Does Language Model Need Better Visual Grounding for Meaning and Understanding?

Saining Xie
Courant Institute, NYU
11/2024

Does AI need sensory grounding for meaning and understanding?

- Stevan Harnad (1990): the symbol grounding problem
- *“Symbols in AI systems must have sensory grounding (or: bodily grounding, or external grounding) to have meaning.”*

Do humans require sensory grounding for meaning and understanding?

- Thomas Aquinas(1200s)
- *"There's nothing in mind that wasn't first in the senses"*

Do humans require sensory grounding for meaning and understanding?

- Diderot, Condilac (1700s): sensim
- *"no cognition without sensasion"*

Do humans require sensory grounding for meaning and understanding?

- Avicenna (Ibn Sina) (1000s): Avicenna's Floating Man
- Avicenna: *"floating man can think about himself without ever sensing (and without sensory capabilities?)"*
- others*: "he could also think about mathematics, logic, philosophy, and could form hypotheses about external reality."*

A less philosophical question

Is sensing useful?

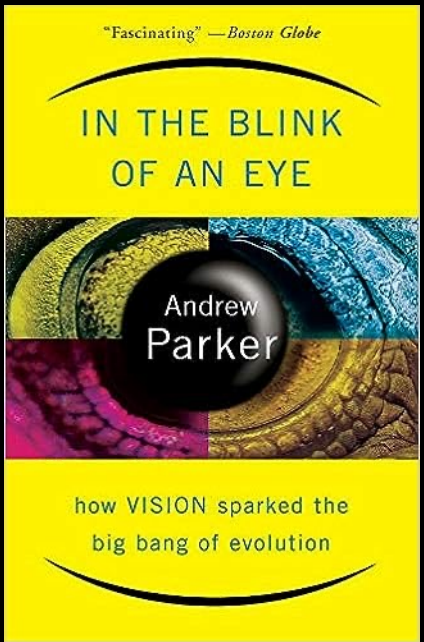
- Does sensing boost thinking (to a new level) in language models?

(and in intelligent creatures generally?)

Credit: David Chalmers

phylogeny of intelligence

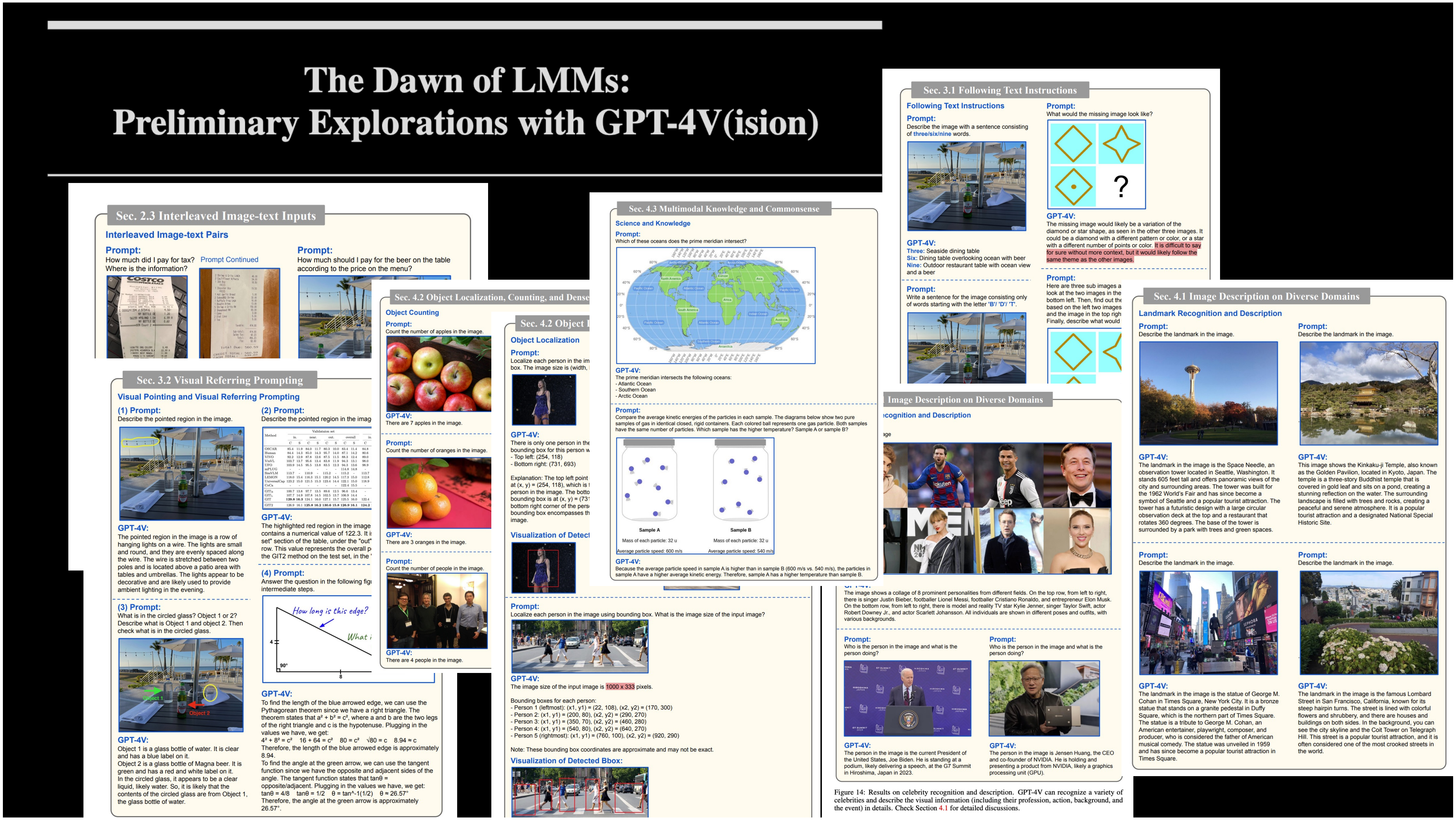
538.8 million years ago
Cambrian era
“biological explosion”



“The evolution of the eye is likely to have been a catalyst for the explosion, initiating an arms race between organisms that were increasingly aware of their surroundings.”

https://www.nhm.ac.uk/discover/eyes-on-the-prize-evolution-of-vision.html

LLMs: knowledgeable but blindfolded


The Dawn of LMMs:
Preliminary Explorations with GPT-4V(ision)
Sec. 2.3 Interleaved Image-text Inputs
Interleaved Image-text Pairs
Sec. 3.2 Visual Referring Prompting
Visual Pointing and Visual Referring Prompting
Sec. 4.2 Object Localization, Counting, and Dense
Object Counting
Object Localization
Sec. 4.3 Multimodal Knowledge and Commonsense
Science and Knowledge
Sec. 3.1 Following Text Instructions
Following Text Instructions
Sec. 4.1 Image Description on Diverse Domains
Landmark Recognition and Description
Figure 14: Results on celebrity recognition and description. GPT-4V can recognize a variety of celebrities and describe the visual information (including their profession, action, background, and the event) in details. Check Section 4.1 for detailed discussions.

Why do we need better visual representations?

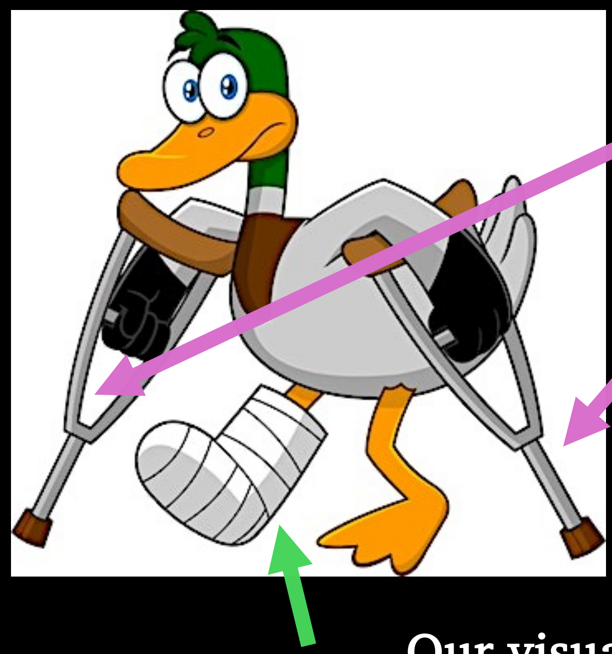
“Most of human knowledge (and almost all of animal knowledge) comes from our sensory experience of the physical world.”

- “Language is the icing on the cake. We need the cake to support the icing.”

LeCun’s cake 2.0 ?

https://lexfridman.com/yann-lecun-3-transcript/

Relying too heavily too early on language can act as a shortcut, compensating for the deficiencies in learning effective visual representations.

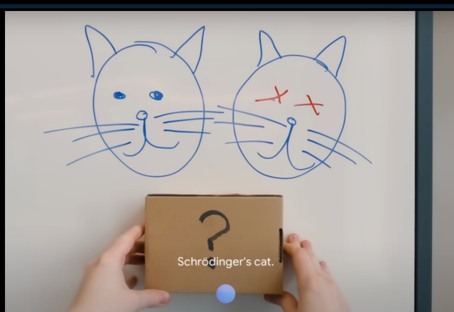


Language vs Visual Intelligence

"Who won the game?"

[GPT-4O, OpenAI]

"what does this remind you of?"

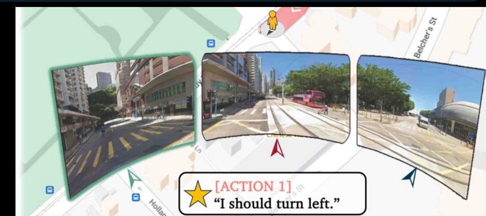


[Project Astra, Google]

"Where can I buy this mug?"

[V* - CVPR 2024]

"Which direction leads home?"



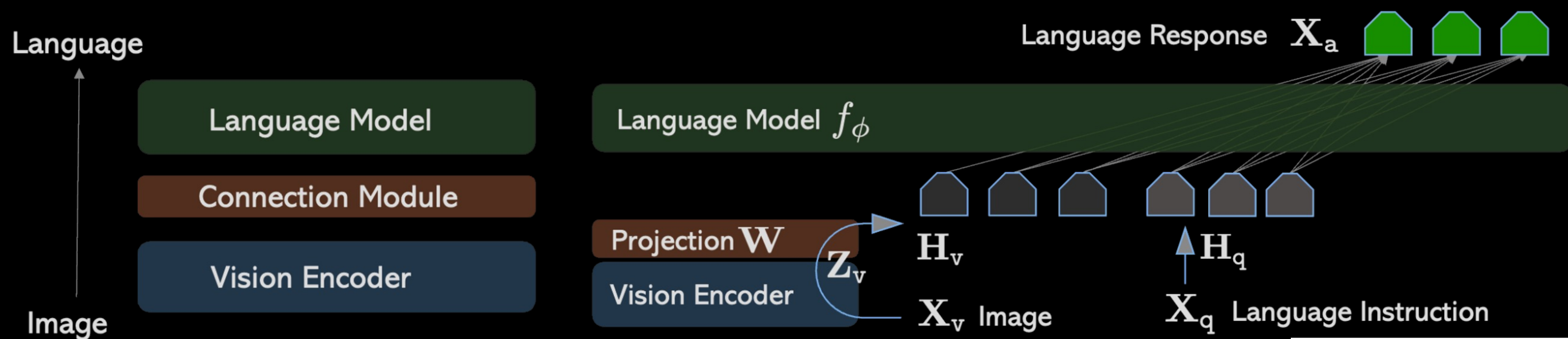
[V-IRL - ECCV 2024]

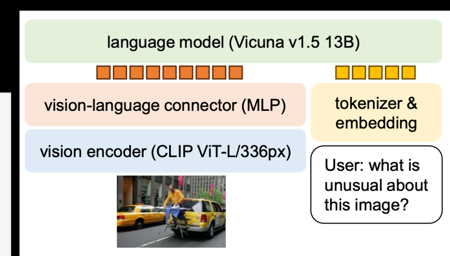
"Thinking in Space"

Tasks Requiring more
Robust Visual-Spatial Intelligence

Tasks Requiring more
Strong Language Capability

A typical MLLM pipeline (LLaVA)

