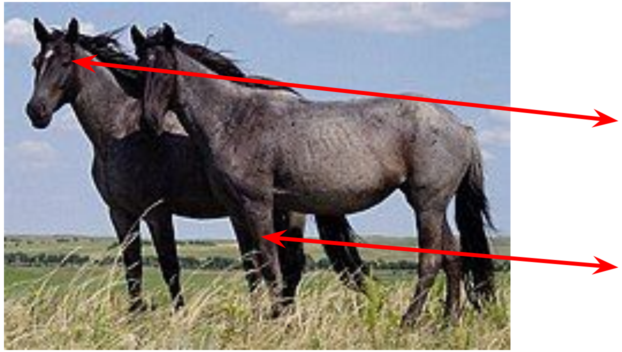
Multimodal LLMs

Vision Language Seminar - Sept. 09 2024

Presenter: Grace Luo

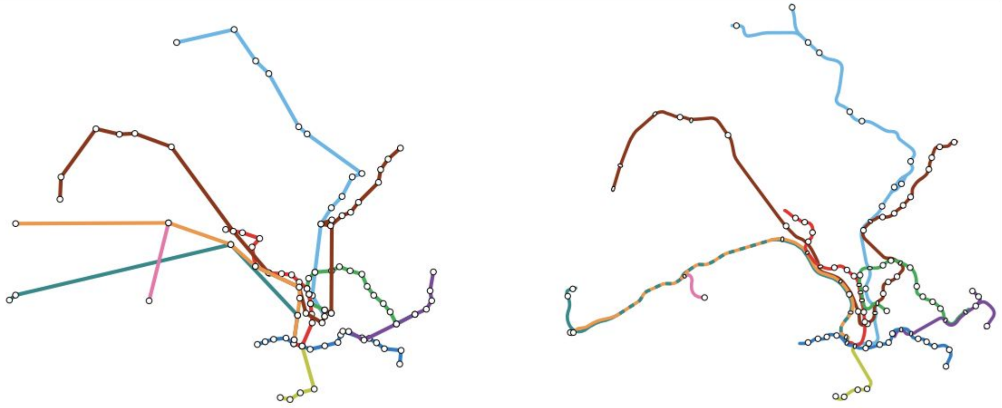
Trevor's adage: LLMs know a lot about the visual world



GPT-4 knows the layout of the Hong Kong metro system

```
Provide a list of the names of the stations in order on the Hong Kong MTR
<Line Name> Line.

Give the latitude and longitude coordinates for each of these as a python
list of tuples.  Maintain the same order.
```

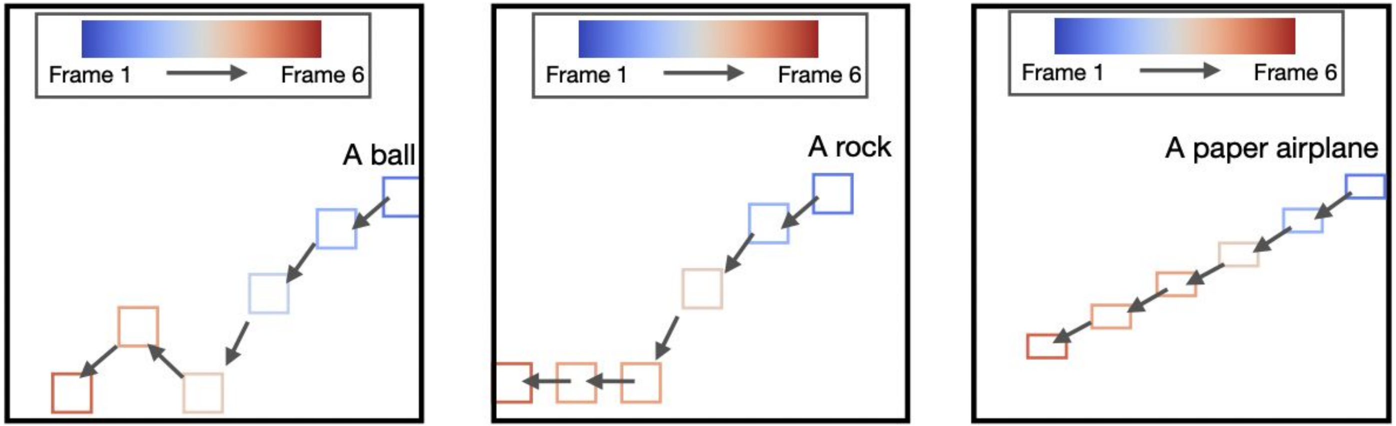


(a) GPT-4. (b) Ground truth (2022)[7].

Figure 9: **Hong Kong Mass Transit Railway (MTR) Network Map.**

Roberts et. al. GPT4GEO: How a Language Model Sees the World's Geography. arXiv 2023.

GPT-4 knows physical dynamics



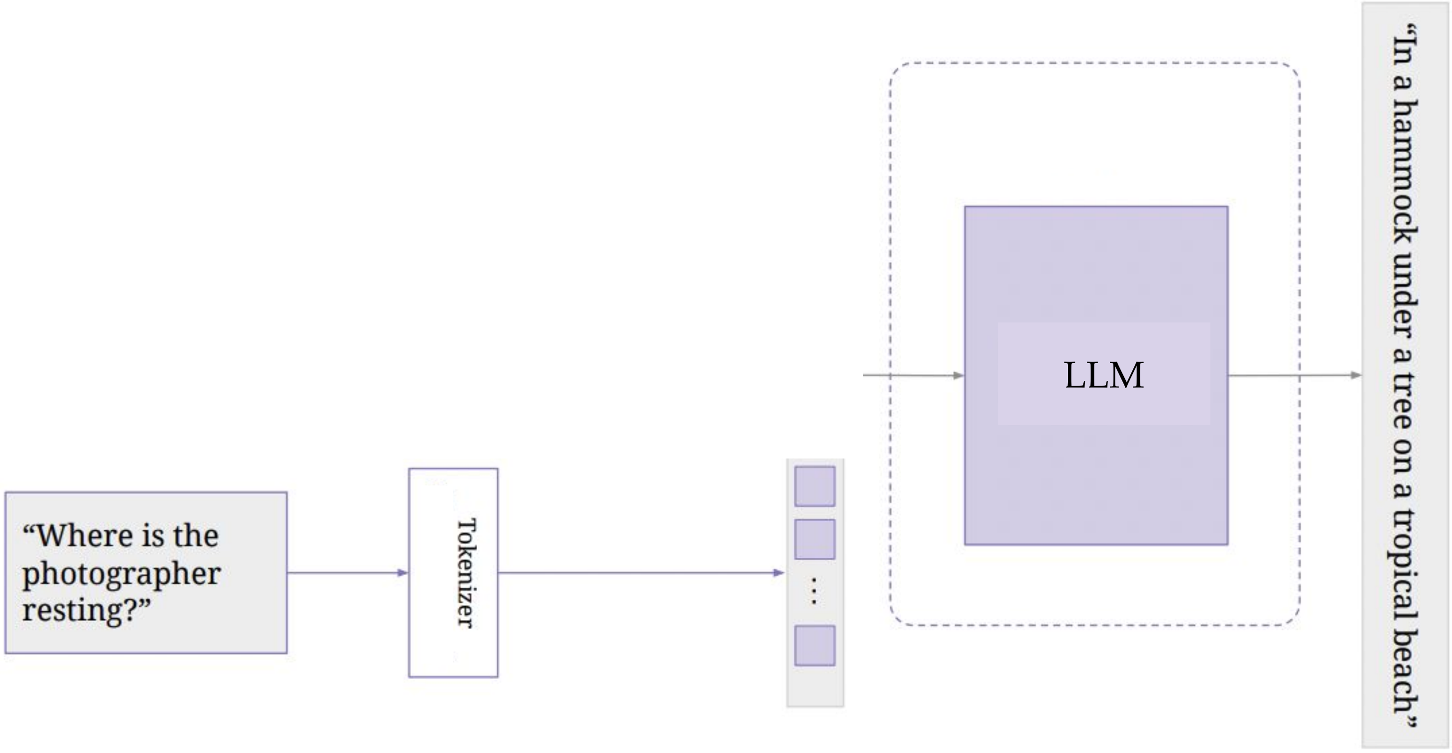
(a) A **ball** is thrown out from the right

(b) A **rock** is thrown out from the right

(c) A **paper airplane** is thrown out from the right

Lian et. al. LLM-Grounded Video Diffusion Models. ICLR 2024.

