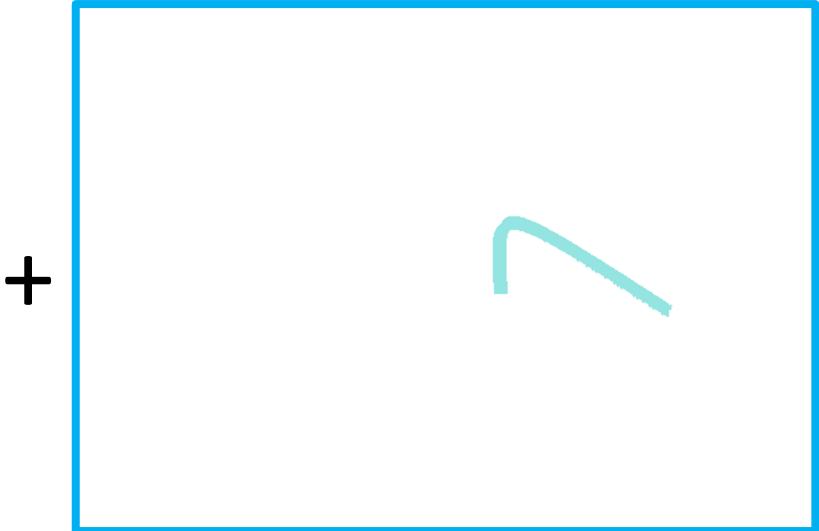
(A) within [100, 200, 150, 310]

(B)



(C) [pos_10, pos_20, pos_15, pos_31]