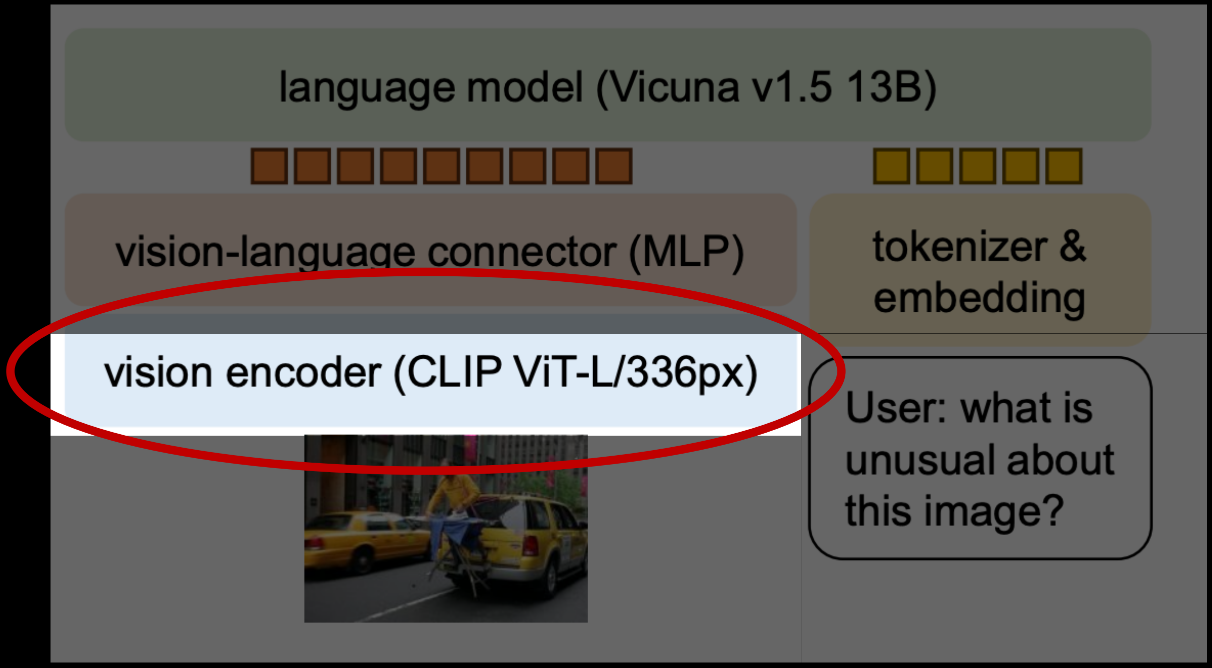




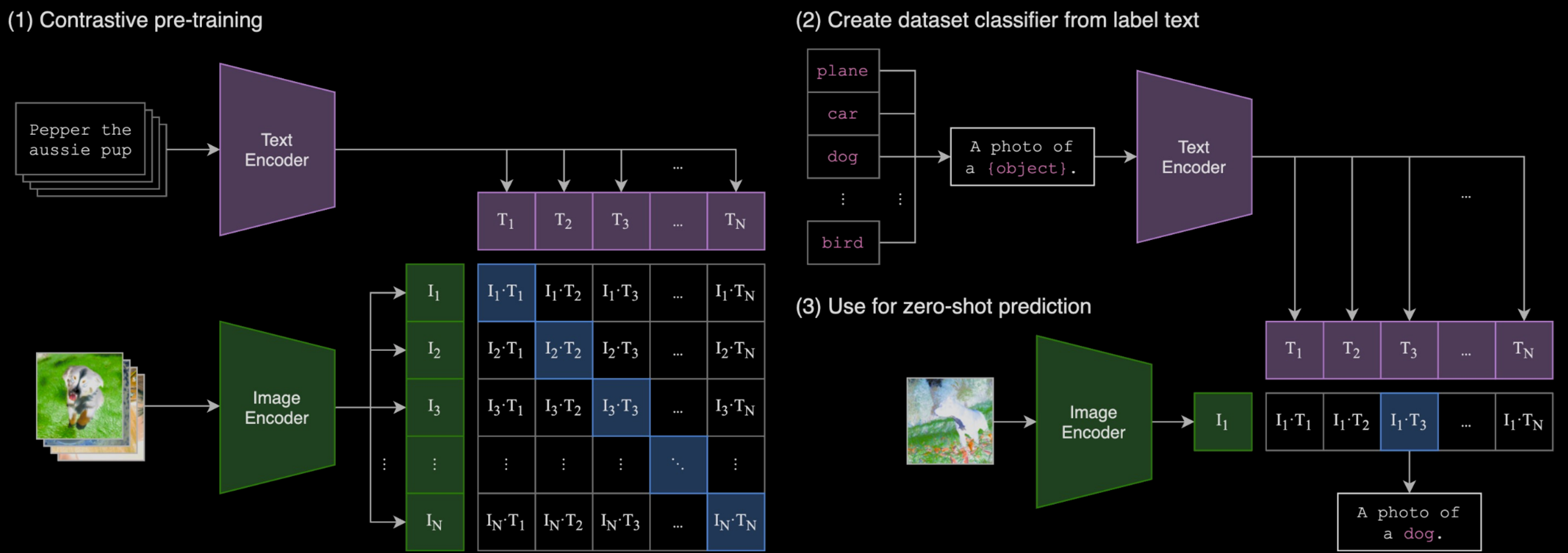
- **Stage 1: Pre-training for Feature Alignment.**
  Only the projection matrix is updated, based on a subset of CC3M.
- **Stage 2: Fine-tuning End-to-End.** Both the projection matrix and LLM are updated
  - **Visual Chat**: Our generated multimodal instruction data for daily user-oriented applications.
  - **Science QA**: Multimodal reasoning dataset for the science domain.

Visual Instruction Tuning, Liu et al.,

A typical MLLM pipeline (LLaVA)



CLIP vison encoder



Learning transferable visual models from natural language supervision, Radford et al. 2021

Key question #1:

Does language model need better visual grounding/representations for meaning and understanding?

Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs

Peter Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, Saining Xie

CVPR 2024 (Oral)

Q: Is the dog facing left or right from the camera's perspective?

The dog is **facing to the right** from the camera's perspective. You can tell by the position of its tail and the orientation of its body.

Q: Do you see any window in this image?

No, there are **no windows visible** in this image. The image shows a traditional red pillar box, which is a type of mailbox found in the United Kingdom.

Q: Do the individuals in the picture face the front or the back?

The individuals in the picture are **facing away from the camera**, so we see them from the back.

Q: In this image, how many eyes can you see on the animal?

The image shows **one eye** of the animal. It's a close-up of a bald eagle facing slightly to its right, which presents only one side of its face to the viewer.

Q: How many wheels can you see in the image?

In the image provided, I can see **two wheels** on the visible side of the car.

Q: Does the keyboard have a backlight?

Based on the image, it appears that the keyboard **does have a backlight**. This is indicated by the illumination seen around the keys, which suggests that there is a light source beneath them to enhance visibility in low-light conditions.

Q: Is the person in the picture on the grass or on the gravel path?

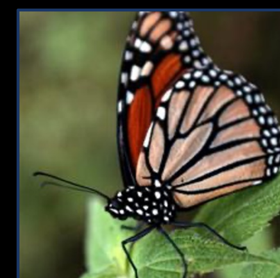
The person in the image appears to be running **on a gravel path**. You can discern this from the texture and color contrast between the light-colored path and the green grass surrounding it.

Q: Is the edge of the heart in the picture white or dark-colored?

The image you've provided is small and of low resolution, but it appears that the hearts have a **dark-colored edge** or outline. This outline varies in visibility, likely due to the quality of the image.

Q: Is the school bus parked facing the camera or away from the camera?

The school bus is parked **facing away from the camera**. You can tell because the rear lights and the back of the exterior rearview mirror are visible. The text "SCHOOL BUS" is also reversed, which is typically seen on the front of a bus, designed so that it reads correctly in a rearview mirror.



Q: Are the butterfly's feet visible?

The butterfly's feet are **not visible** in this image. The wings and part of the body of the butterfly are in view, but the feet are either hidden beneath the body or out of the frame of the picture. Butterflies have delicate, slender feet which are often not the most noticeable feature when they are perched, especially from the angle shown in this image.

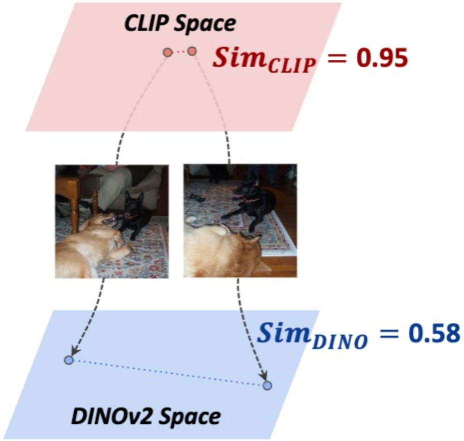
We can systematically identify instances where the visual question answering (VQA) capabilities of GPT-4V fall short.

Constructing MMVP Benchmark via CLIP-blind Pairs.

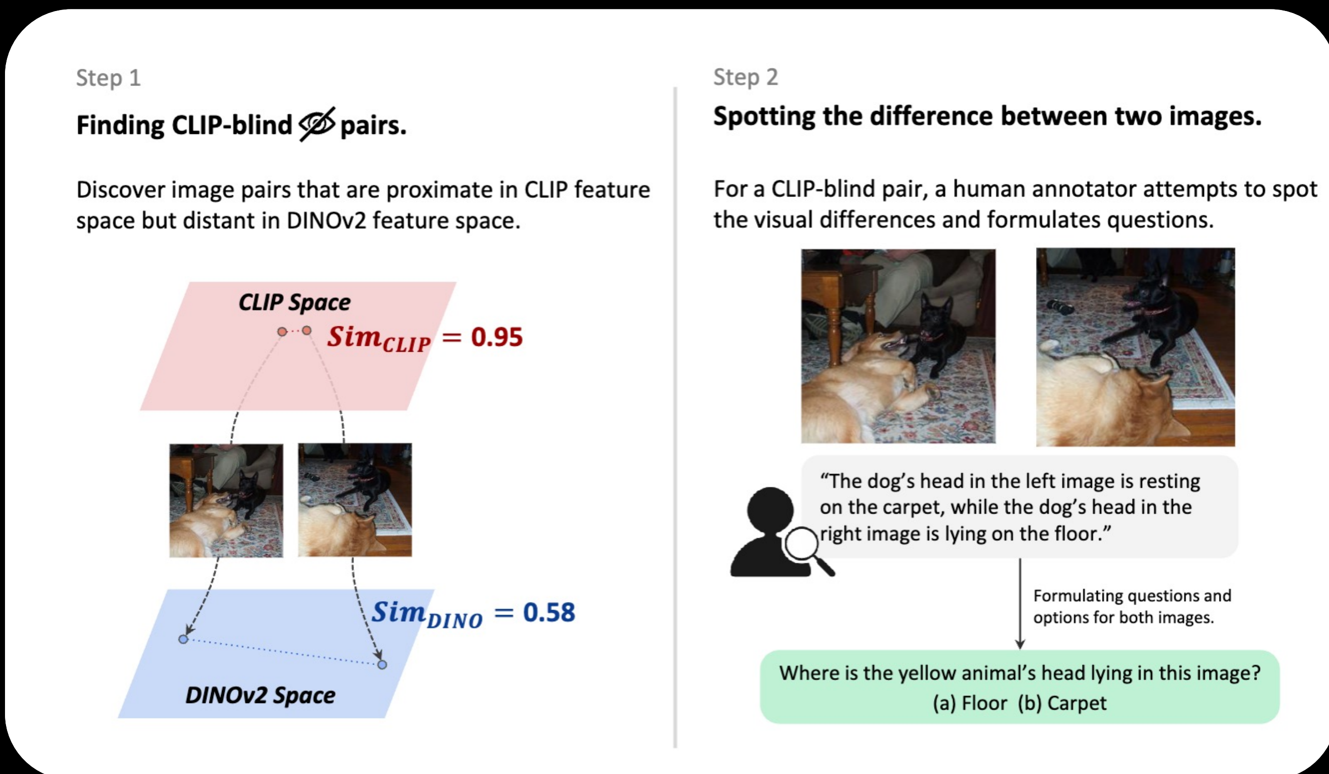
Step 1

Finding CLIP-blind pairs.

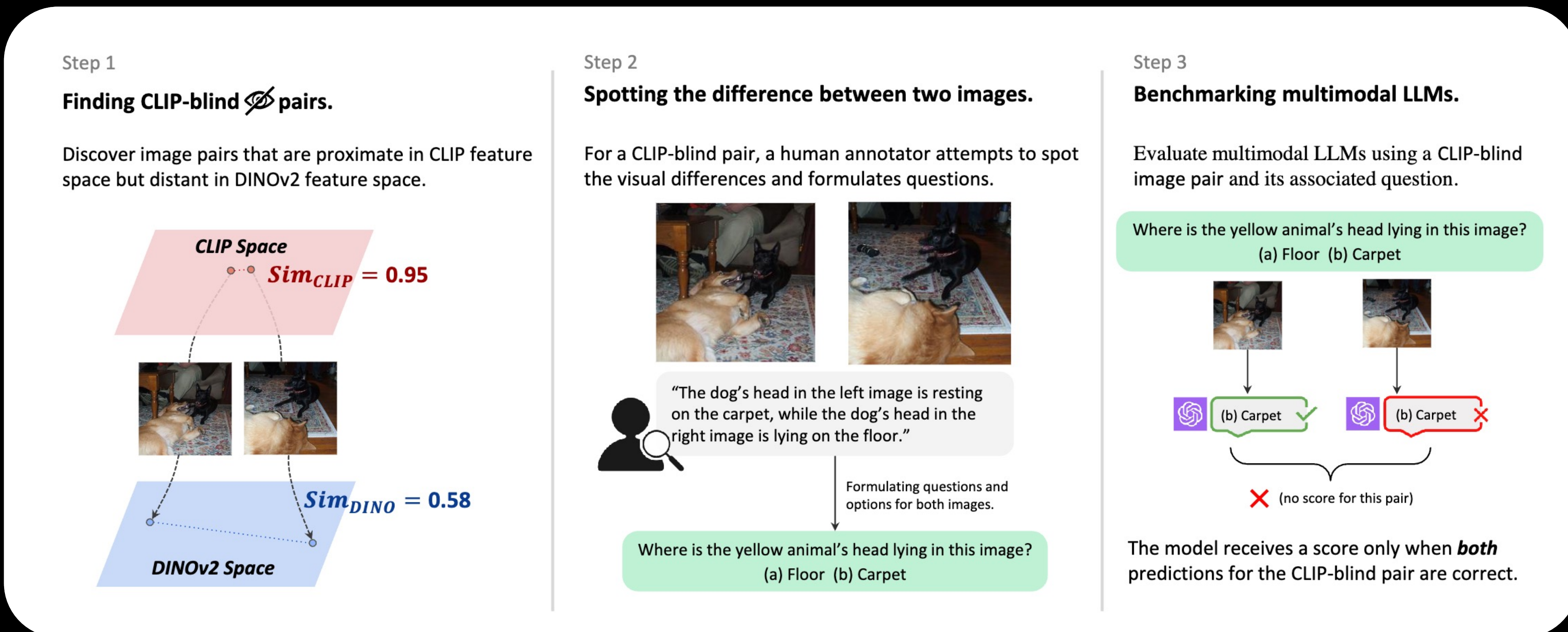
Discover image pairs that are proximate in CLIP feature space but distant in DINOv2 feature space.