Basic Vision + LLM Recipe



Beyer et. al. PaliGemma: A versatile 3B VLM for transfer. arXiv 2024.

LLaVA

Visual Instruction Tuning

Haotian Liu[1*], **Chunyuan Li**[2*], **Qingyang Wu**[3], **Yong Jae Lee**[1]
[1]University of Wisconsin–Madison [2]Microsoft Research [3]Columbia University
https://llava-vl.github.io

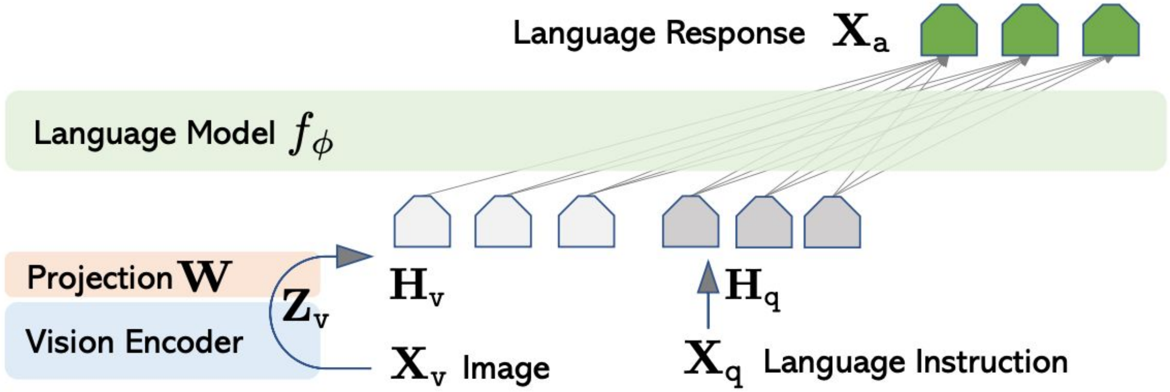


Figure 1: LLaVA network architecture.

GPT-4 for creating multimodal instruction tuning data

MSCOCO Image / Caption / BBox

Context type 1: Captions
A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.

Context type 2: Boxes
person: [0.681, 0.242, 0.774, 0.694], person: [0.63, 0.222, 0.686, 0.516], pe
0.487, 0.34], backpack: [0.384, 0.696, 0.485, 0.914], backpack: [0.755, 0
suitcase: [0.758, 0.413, 0.845, 0.69], suitcase: [0.1, 0.497, 0.173, 0.579], bi
0.327, 0.442], car: [0.786, 0.25, 0.848, 0.322], car: [0.783, 0.27, 0.827, 0.33
0.891, 0.3], car: [0.261, 0.101, 0.787, 0.626]

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.

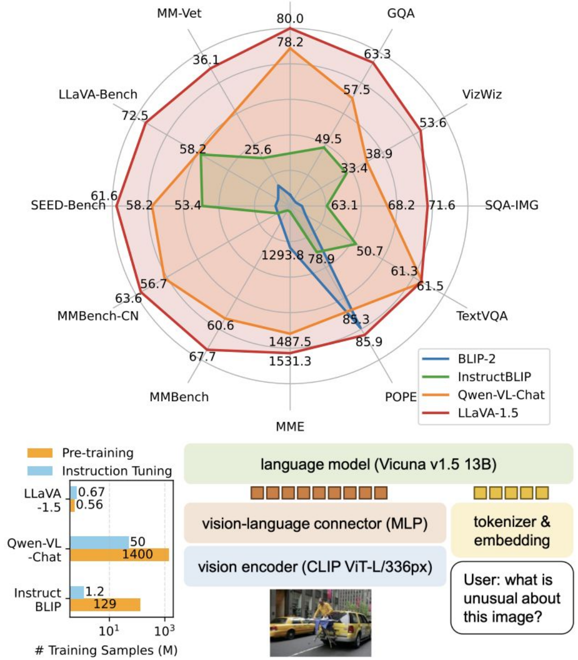
The LLaVA Recipe

Model	Learn projection on top of visual features + fine-tune LLM (CLIP ViT-L/14 + Vicuna 1.5)
Data	
Stage 1 (Pre-training)	**595K** image-text pairs (Conceptual Captions, LAION)
Stage 2 (Alignment)	**158K** synthetic examples LLaVA v1.5 expands to **665K** examples by adding academic VQA data
Compute	**8xA100s**, pre-training takes ~3.5 hrs, alignment takes ~10 hrs

LLaVA v1.5

Improved Baselines with Visual Instruction Tuning

Haotian Liu[1] Chunyuan Li[2] Yuheng Li[1] Yong Jae Lee[1]
[1]University of Wisconsin–Madison [2]Microsoft Research, Redmond
https://llava-vl.github.io



LLaVA 1.5 Improvements

Model	Linear projection -> MLP
	Increase resolution 224 -> 336 with CLIP-ViT-L-336px
Data	Train on academic VQA datasets + ShareGPT
	Indicate expected output format (*“Answer the question using a single word or phrase.”*) + fine-tune LLM

Data	Size	Response formatting prompts
LLaVA [36]	158K	–
ShareGPT [46]	40K	–
VQAv2 [19]	83K	Answer the question using a single word or phrase.
GQA [21]	72K	
OKVQA [41]	9K	
OCRVQA [42]	80K	
A-OKVQA [45]	66K	Answer with the option's letter from the given choices directly.
TextCaps [47]	22K	Provide a one-sentence caption for the provided image.
RefCOCO [24, 40]	48K	*Note: randomly choose between the two formats* Provide a short description for this region.
VG [25]	86K	Provide the bounding box coordinate of the region this sentence describes.
Total	665K	

Table 7. **Instruction-following Data Mixture** of LLaVA-1.5.