ViP-LLaVA: Understanding Freeform Visual Prompts



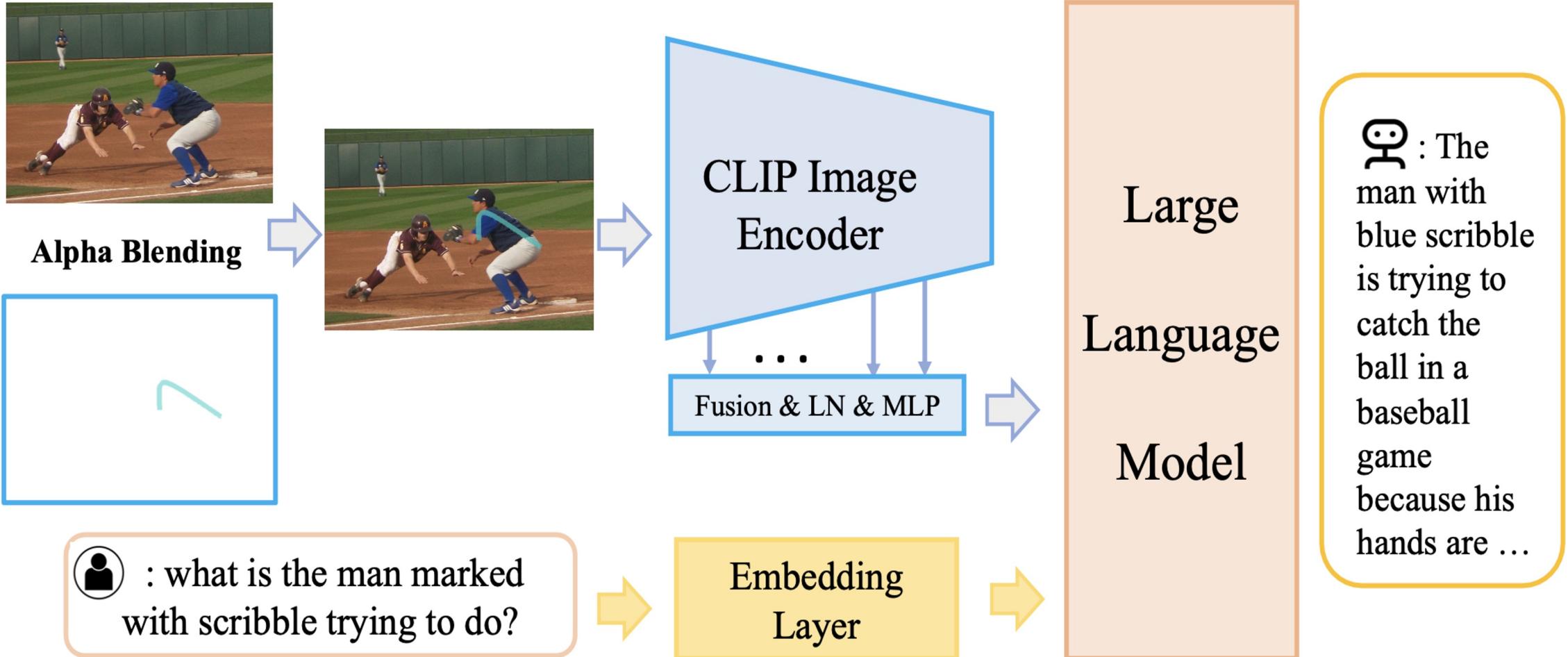
Separate Images

Overlay

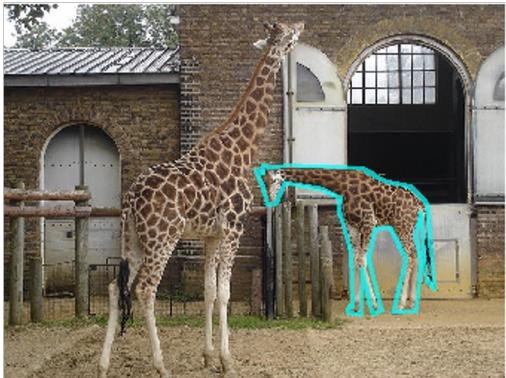
- Super simple!
- A portion of CLIP pretraining data has visual prompt annotations

[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

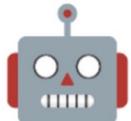


Mask

- 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.

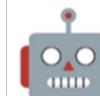
Describe the instance pointed to by the red arrow.



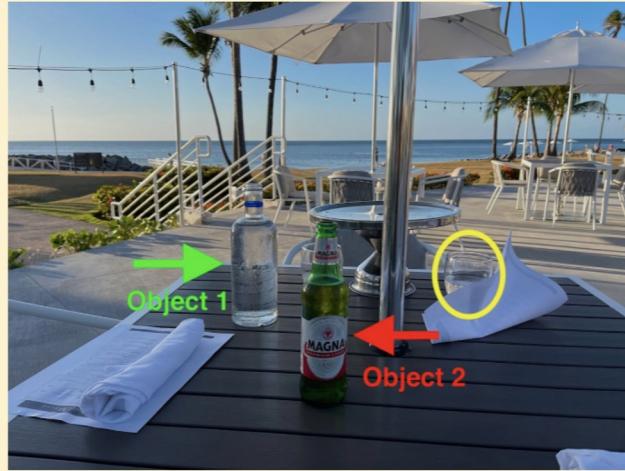
Describe the instance pointed to by the red arrow.



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**

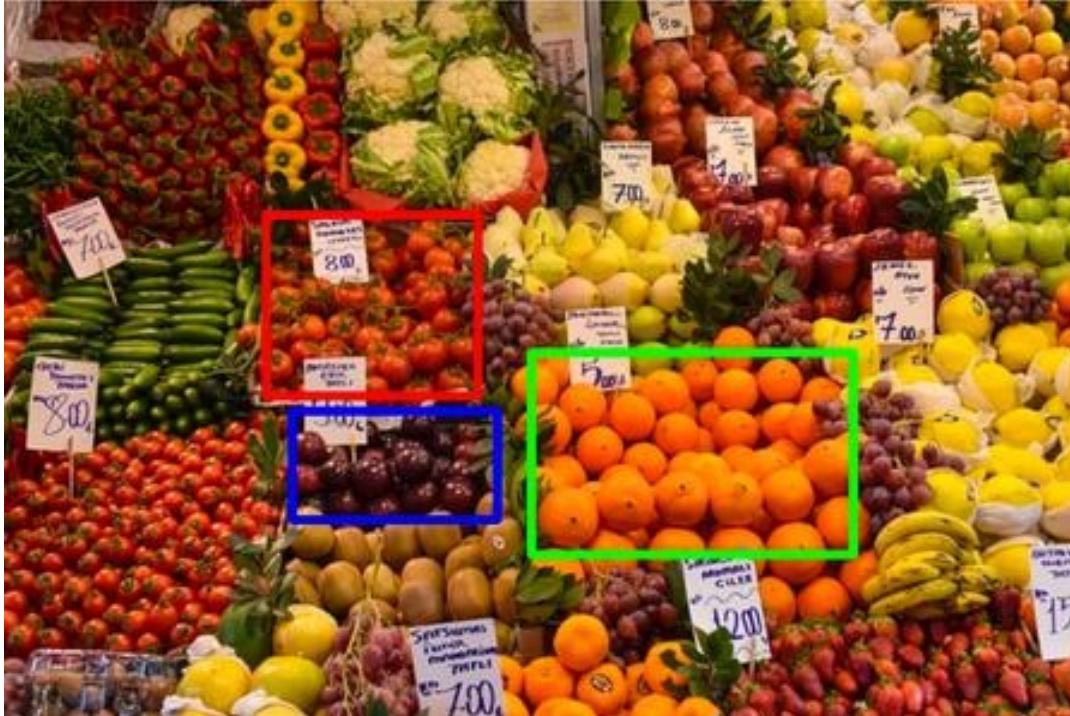
Model	$Q \rightarrow AR$ (%)
ViLBERT	54.0
Unicoder-VL	54.5
VLBERT-L	58.9
GPT4RoI-7B	78.6
Ours-7B	78.93

**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 for image regions

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?