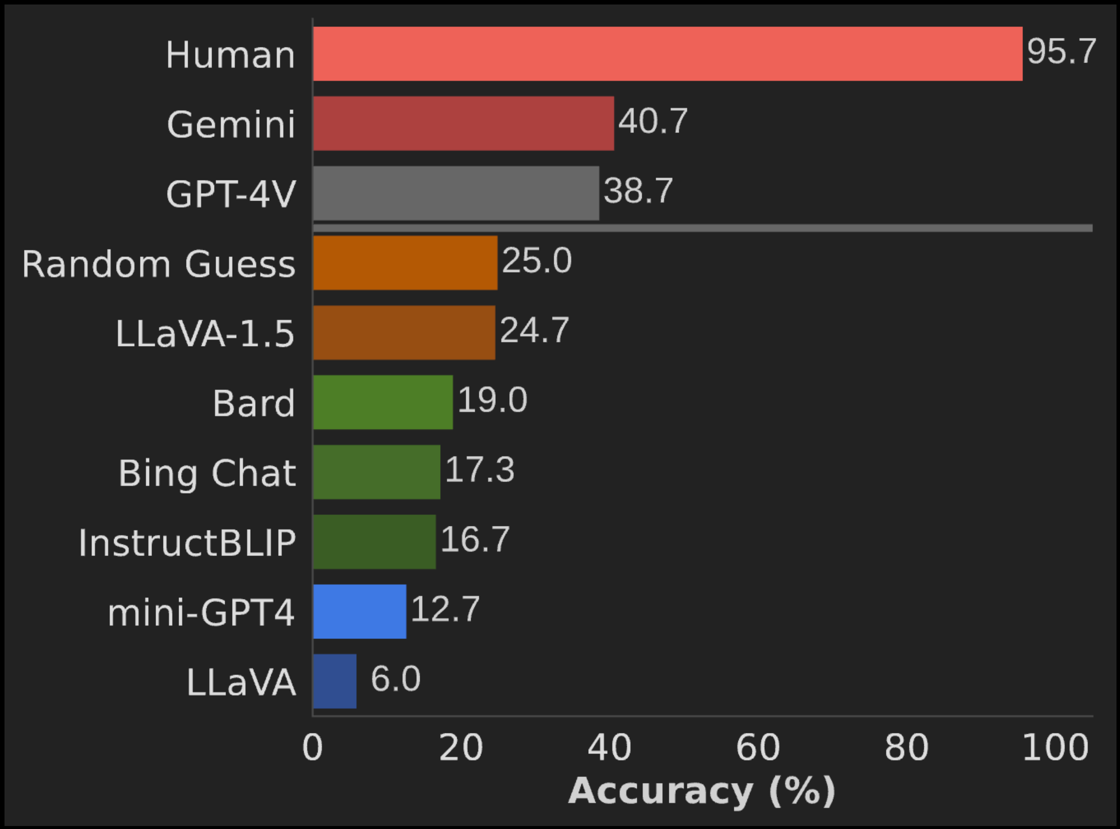
Constructing MMVP Benchmark via CLIP-blind Pairs

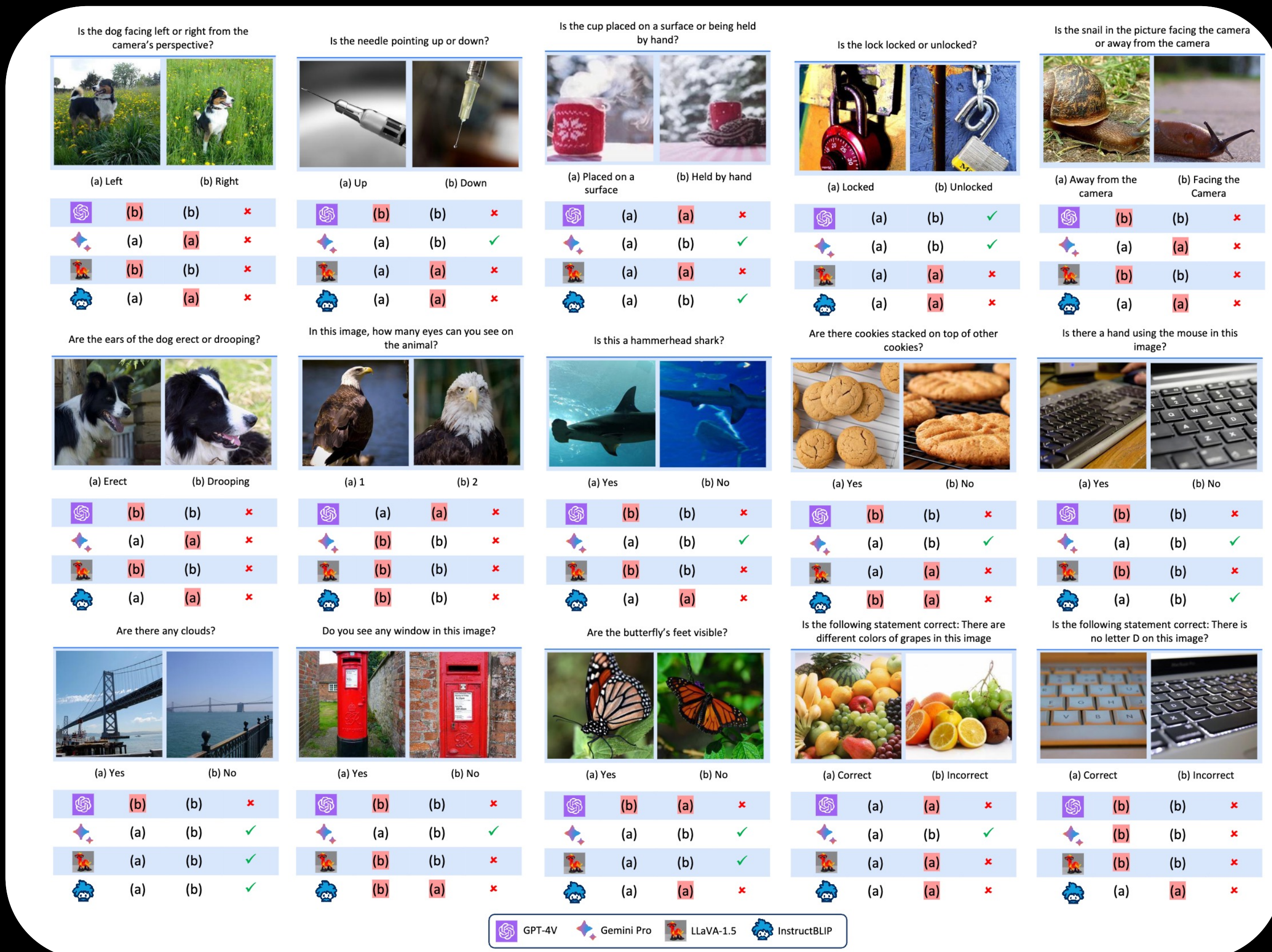


Constructing MMVP Benchmark via CLIP-blind Pairs



Results Of Current MLLM Models (And Humans)





Is the dog facing left or right from the camera's perspective?
(a) Left
(b) Right
(b) (b) ✗
(a) (a) ✗
(b) (b) ✗
(a) (a) ✗
Is the needle pointing up or down?
(a) Up
(b) Down
(b) (b) ✗
(a) (b) ✓
(a) (a) ✗
(a) (a) ✗
Is the cup placed on a surface or being held by hand?
(a) Placed on a surface
(b) Held by hand
(a) (a) ✗
(a) (b) ✓
(a) (a) ✗
(a) (b) ✓
Is the lock locked or unlocked?
(a) Locked
(b) Unlocked
(a) (b) ✓
(a) (b) ✓
(a) (a) ✗
(a) (a) ✗
Is the snail in the picture facing the camera or away from the camera
(a) Away from the camera
(b) Facing the Camera
(b) (b) ✗
(a) (a) ✗
(b) (b) ✗
(a) (a) ✗
Are the ears of the dog erect or drooping?
(a) Erect
(b) Drooping
(b) (b) ✗
(a) (a) ✗
(b) (b) ✗
(a) (a) ✗
In this image, how many eyes can you see on the animal?
(a) 1
(b) 2
(a) (a) ✗
(b) (b) ✗
(b) (b) ✗
(b) (b) ✗
Is this a hammerhead shark?
(a) Yes
(b) No
(b) (b) ✗
(a) (b) ✓
(b) (b) ✗
(a) (a) ✗
Are there cookies stacked on top of other cookies?
(a) Yes
(b) No
(b) (b) ✗
(a) (b) ✓
(a) (a) ✗
(b) (a) ✗
Is there a hand using the mouse in this image?
(a) Yes
(b) No
(b) (b) ✗
(a) (b) ✓
(b) (b) ✗
(a) (b) ✓
Are there any clouds?
(a) Yes
(b) No
(b) (b) ✗
(a) (b) ✓
(a) (b) ✓
(a) (b) ✓
Do you see any window in this image?
(a) Yes
(b) No
(b) (b) ✗
(a) (b) ✓
(b) (b) ✗
(b) (a) ✗
Are the butterfly's feet visible?
(a) Yes
(b) No
(b) (a) ✗
(a) (b) ✓
(a) (b) ✓
(a) (a) ✗
Is the following statement correct: There are different colors of grapes in this image
(a) Correct
(b) Incorrect
(a) (a) ✗
(a) (b) ✓
(a) (a) ✗
(a) (a) ✗
Is the following statement correct: There is no letter D on this image?
(a) Correct
(b) Incorrect
(b) (b) ✗
(b) (b) ✗
(b) (b) ✗
(a) (a) ✗
GPT-4V
Gemini Pro
LLaVA-1.5
InstructBLIP

So, what is going on?

User

I am analyzing an image embedding model. Can you go through the questions and options, trying to figure out some general patterns that the embedding model struggles with? Please focus on the visual features and generalize patterns that are important to vision models
[MMVP Questions and Options]

We identify 9 visual patterns:

- Orientation and Direction
- Presence of Specific Features
- State and Condition
- Quantity and Count
- Positional and Relational Context
- Color and Appearance
- Structural and Physical Characteristics
- Text
- Viewpoint and Perspective

Visual patterns in CLIP-blind pairs

Systematic Failures in CLIP

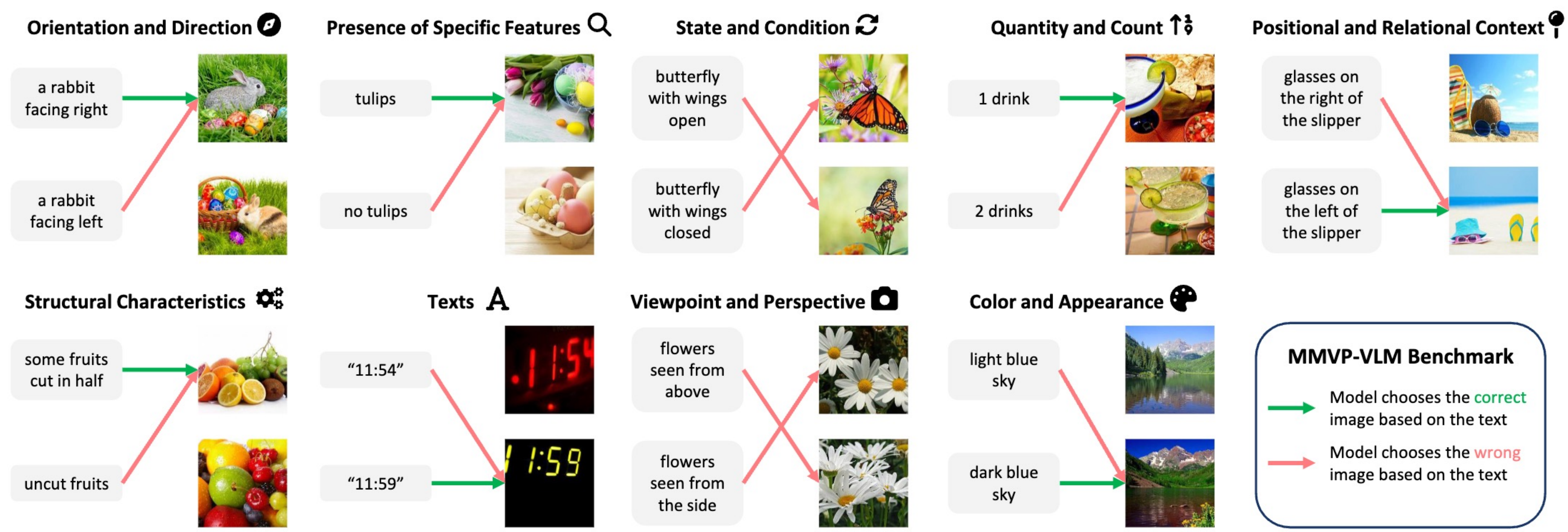


Figure 5. **Examples from MMVP-VLM**. MMVP-VLM consists of image pairs across nine visual patterns. The examples in the figure are from EVA01 ViT-g-14 model [54], one of the largest CLIP models that also fails to choose the right image given the text description.