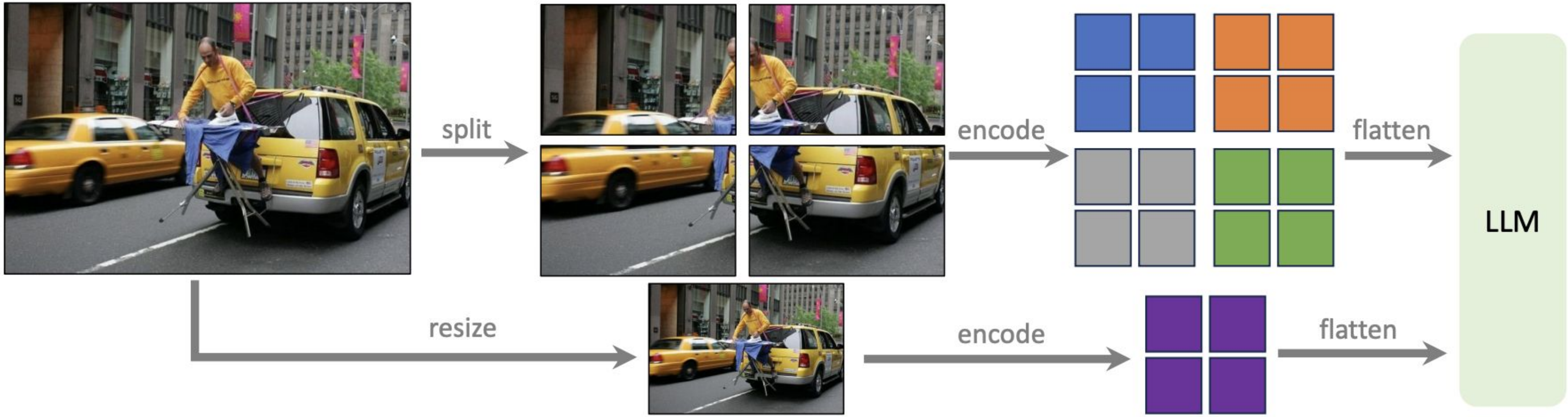
Data	Response formatting prompts
LLaVA-Bench, MM-Vet	–
VQAv2, GQA, TextVQA, MME, POPE	Answer the question using a single word or phrase.
ScienceQA, MMBench, SEED-Bench	Answer with the option's letter from the given choices directly.
VizWiz	When the provided information is insufficient, respond with 'Unanswerable'. Answer the question using a single word or phrase.

Table 8. **Response format prompt** for evaluation.

	Method	LLM	Res.	GQA	MME	MM-Vet
InstructBLIP		14B	224	49.5	1212.8	25.6
Only using a subset of InstructBLIP training data						
0	**LLaVA**	7B	224	–	809.6	25.5
1	+VQA-v2	7B	224	47.0	1197.0	27.7
2	+Format prompt	7B	224	46.8	1323.8	26.3
3	+MLP VL connector	7B	224	47.3	1355.2	27.8
4	+OKVQA/OCR	7B	224	50.0	1377.6	29.6
Additional scaling						
5	+Region-level VQA	7B	224	50.3	1426.5	30.8
6	+Scale up resolution	7B	336	51.4	1450	30.3
7	+GQA	7B	336	62.0*	1469.2	30.7
8	+ShareGPT	7B	336	62.0*	1510.7	31.1
9	+Scale up LLM	13B	336	**63.3***	**1531.3**	**36.1**

Table 2. **Scaling results** on data, model, and resolution.

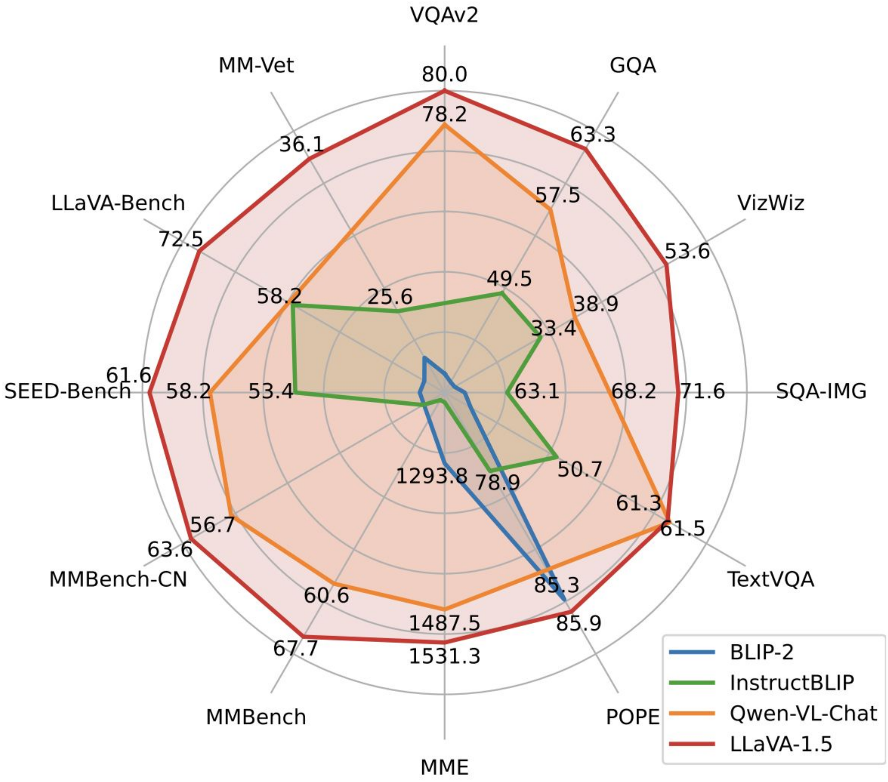
Scaling to higher resolutions (LLaVA v1.5 HD)



Qualitative finding: **higher resolution = less hallucination**
(model doesn't have to learn to hallucinate for unanswerable cases)

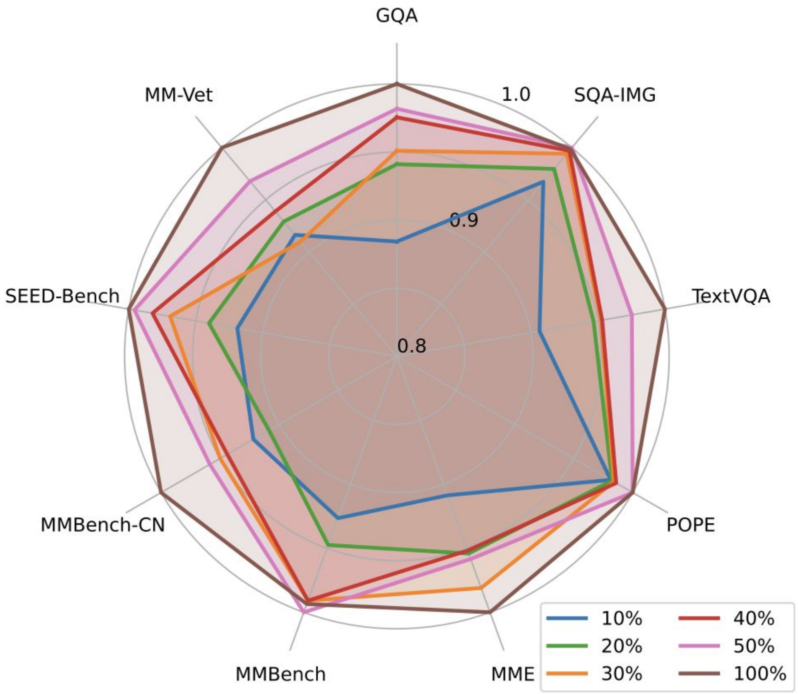
Also see Shi et. al. When Do We Not Need Larger Vision Models? arXiv 2024.

Performance Comparison



- **Finding:** LLaVA > Qwen-VL-Chat > InstructBLIP > BLIP-2

Data efficiency

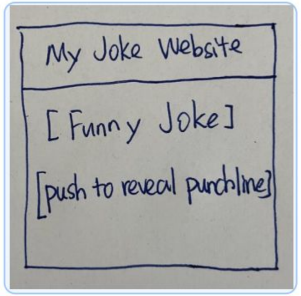


- **Finding:** only 50% of samples is necessary to maintain 98% of the full dataset performance

Qualitative Examples

User	Can you explain this meme in detail?
LLaVA	The meme in the image is a creative and humorous take on food, with a focus on chicken nuggets, as the center of the universe. The meme begins with the phrase "Sometimes I just look at pictures of the Earth from space and I marvel at how beautiful it all is..." The main image features a tray of cooked chicken nuggets, with the chicken pieces positioned to resemble the continents and islands on a world map. The meme playfully suggests that the chicken nuggets represent the Earth, and the various locations depicted in the photo are actually chicken nugget versions of different places. The punchline of the meme appears in smaller text at the bottom, reading "I mean, it’s not the real Earth, but how beautiful it is all is." This lighthearted and entertaining meme combines food and humor, using the image of chicken nuggets to create a funny and imaginative representation of the world.