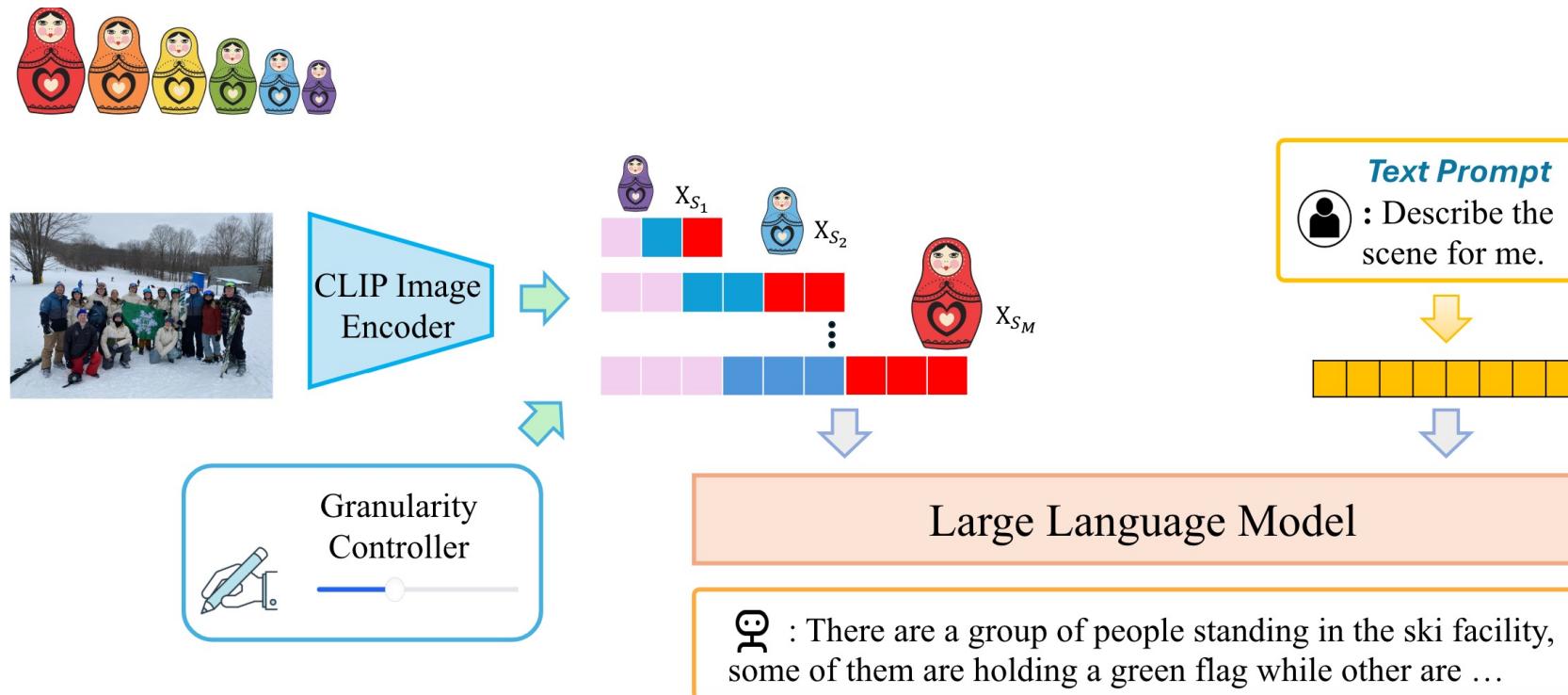
<thao> is in the photo, and she is standing next to another person. They both appear to be having a good time, possibly at a party or event. <thao> is smiling and making a peace sign with her hand, which suggests a positive and fun atmosphere.



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

Matryosha Multimodal Models

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



"Matryoshka Multimodal Models" Mu Cai et al. arXiv 2024

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
 - ...

Conclusions

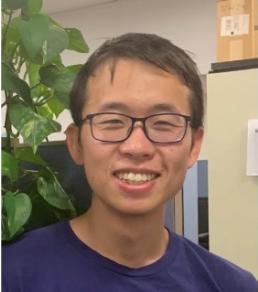
- From *specialist* to *generalist* multi-modal models
- Steerable (“aligned”) image understanding for open-world concepts
- Understand visual prompts; personalization; adaptive token lengths
- Code, models, online demo available:
<https://llava-vl.github.io/>, <https://vip-llava.github.io/>

Thank you

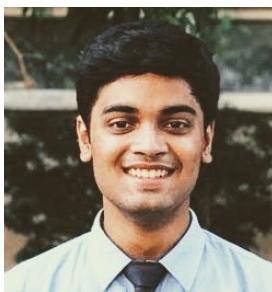
- Haotian Liu, Yuheng Li, Utkarsh Ojha, Mu Cai, Thao Nguyen, Chunyuan Li, Jianwei Yang, Jianfeng Gao , Siva Mustikovela, Greg Meyer, Yuning Chai, Dennis Park



Haotian Liu



Yuheng Li



Utkarsh Ojha



Mu Cai



Thao Nguyen