Systematic Failures in CLIP-like models

	Image Size	Params (M)	IN-1k ZeroShot										MMVP Average
OpenAI ViT-L-14 [43]	224^2	427.6	75.5	13.3	13.3	20.0	20.0	13.3	53.3	20.0	6.7	13.3	19.3
OpenAI ViT-L-14 [43]	336^2	427.9	76.6	0.0	20.0	40.0	20.0	6.7	20.0	33.3	6.7	33.3	20.0
SigLIP ViT-SO-14 [66]	224^2	877.4	82.0	26.7	20.0	53.3	40.0	20.0	66.7	40.0	20.0	53.3	37.8
SigLIP ViT-SO-14 [66]	384^2	878.0	83.1	20.0	26.7	60.0	33.3	13.3	66.7	33.3	26.7	53.3	37.0
DFN ViT-H-14 [10]	224^2	986.1	83.4	20.0	26.7	73.3	26.7	26.7	66.7	46.7	13.3	53.3	39.3
DFN ViT-H-14 [10]	378^2	986.7	84.4	13.3	20.0	53.3	33.3	26.7	66.7	40.0	20.0	40.0	34.8
MetaCLIP ViT-L-14 [62]	224^2	427.6	79.2	13.3	6.7	66.7	6.7	33.3	46.7	20.0	6.7	13.3	23.7
MetaCLIP ViT-H-14 [62]	224^2	986.1	80.6	6.7	13.3	60.0	13.3	6.7	53.3	26.7	13.3	33.3	25.2
EVA01 ViT-g-14 [54]	224^2	1136.4	78.5	6.7	26.7	40.0	6.7	13.3	66.7	13.3	13.3	20.0	23.0
EVA02 ViT-bigE-14+ [54]	224^2	5044.9	82.0	13.3	20.0	66.7	26.7	26.7	66.7	26.7	20.0	33.3	33.3

CLIP Failures and MLLM Failures are Correlated

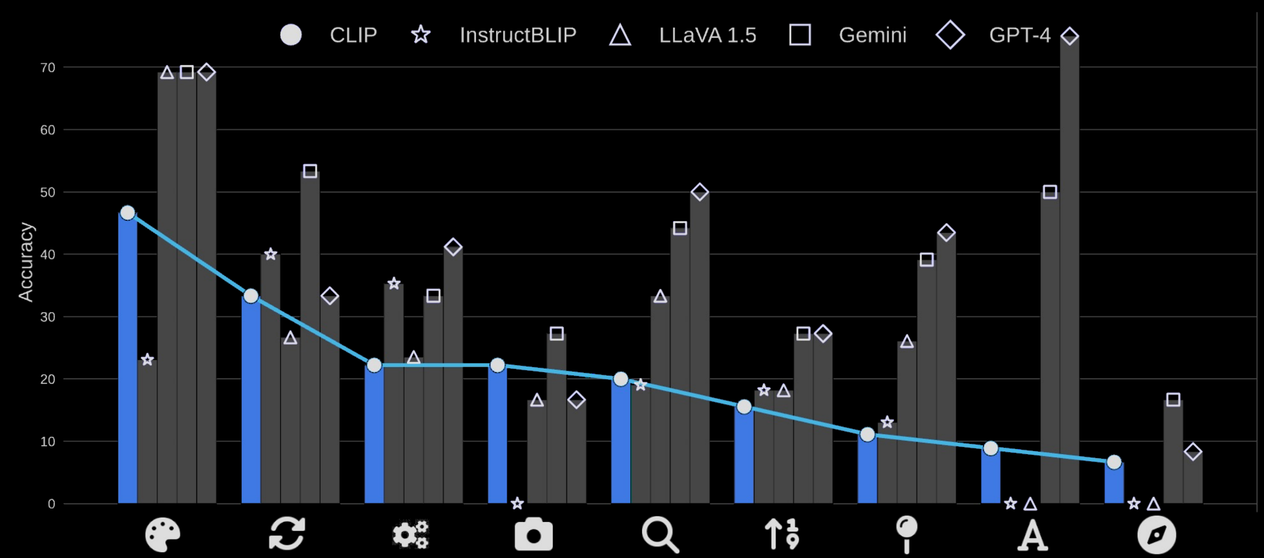
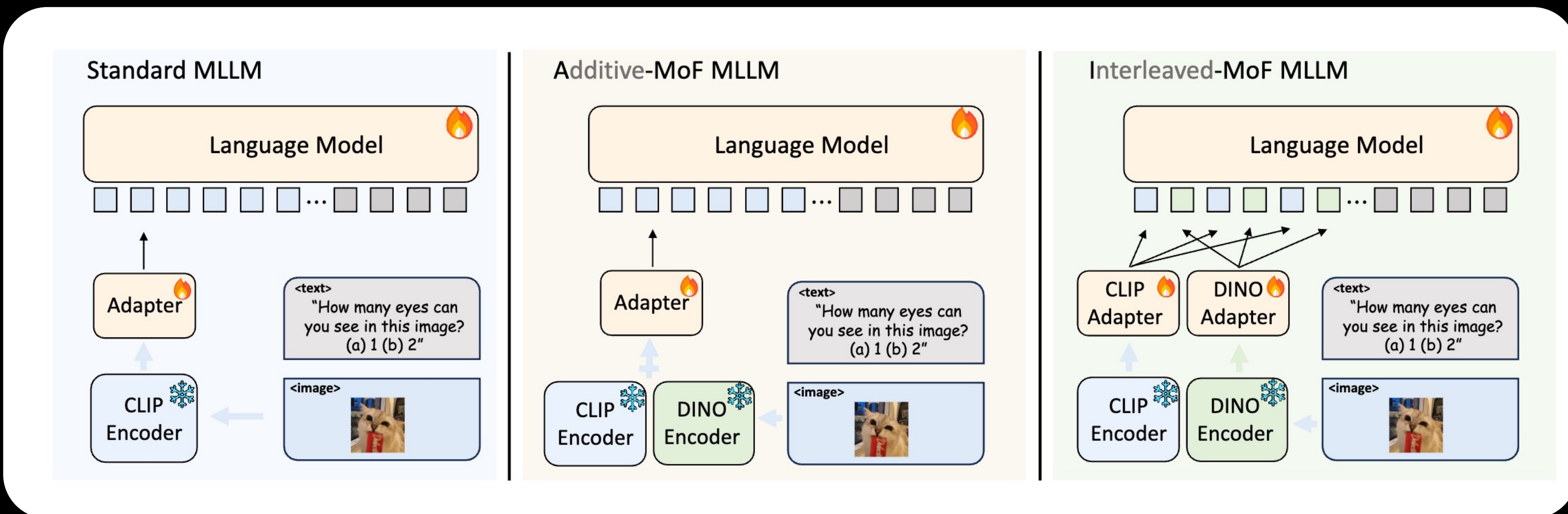


Figure 6. **CLIP and MLLM's performance on visual patterns.** If CLIP performs poorly on a visual pattern such as " orientation", MLLMs also underperform on the visual pattern.

Mixture-of-Features (MoF) for MLLM



Additive MoF

method	SSL ratio	MMVP	LLaVA
LLaVA	0.0	5.5	**81.8**
LLaVA + A-MoF	0.25	7.9 (+2.4)	79.4 (-2.4)
	0.5	12.0 (+6.5)	78.6 (-3.2)
	0.625	15.0 (+9.5)	76.4 (-5.4)
	0.75	**18.7** (+13.2)	75.8 (-6.0)
	0.875	16.5 (+11.0)	69.3 (-12.5)
	1.0	13.4 (+7.9)	68.5 (-13.3)

Table 2. **Empirical Results of Additive MoF.** We use DINOv2 as the image SSL model in our work. With more DINOv2 features added, there is an improvement in visual grounding, while a decline in instruction following ability.

Interleaved MoF

method	res	#tokens	MMVP	LLaVA	POPE
LLaVA	224^2	256	5.5	81.8	50.0
LLaVA	336^2	576	6.0	81.4	50.1
LLaVA + I-MoF	224^2	512	16.7 (+10.7)	82.8	51.0
LLaVA$^{1.5}$	336^2	576	24.7	84.7	85.9
LLaVA$^{1.5}$ + I-MoF	224^2	512	28.0 (+3.3)	82.7	86.3

Table 3. **Empirical Results of Interleaved MoF.** Interleaved MoF improves visual grounding while maintaining same level of instruction following ability.

Other SSL backbones can work too

method	SSL Model	res	#tokens	MMVP	POPE
LLaVA$^{1.5}$	None	336^2	576	24.7	85.9
LLaVA$^{1.5}$ + I-MoF	MoCov3	224^2	512	26.7 (+2.0)	86.1
LLaVA$^{1.5}$ + I-MoF	MAE	224^2	512	27.3 (+2.6)	86.1
LLaVA$^{1.5}$ + I-MoF	DINOv2	224^2	512	28.0 (+3.3)	86.3

Table 6. Results of Interleaved MoF with different vision-only SSL model

Takeaways

- Visual grounding is important for language understanding and meaning.

- CLIP's been lounging around for too long! (opportunities!)

- Vision SSL FTW!

  (but we need fundamentally different ways to pursue the problem.)

Key Question #2:

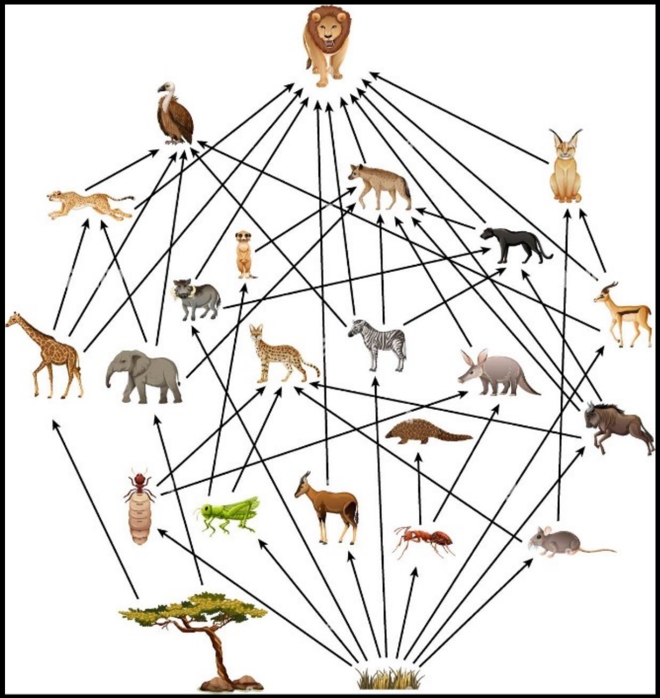
Better visual representations: beyond just static, global image feature extractors?