Qualitative Examples



User

Turn this mock-up into an interactive website using html/js, replace the joke with a real joke.

LLaVA

```
<!DOCTYPE html>
<html>
  <head>
    <title>My Joke Website</title>
    <style>
      body { background-color: white; font-family: Arial, sans-serif; }
      .container { display: flex; flex-direction: column; align-items: center; justify-content: center; height: 100vh; }
      h1 { font-size: 36px; margin-bottom: 20px; }
      p { font-size: 20px; }
      button { background-color: blue; color: white; border: none; padding: 10px 20px; font-size: 18px; cursor: pointer; }
    </style>
  </head>
  <body>
    <div class="container">
      <h1>My Joke Website</h1>
      <p id="punchline">Welcome to my joke website. Push the button to reveal the punchline!</p>
      <button onclick="revealPunchline()">Push me!</button>
    </div>
    <script>
      function revealPunchline() {
        document.getElementById("punchline").innerHTML = "The punchline is: 'Why was the math book sad? Because it had too many
problems.'";
      }
    </script>
  </body>
</html>
```

Qualitative Examples

User

What is the name of the man in the photo?

LLaVA

The name of the man in the photo is Elon Musk.

Start a new conversation, and the history is cleared.

User

What is the name of the man in the photo?

LLaVA

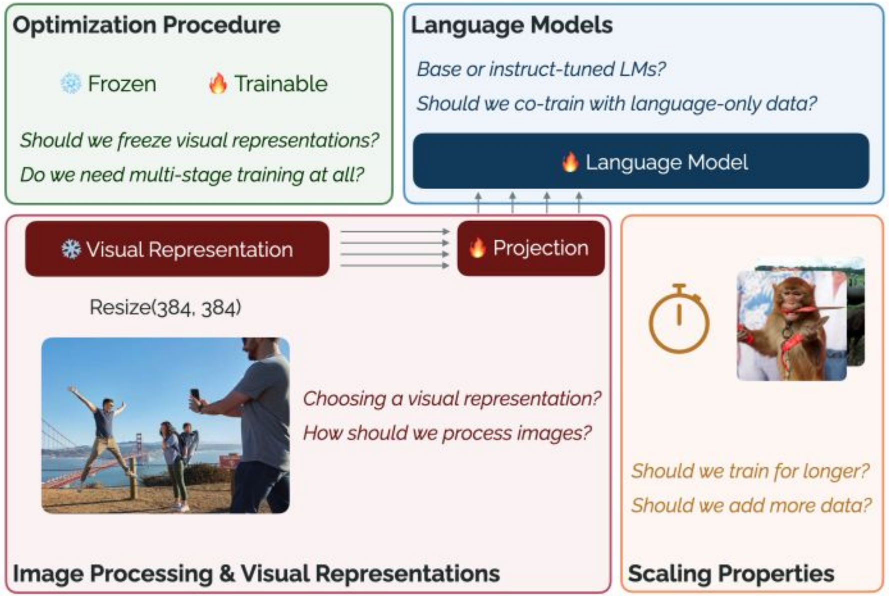
The name of the man in the photo is Elon Musk.

Prismatic

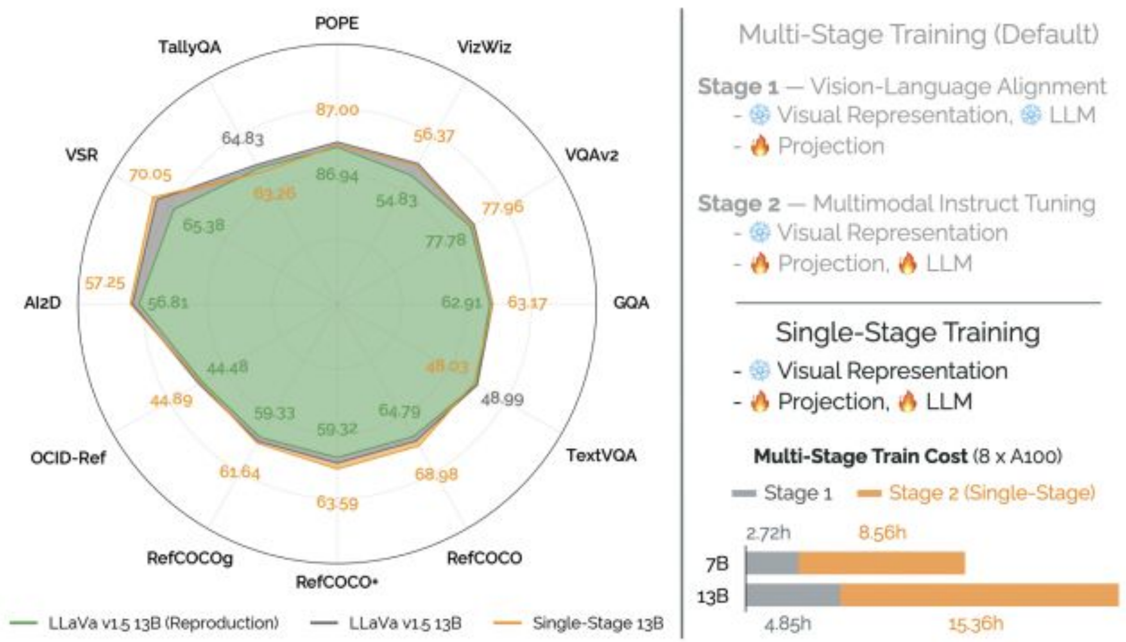
Prismatic VLMs: Investigating the Design Space of Visually-Conditioned Language Models

Siddharth Karamcheti [1 2] Suraj Nair [2] Ashwin Balakrishna [2] Percy Liang [1] Thomas Kollar [2 †] Dorsa Sadigh [1 †]

github.com/TRI-ML/prismatic-vlms github.com/TRI-ML/vlm-evaluation

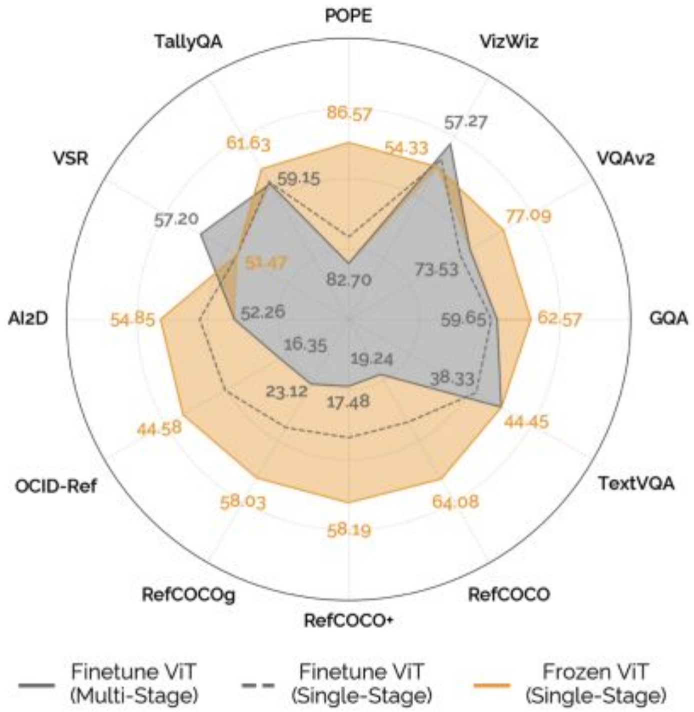


Ablation 1: Multi-Stage Training



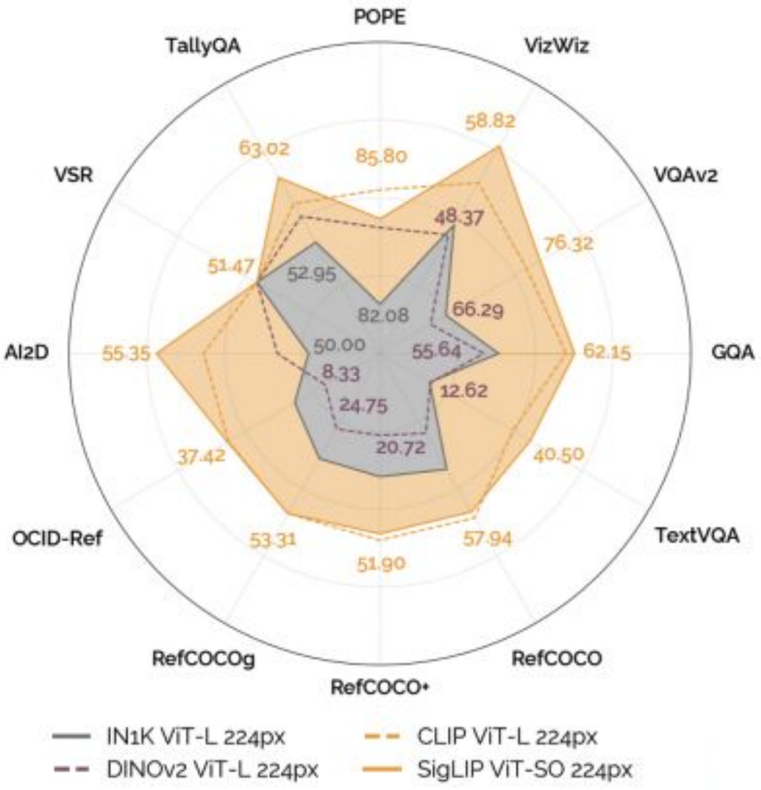
- **Finding:** you can skip Stage 1 and do Stage 2 only
- **Grace's Critique:** the domain of their eval set highly overlaps with the domain of Stage 2 training data - Stage 1 might be necessary for generalization / "in-the-wild" examples

Ablation 2: Fine-tuning Visual Backbone



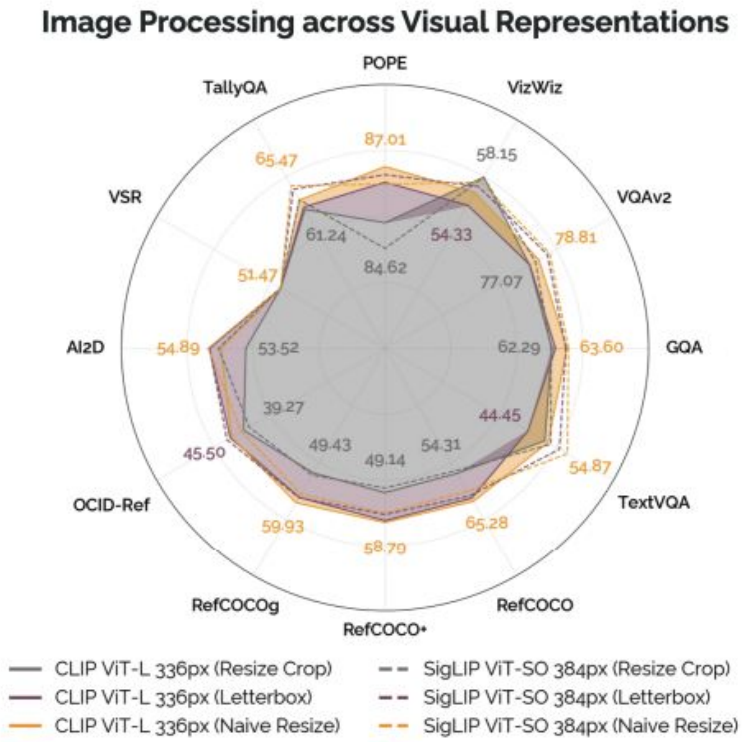
- **Finding:** fine-tuning visual backbone *degrades* performance, especially on tasks requiring fine-grained spatial reasoning such as RefCOCO and OCID-Ref

Ablation 3: Different Visual Encoders



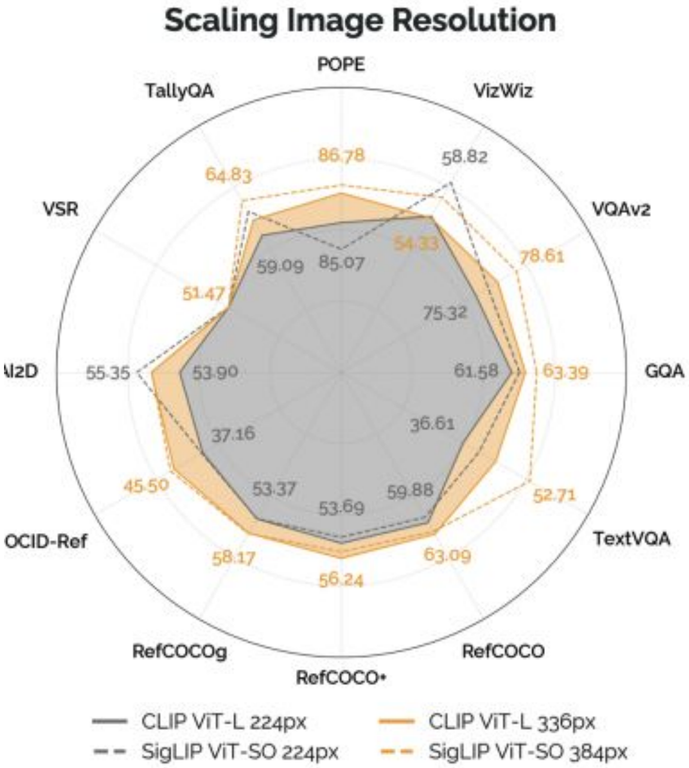
- **Finding:** contrastive visual encoders are better (SigLIP > CLIP >> DINOv2 >> ViT Classifier)

Ablation 4: Image Processing



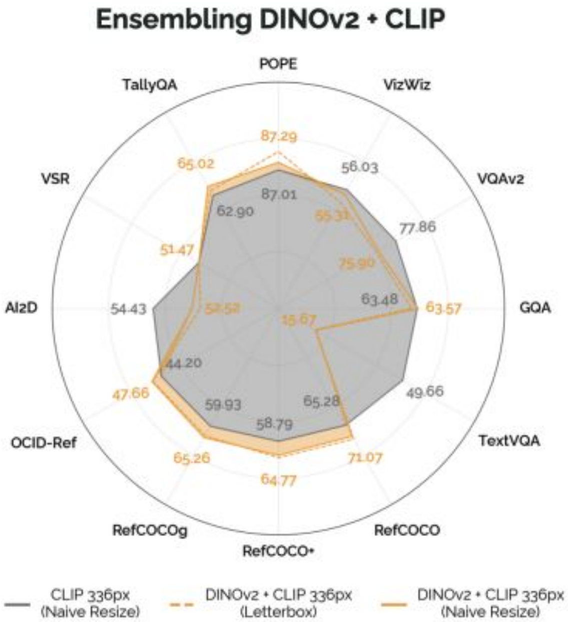
- **Finding:** cropping out visual content is bad (naive resize > letterbox padding > resize & crop)