V^*: Guided Visual Search as a Core Mechanism in Multimodal LLMs

Penghao Wu, Saining Xie

CVPR 2024

“Deliberate” visual processing



A Concrete Example

plastic
straw

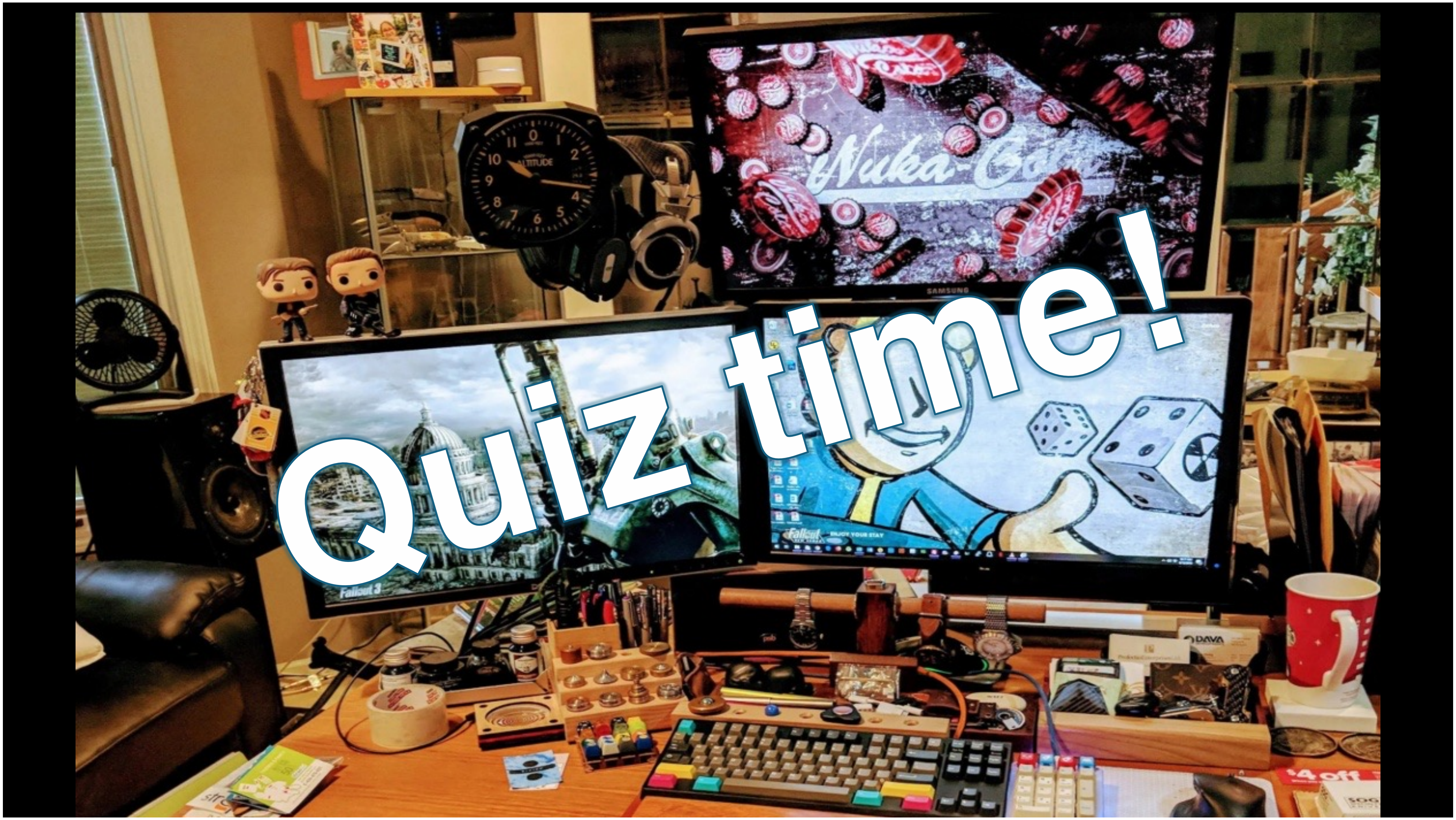
plastic
straw

plastic
straw

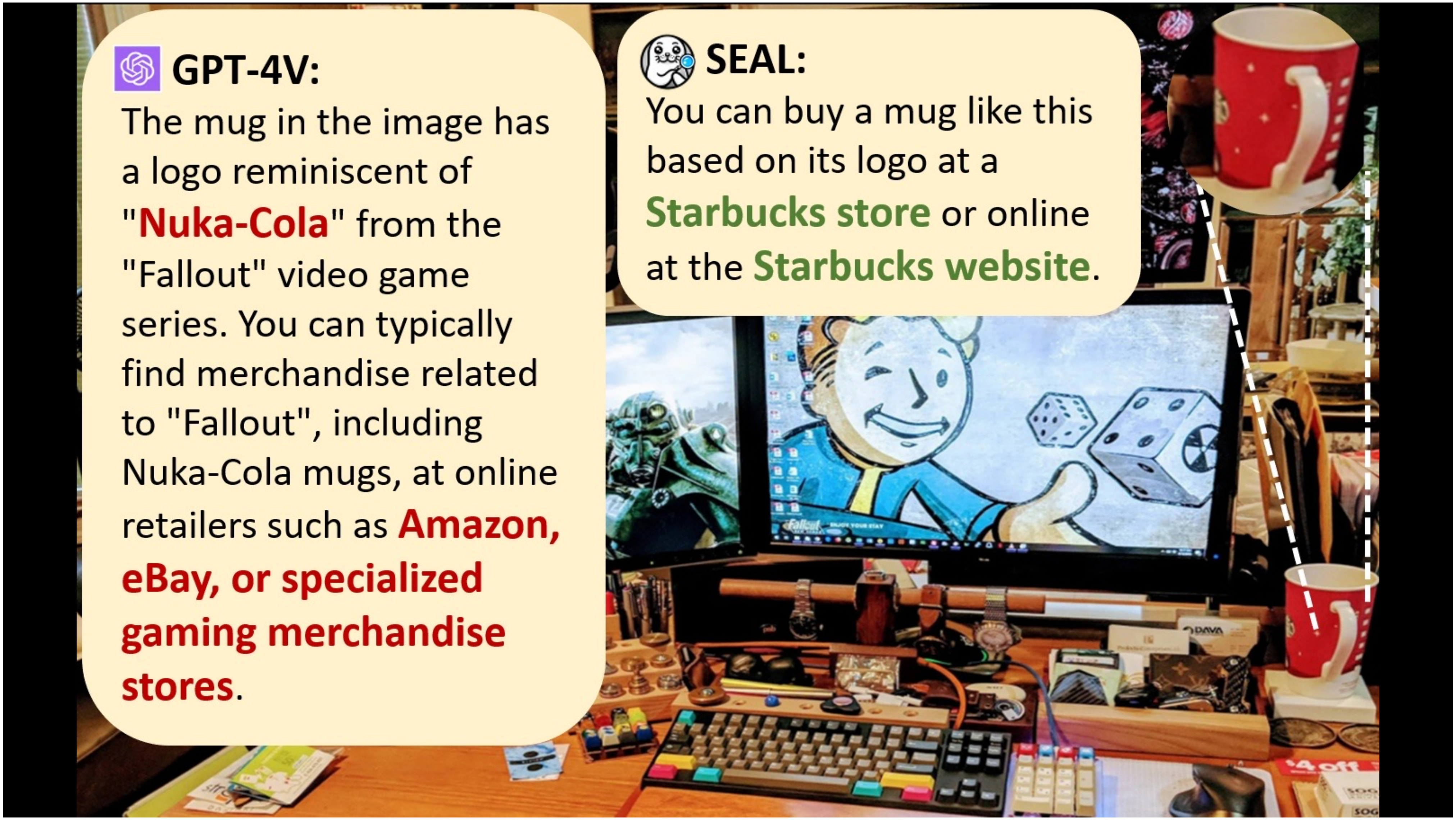
plastic
straw

plastic
straw

plastic
straw

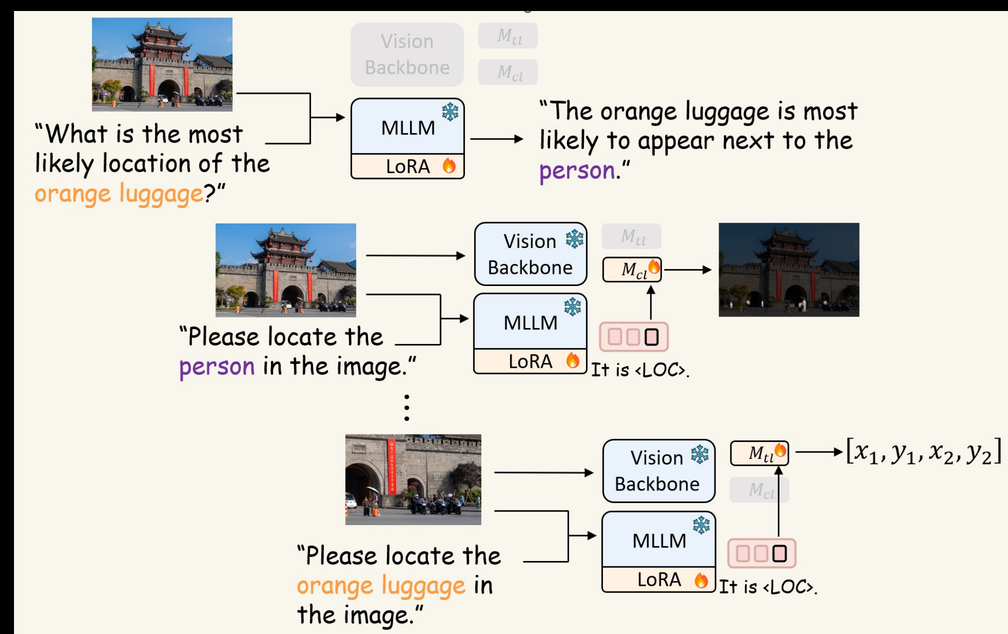

Quiz time!
Nuka-Cola
Fallout 3
ALTITUDE
SAMSUNG


DAVA


GPT-4V:
The mug in the image has a logo reminiscent of "Nuka-Cola" from the "Fallout" video game series. You can typically find merchandise related to "Fallout", including Nuka-Cola mugs, at online retailers such as Amazon, eBay, or specialized gaming merchandise stores.
SEAL:
You can buy a mug like this based on its logo at a Starbucks store or online at the Starbucks website.

Visual Search Model for MLLM

Is the orange luggage on the left side of the black luggage?



Visual Search Examples

Search for the **guitar**

Visual Search Examples

Contextual cue

The guitar is most likely to appear on the **stage**.

Visual Search Examples

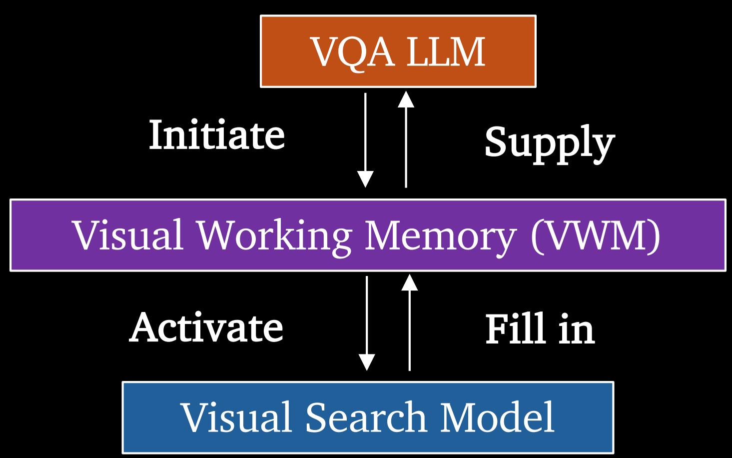
Target-specific cue

Visual Search Examples

Target Spotted!

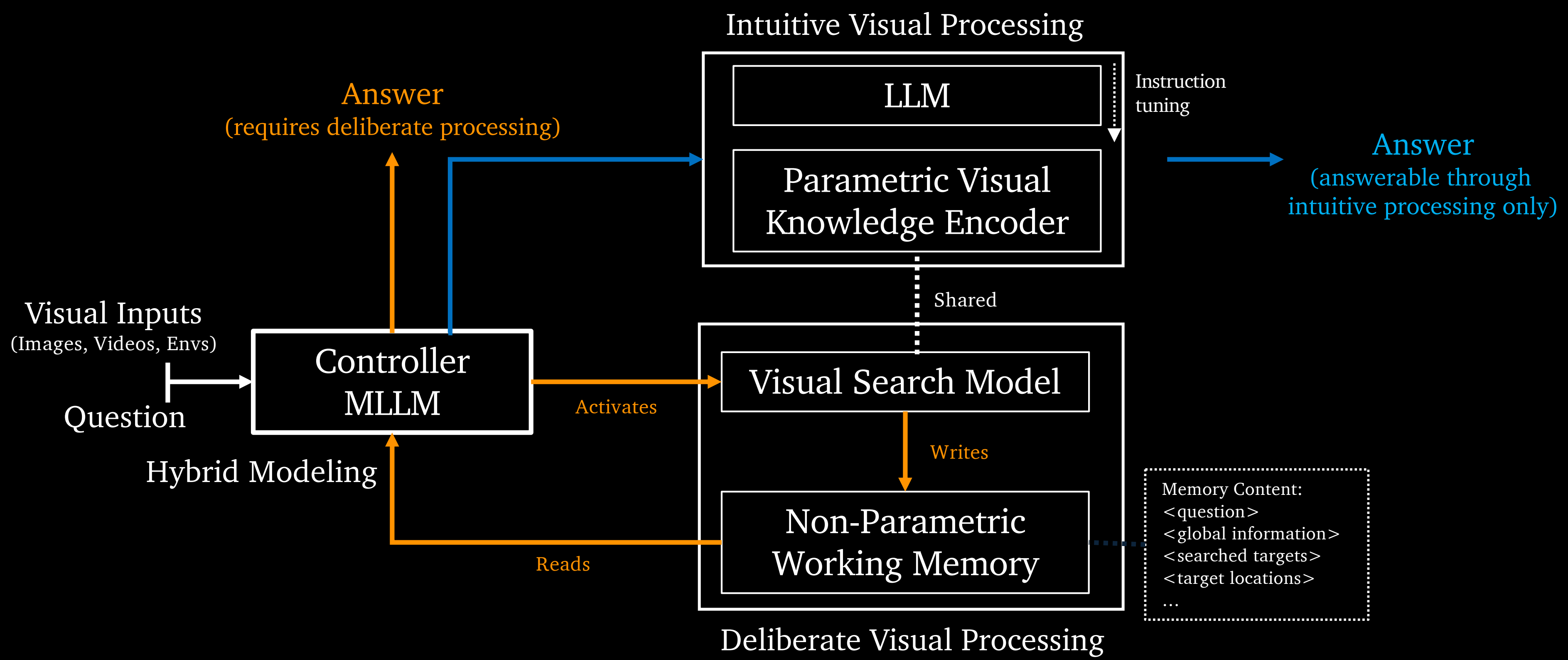
SEAL is a *Meta* Architecture for Multimodal LLMs

Show, s**EA**rch, and **T**el**L**



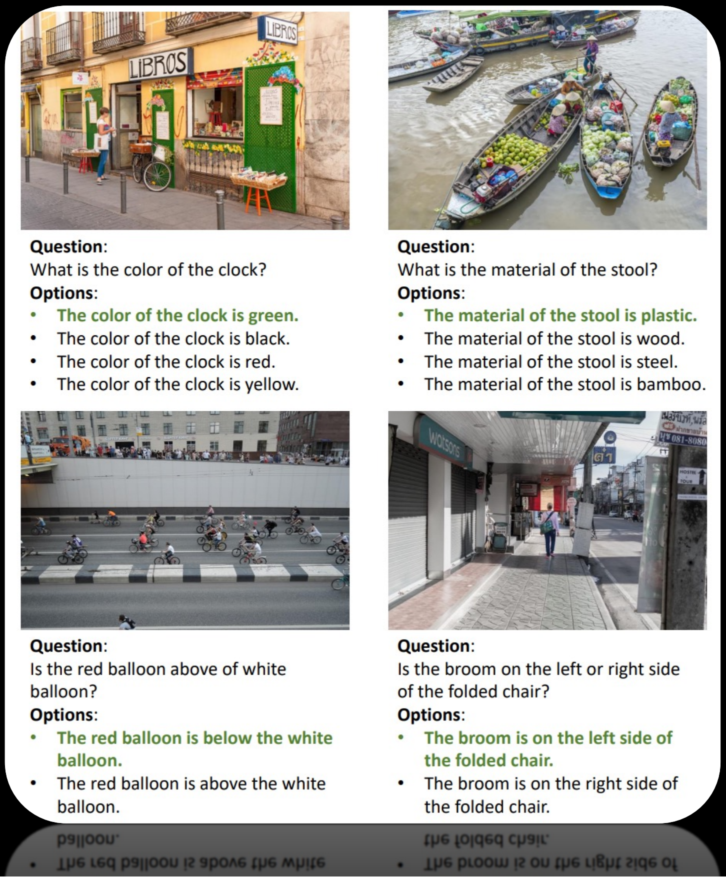
VWM:
<question>
<global image>
<searched targets>
<target locations>

An architecture for hybrid visual processing



V^*Bench – a Vision-centric VQA Benchmark

- 191 high-resolution images from SA-1B dataset
- Attribute recognition task (115 samples) & spatial relationship reasoning task (76 samples)
- Average image resolution 2246×1582, average target resolution 57×62
- Multiple-choice questions



V^*Bench – Evaluation Results of Multimodal Systems

	Attribute (%)	Spatial (%)	Overall (%)
Human	98.26	100.00	98.95
Random Guess	26.73	50.00	35.99
Open-source end-to-end MLLMs			
BLIP2 [23]	26.95	53.94	37.69
MiniGPT-4 [63]	30.43	50.00	38.22
LLaVA [28]	23.47	53.94	35.59
InstructBLIP [8]	25.21	47.36	34.02
Otter [22]	26.95	56.57	38.74
LLaVA-1.5 [27]	43.47	56.57	48.68
LLM tool-using pipelines			
MM-React [53]	34.78	51.31	41.36
VisualChatGPT [54]	30.43	48.68	37.69
Visprog [12]	31.30	56.57	41.36
Commercial chatbot systems			
Bard [11]	31.30	46.05	37.17
Gemini Pro [9]	40.86	59.21	48.16
GPT-4V [35]	51.30	60.52	54.97
SEAL (Ours)	**74.78**	**76.31**	**75.39**

Random Guess Level

Takeaways:

A good MLLM should be able to:

1. Acknowledge that initial visual information is NOT enough, and they cannot see.
2. Explicitly list additional visual information needed.
3. Understand and integrate the search results after the visual search process.
4. Allocate more FLOPS for more complex tasks during test time.
   (**After ChatGPT-o1🍓, I think people will be more open to this.)*

Note: these are not engineering hacks!