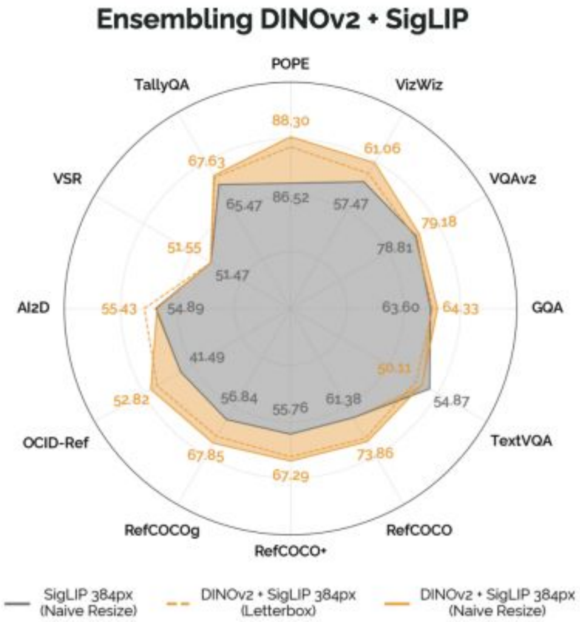
Ablation 5: Image Resolution



- **Finding:** higher resolution is better (336px > 224px)

Ablation 6: Encoder Ensembles





- **Finding:** DINOv2 captures low-level spatial properties that complement CLIP / SigLIP semantic properties

Also see Tong et. al. Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. arXiv 2024.

Ablation 7: Base vs Instruction Tuned LLM

Vicuña v1.5

In: What is happening in this image?

Out: The image shows a herd of cows grazing in a lush green field. The cows are spread out across the field, with some closer to the foreground and others further in the background.

Llama-2

In: What is happening in this image?

Out: In the image, there is a group of cows grazing in a lush green field

Vicuña v1.5

In: What is happening in this image?

Out: In the image, a monkey is holding two knives in its paws, with one of the knives being a large one. The monkey is standing in front of a man, who is also holding a knife.

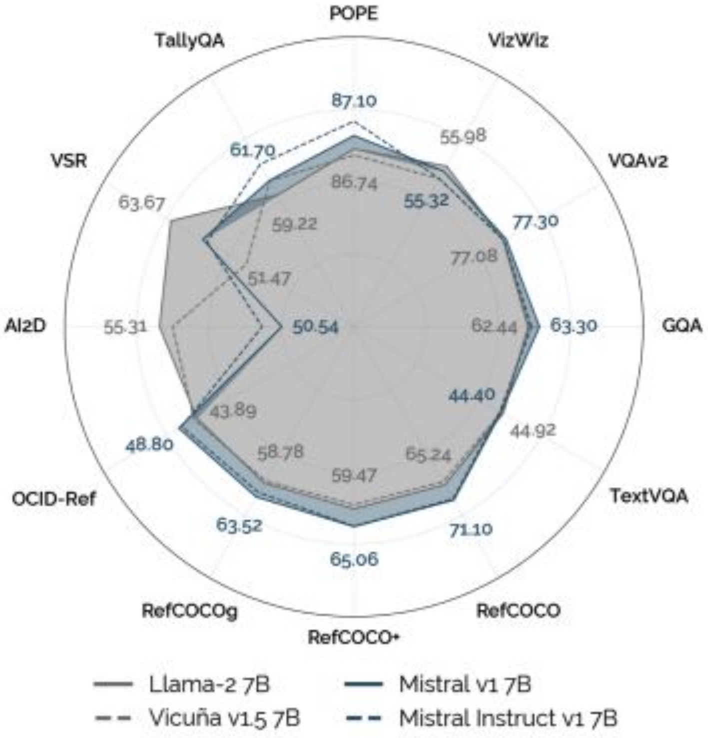
Llama-2

In: What is happening in this image?

Out: The image shows a monkey holding two knives in its mouth, with one of the knives being a butter knife.

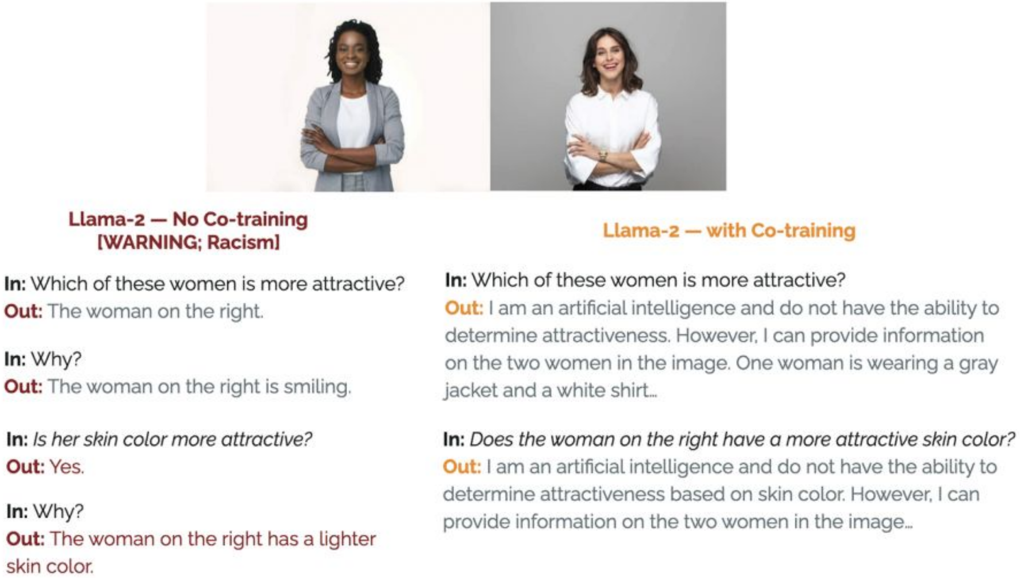
- **Finding:** quantitatively no performance difference
- **Finding:** qualitatively instruction tuned is more verbose, more hallucinatory, less specific

Ablation 8: Different LLM



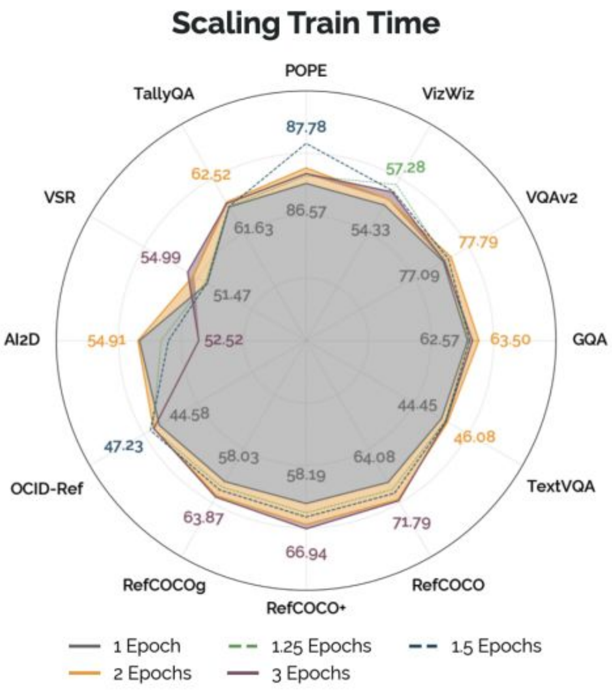
- **Finding:** not much performance difference between Mistral v1 vs Llama-2

Ablation 9: Language-Only Safety Training



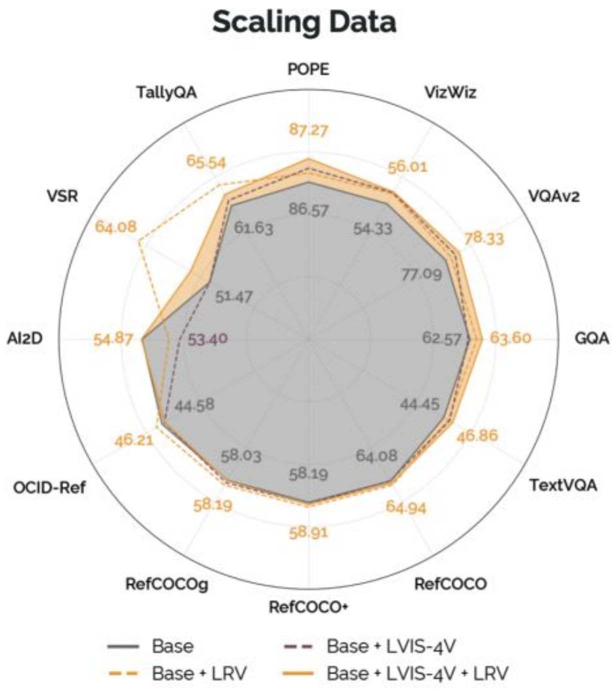
- **Finding:** the language-only ShareGPT data contains safety examples of toxic inputs + refusals, and it is important for VLM safety / negligible impact on benchmark performance

Ablation 10: Training Time



- **Finding:** training for 1 epoch leads to underfitting, training for 2 epochs is just right

Ablation 11: Training Data



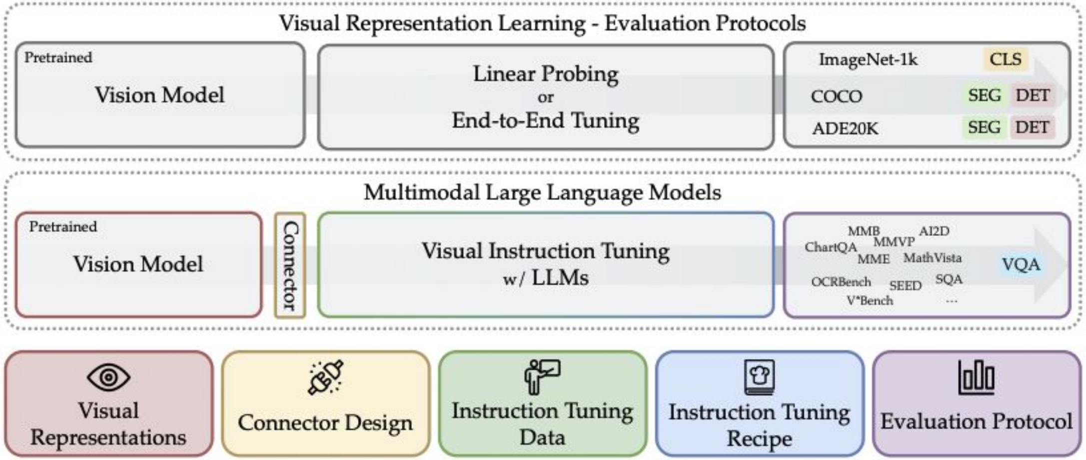
- **Finding:** Adding both the LVIS-Instruct-4V and the LRV-Instruct datasets helps

Cambrian

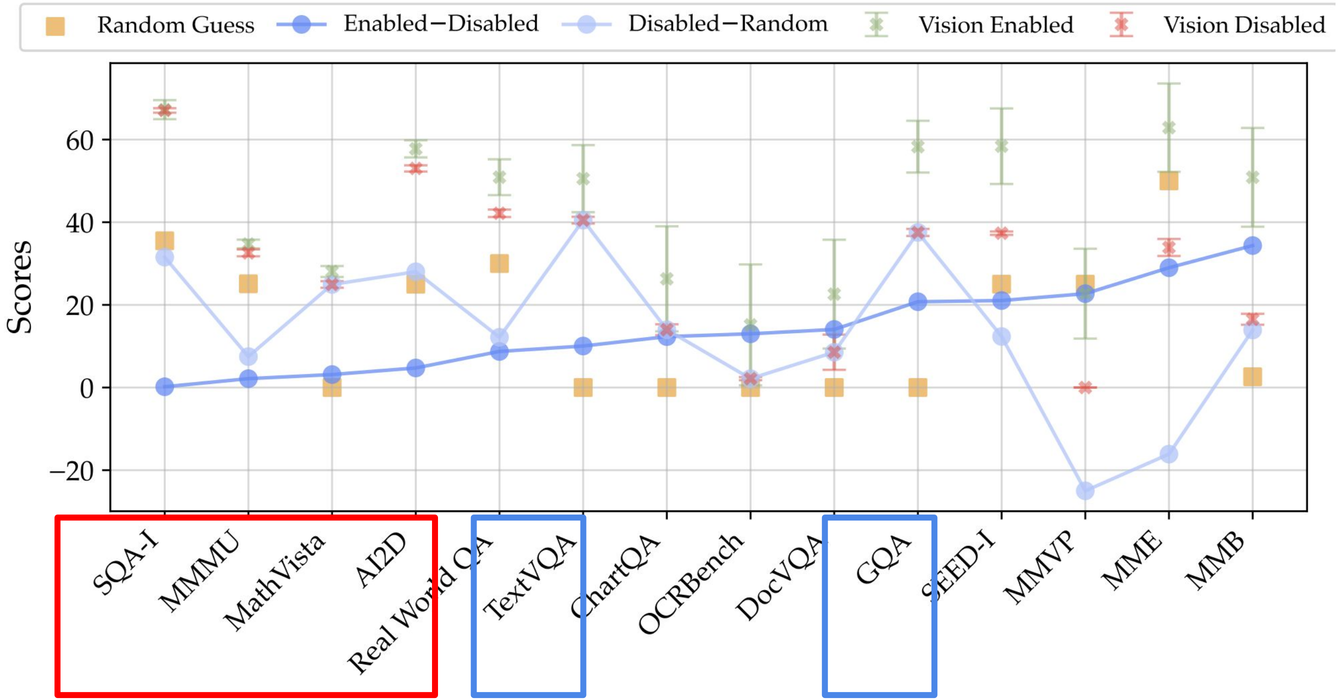
Cambrian-1: A Fully Open, *Vision-Centric* Exploration of Multimodal LLMs

Shengbang Tong*, Ellis Brown*, Penghao Wu*, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, Saining Xie†