Key Question #3:

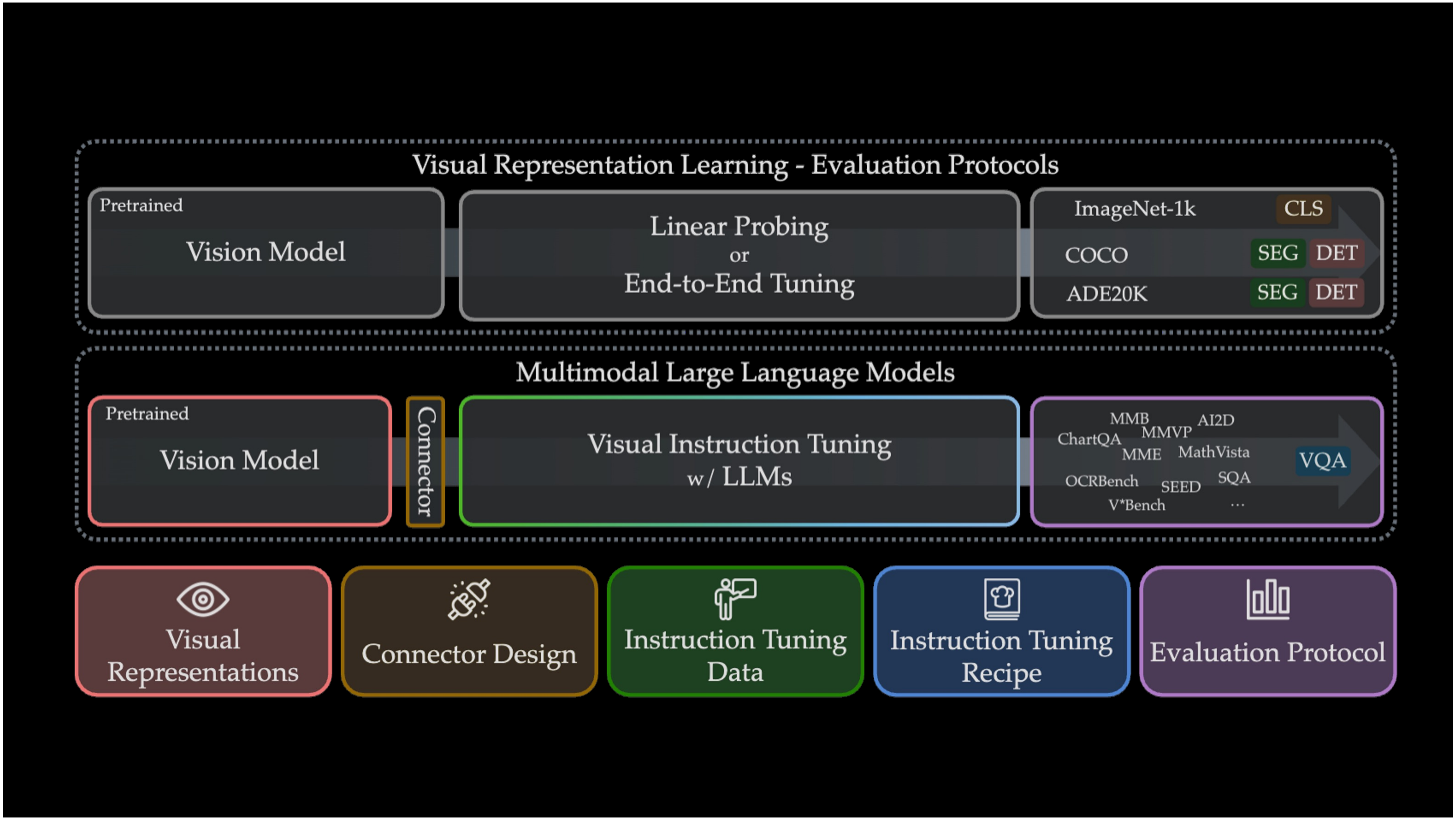
What are the necessary components towards building better, vision-centric MLLMs?

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

NeurIPS 2024 (Oral)

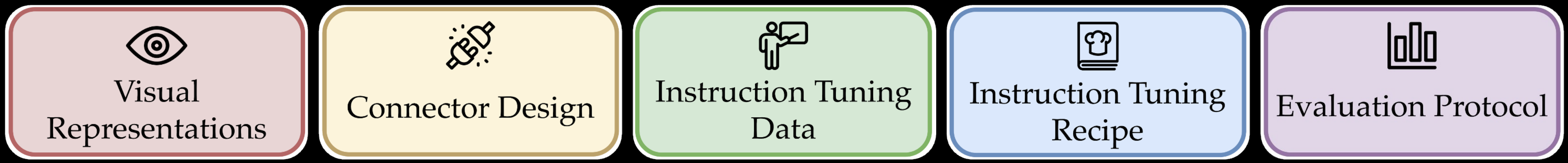

Visual Representation Learning - Evaluation Protocols
Pretrained
Vision Model
Linear Probing
or
End-to-End Tuning
ImageNet-1k
CLS
COCO
SEG
DET
ADE20K
SEG
DET
Multimodal Large Language Models
Pretrained
Vision Model
Connector
Visual Instruction Tuning
w/ LLMs
MMB
AI2D
ChartQA
MMVP
MME
MathVista
OCRBench
SEED
SQA
V*Bench
...
VQA
Visual Representations
Connector Design
Instruction Tuning Data
Instruction Tuning Recipe
Evaluation Protocol

Overview: Core Components in MLLMs

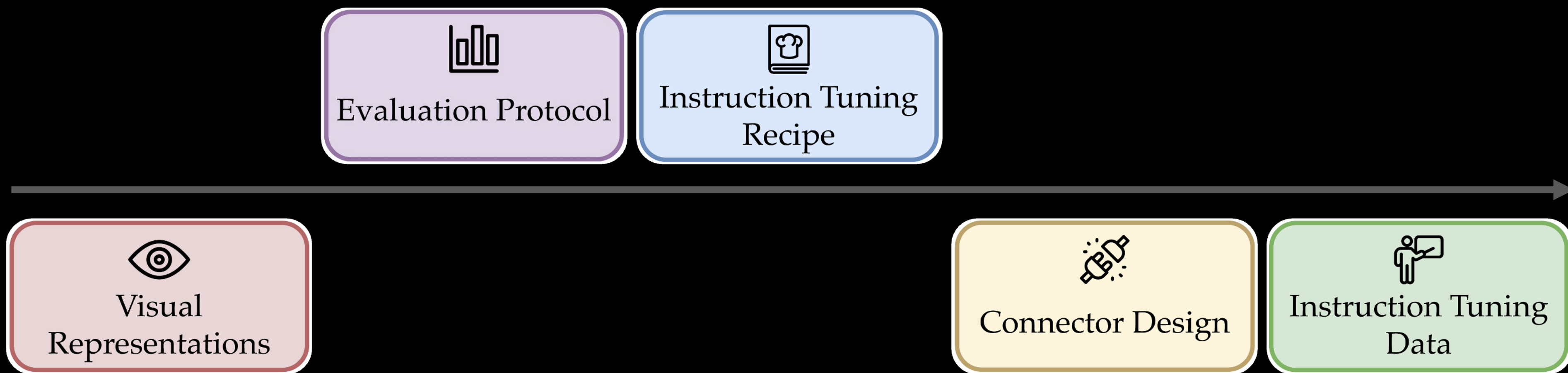
- LLM
- Vision Backbone
- Vision-Language Connector
- Instruction Tuning Data
- Instruction Tuning Recipe
- Evaluation Protocol

Overview: Core Components in MLLMs

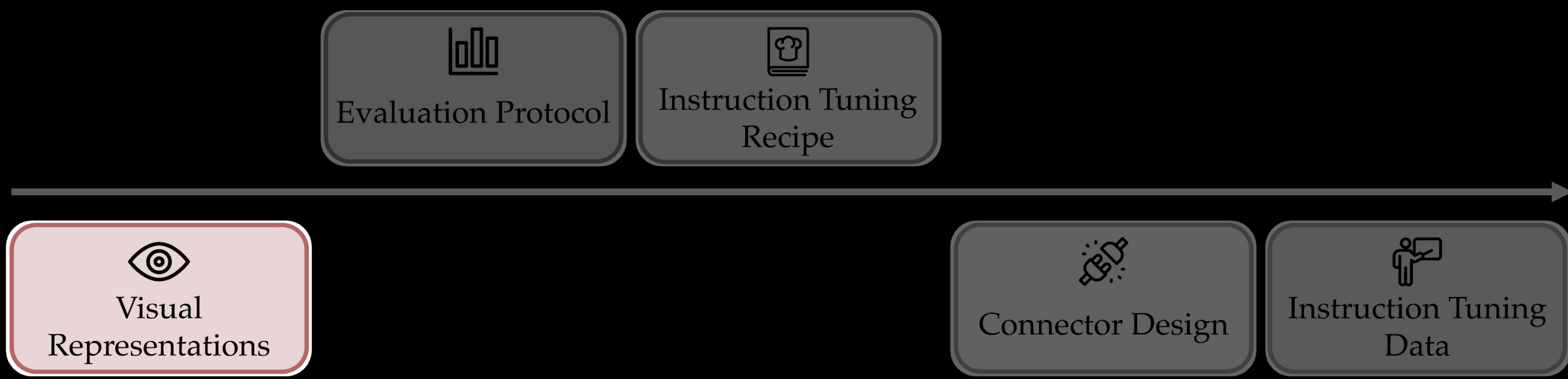
- ~~LLM~~
- Vision Backbone
- Vision-Language Connector
- Instruction Tuning Data
- Instruction Tuning Recipe
- Evaluation Protocol



Overview: Core Components in MLLMs



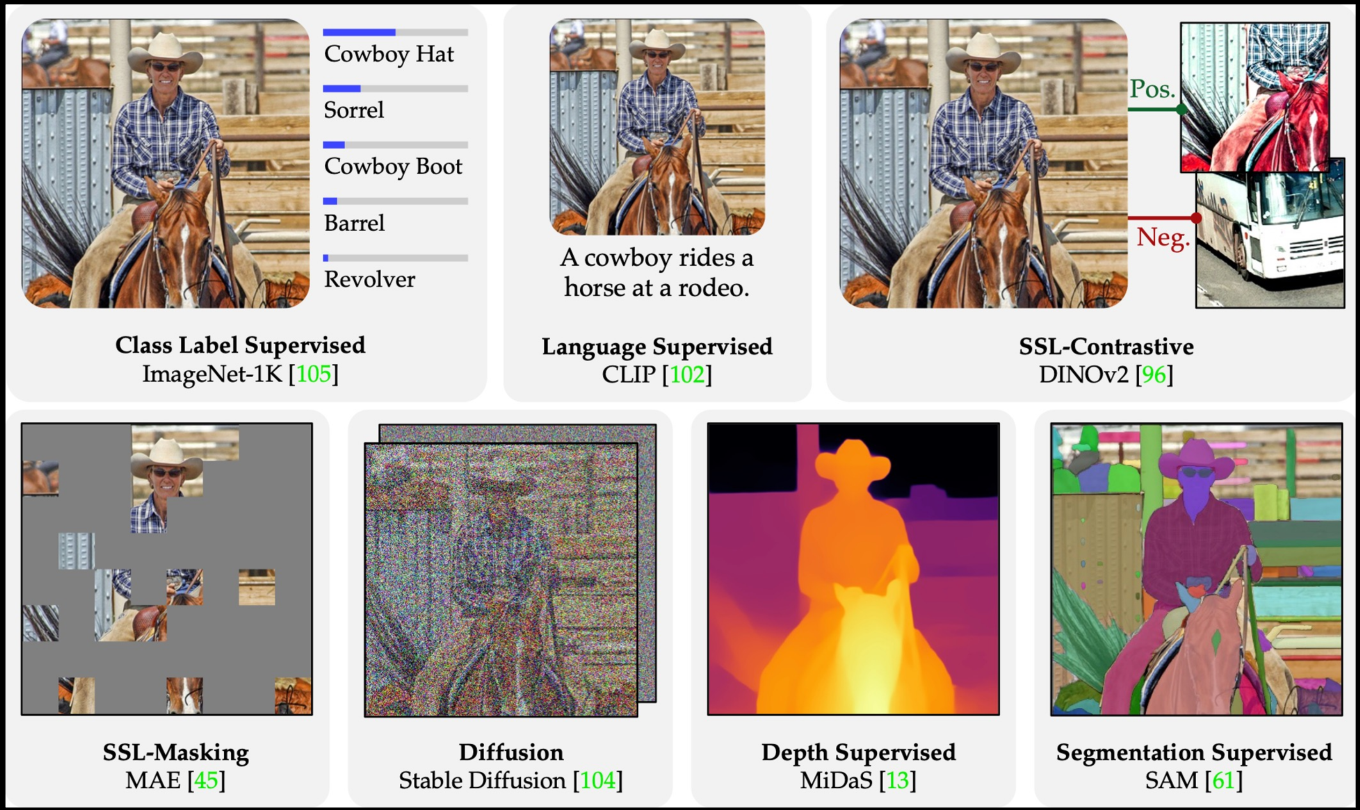
Overview


Evaluation Protocol
Instruction Tuning Recipe
Visual Representations
Connector Design
Instruction Tuning Data

Visual Representations for MLLMs

How to evaluate visual reprs.?

What visual reprs. to use?



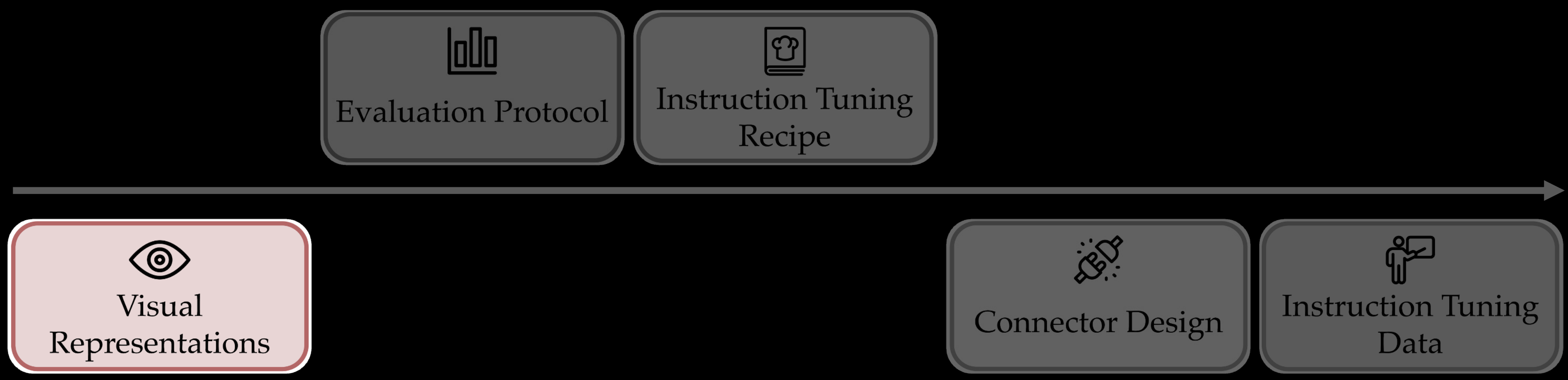
Visual Representations for MLLMs

Supervision Type	Method	Architecture	Patch Size	Res.	# Tok.	Hidden Size
Language-Supervised						
Language	OpenAI CLIP	ViT-L	14	336	576	768
	DFN-CLIP	ViT-L	14	224	256	1024
	DFN-CLIP	ViT-H	14	378	729	1280
	EVA-CLIP-02	ViT-L	14	336	576	1024
	SigLIP	ViT-L	16	384	576	1024
	SigLIP	ViT-SO400M	14	384	729	1152
	OpenCLIP	ConvNeXT-L	-	512	[I]576	1536
	OpenCLIP	ConvNeXT-L	-	1024	[I]576	1536
	OpenCLIP	ConvNeXT-XXL	-	1024	[I]576	3072
Self-Supervised						
Contrastive	DINOv2	ViT-L	14	336	576	1024
	DINOv2	ViT-L	14	518	[I]576	1024
	MoCo v3	ViT-B	16	224	196	768
	MoCo v3	ViT-L	16	224	196	1024
Masked	MAE	ViT-L	16	224	196	1024
	MAE	ViT-H	14	224	256	1280
JEPA	I-JEPA	ViT-H	14	224	256	1280
Other						
Segmentation	SAM	ViT-L	16	1024	[I]576	1024
	SAM	ViT-L	16	1024	[I]576	1280
Depth	MiDaS 3.0	ViT-L	16	384	576	1024
	MiDaS 3.1	ViT-L	16	518	1024	1024
Diffusion	Stable Diffusion 2.1	VAE+UNet	16	512	1024	3520
Class Labels	SupViT	ViT-L	16	224	196	1024
	SupViT	ViT-H	14	224	256	1280

Table 9 | Catalog of all vision backbones tested. [I] denotes that the visual tokens have been interpolated down to the specified length.

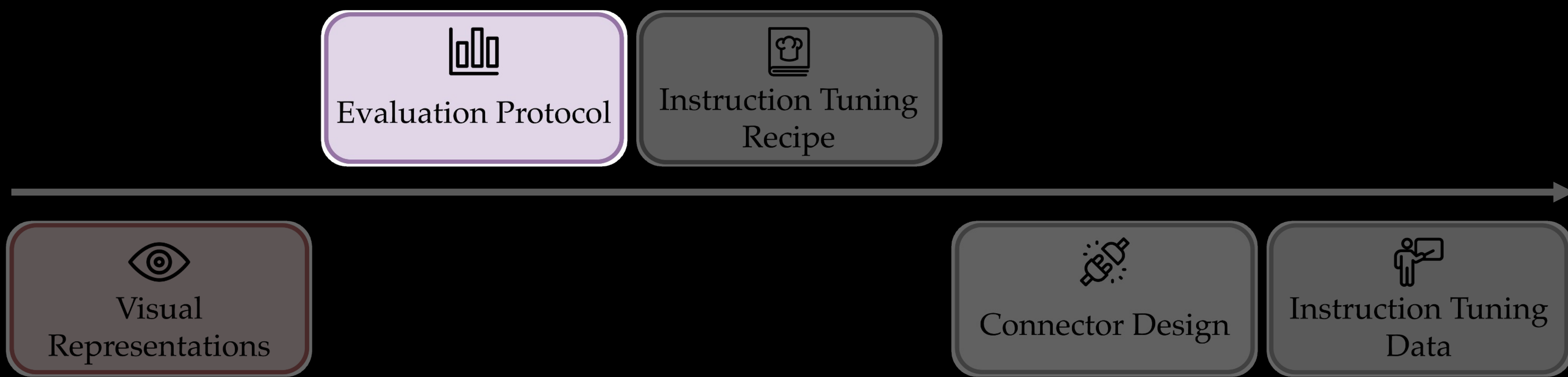
$\left.\right\}$ 23 models!

Overview

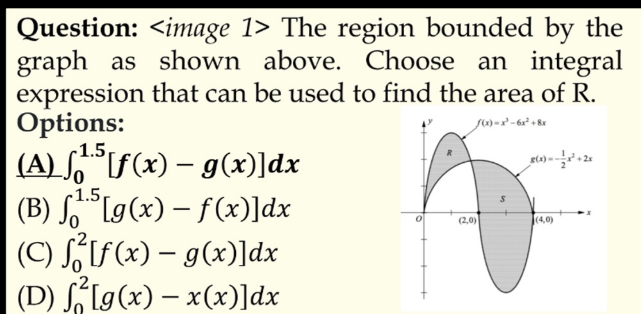


Evaluation Protocol
Instruction Tuning Recipe
Visual Representations
Connector Design
Instruction Tuning Data

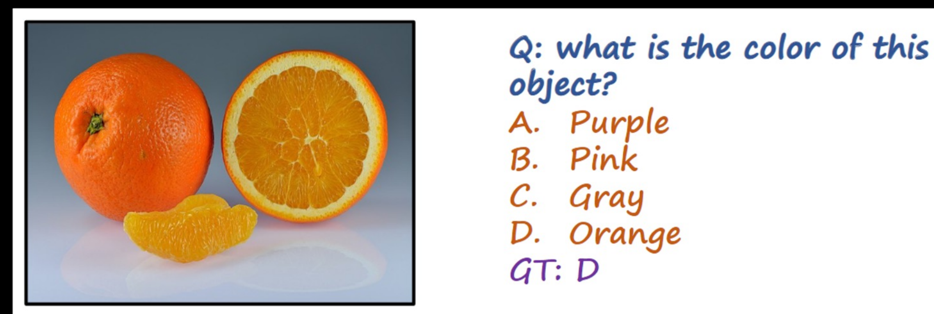
Overview


Evaluation Protocol
Instruction Tuning Recipe
Visual Representations
Connector Design
Instruction Tuning Data

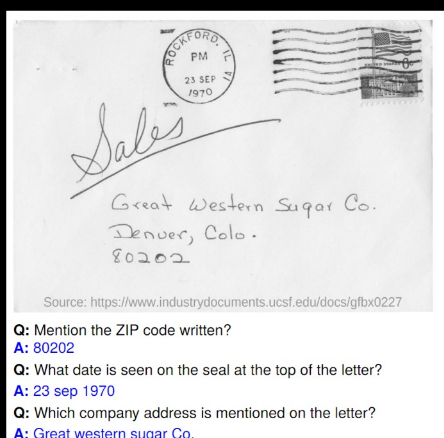
Evaluation Protocol



MMMU [Yue, et al. 2024]

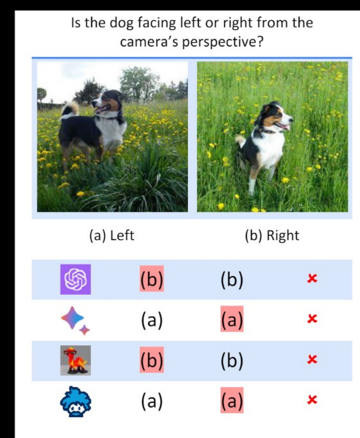


MM-Bench [Liu, et al. 2024]



DocVQA [Mathew, et al. 2020]

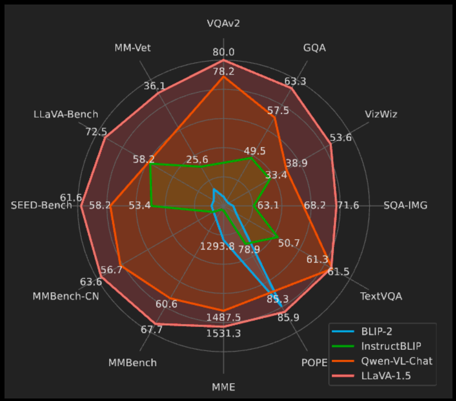
RealWorldQA [Grok, et al. 2024]



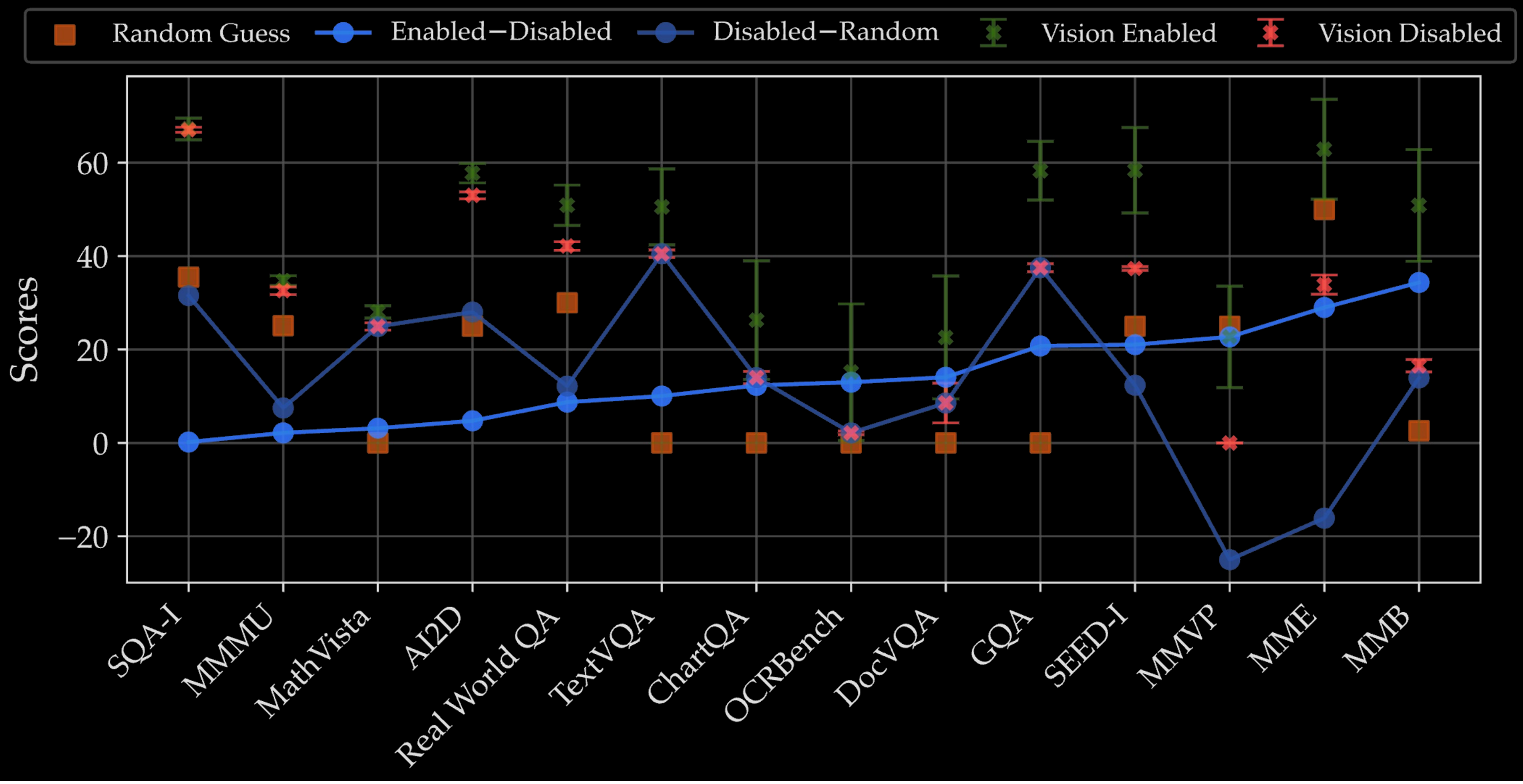
MMVP [Tong, et al. 2024]

and a lot more…

How should we systematically evaluate an MLLM and interpret the evaluation results?