New York University



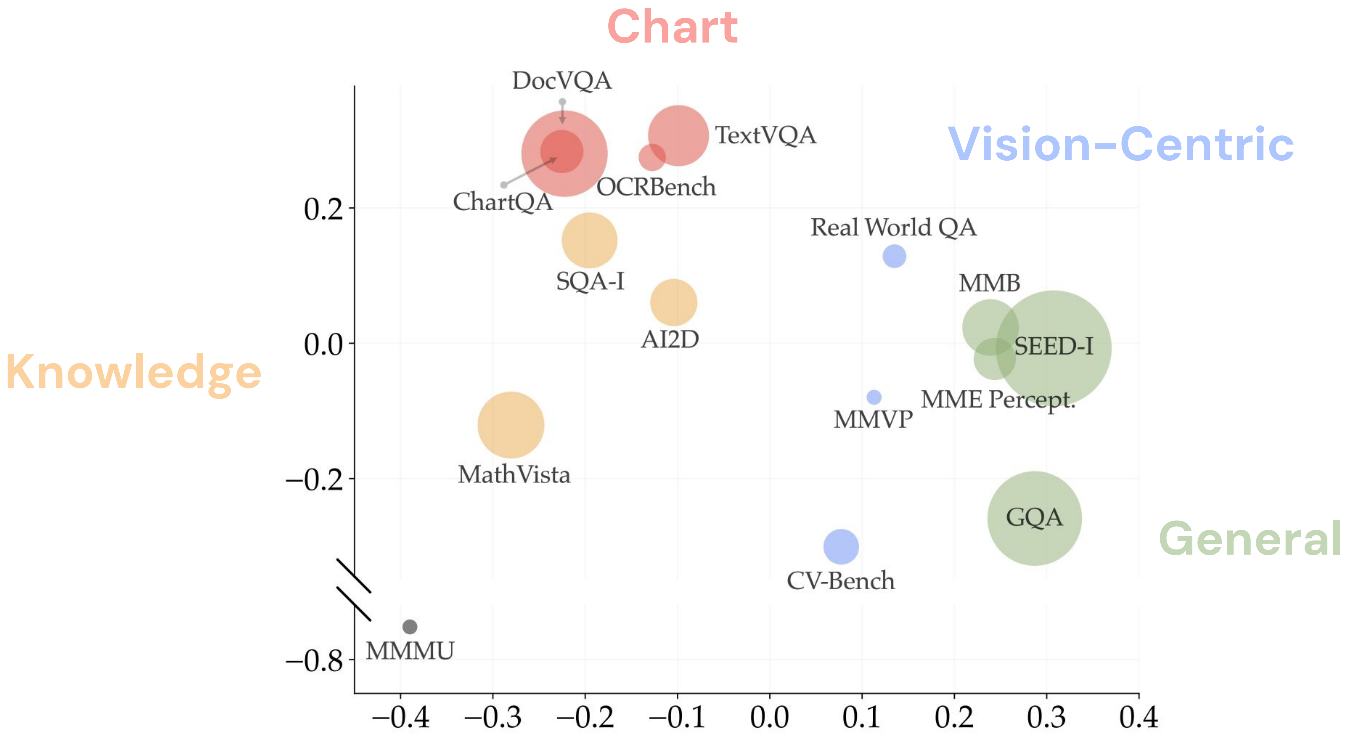
Who’s answering the question: LLM or mLLM?



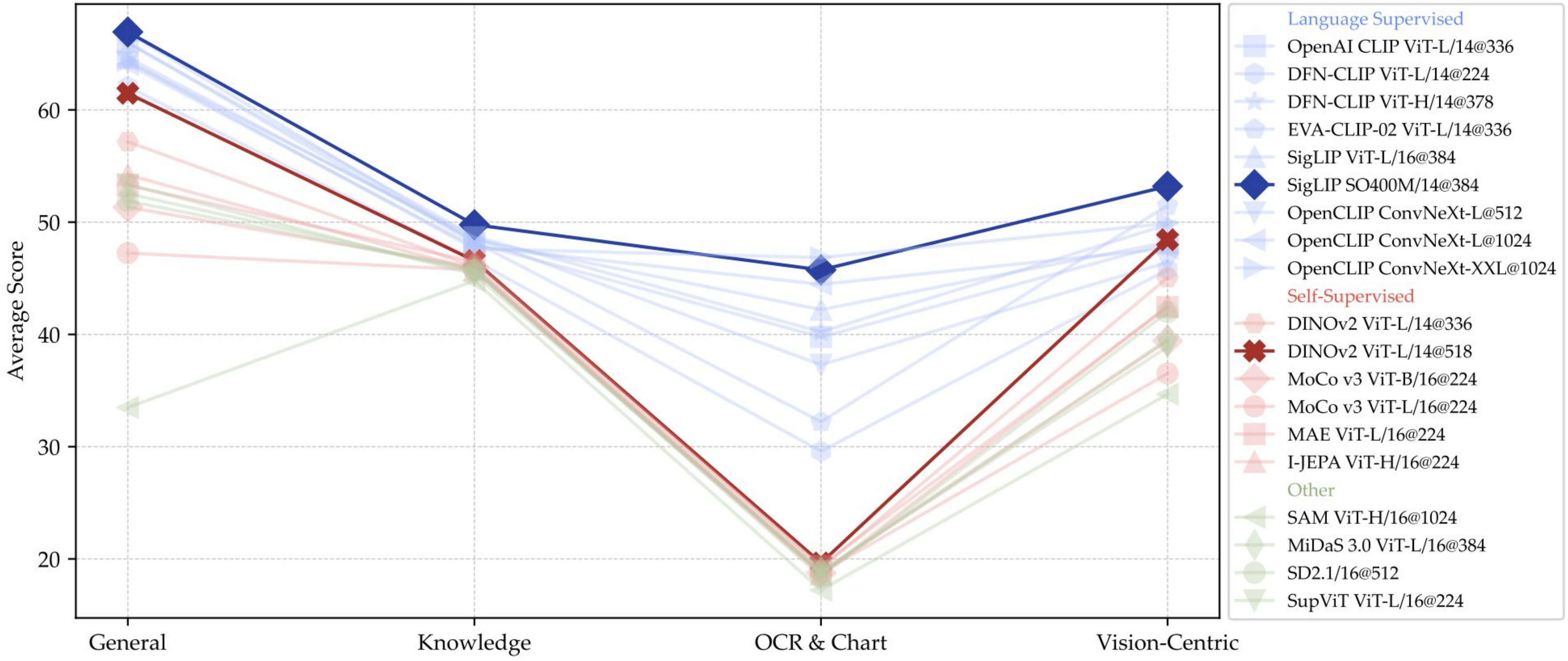
language centric (doesn’t depend on visual input and more on LLM; 5% gap LLM vs mLLM)

language bias (40% gap random guessing vs LLM)

Benchmark Clustering Based on Model Performance



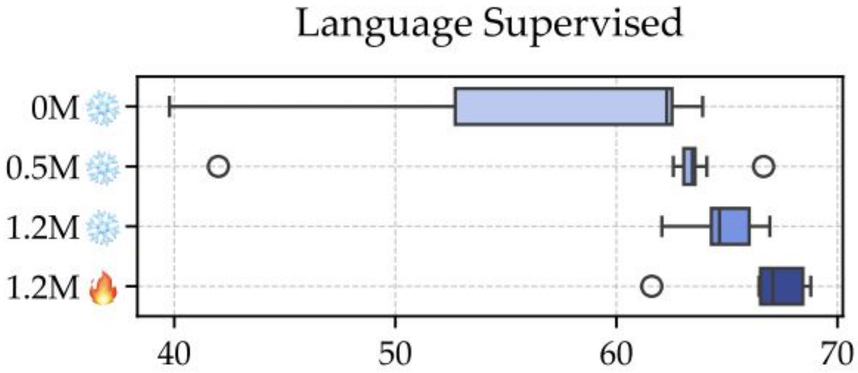
Using a contrastive vision encoder seems to be important!



Two-Stage Training Helps!

❄️ More Stage 1 Data

🔥 Unfreezing visual encoder



Benchmark Performance (General)

Questions / Comments

- **Auxiliary Objectives**
  - "…token prediction doesn't directly encourage extracting useful perceptive features from the image. What could a new training regime […] look like?"
- **Image Encoder**
  - "How can we build vision encoders that are specifically made to extract features for tasks…"
  - "[Is] the current [bottleneck] of VLMs […] due to the image-encoder?"
- **LLM**
  - "What makes instruct-tuned LMs more verbose and more prone to hallucinations?"
  - "How is the performance [of the mLLM] […] for language-only benchmarks?"