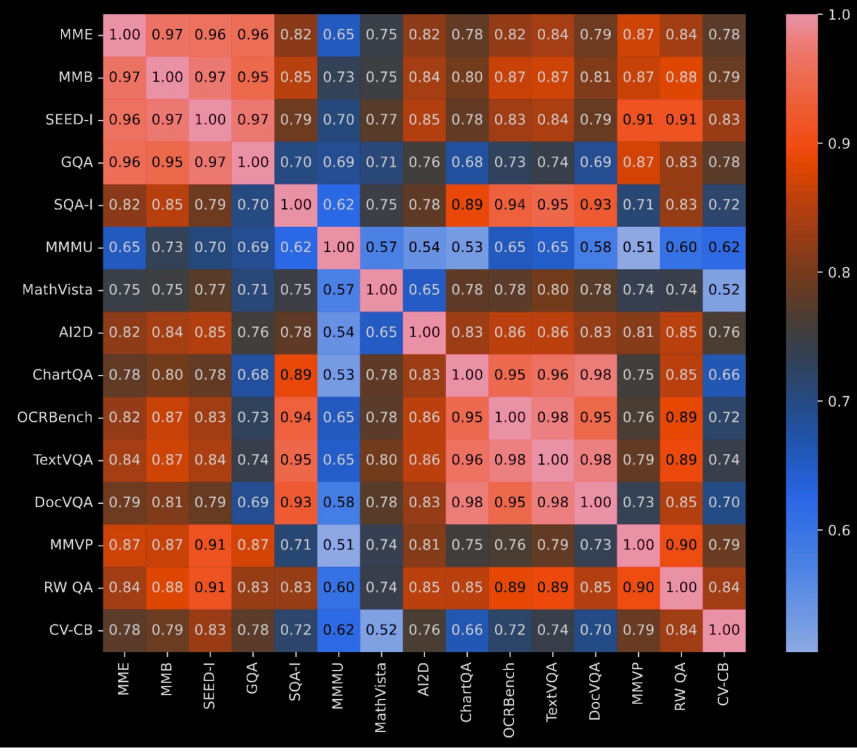


Who's answering the question: the **LLM** or **MLLM**?

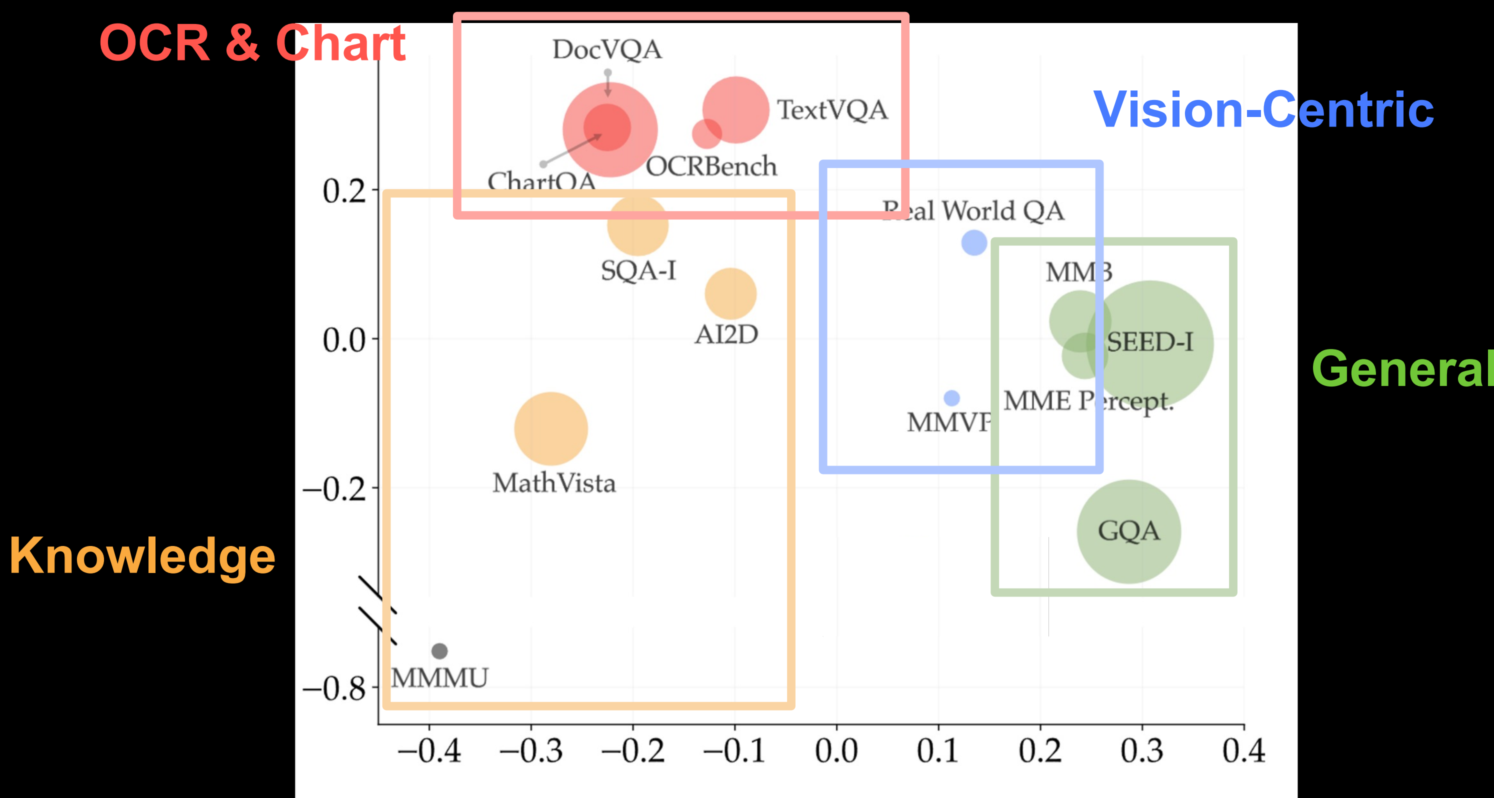


Group Benchmarks by Correlation

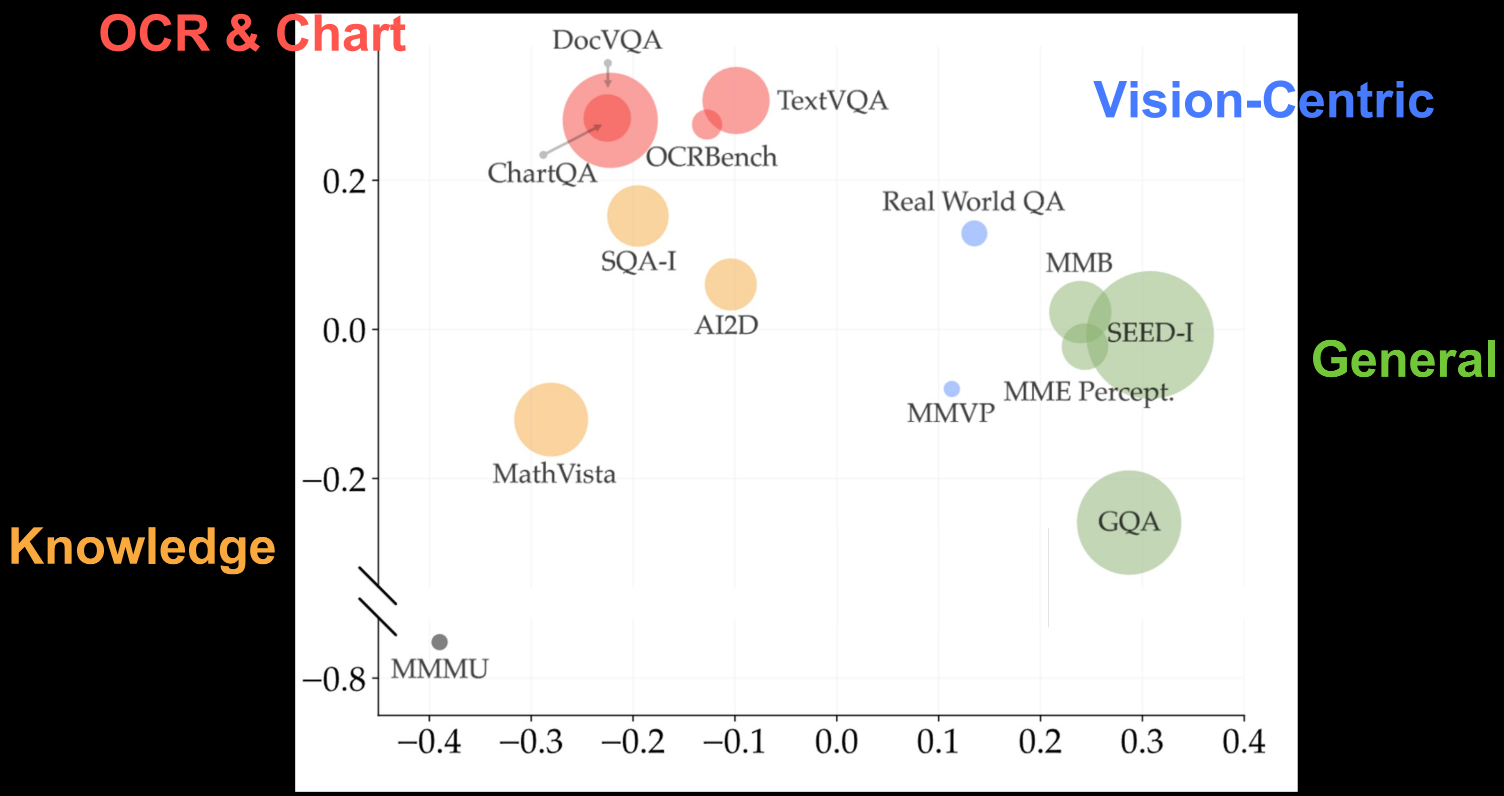
If two benchmarks evaluate on similar domains, they should have a strong correlation



Group Benchmarks by Correlation

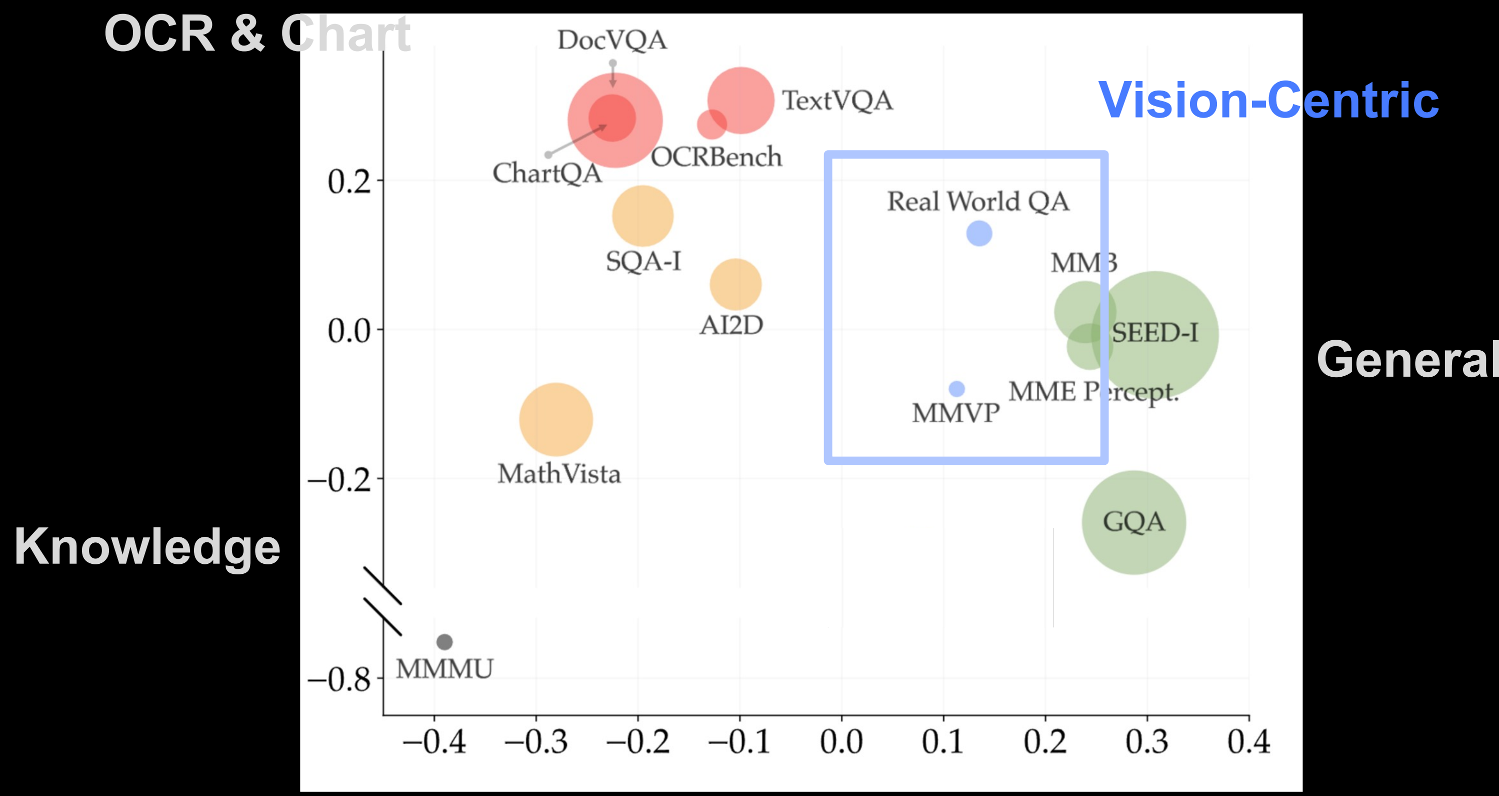


Group Benchmarks by Correlation



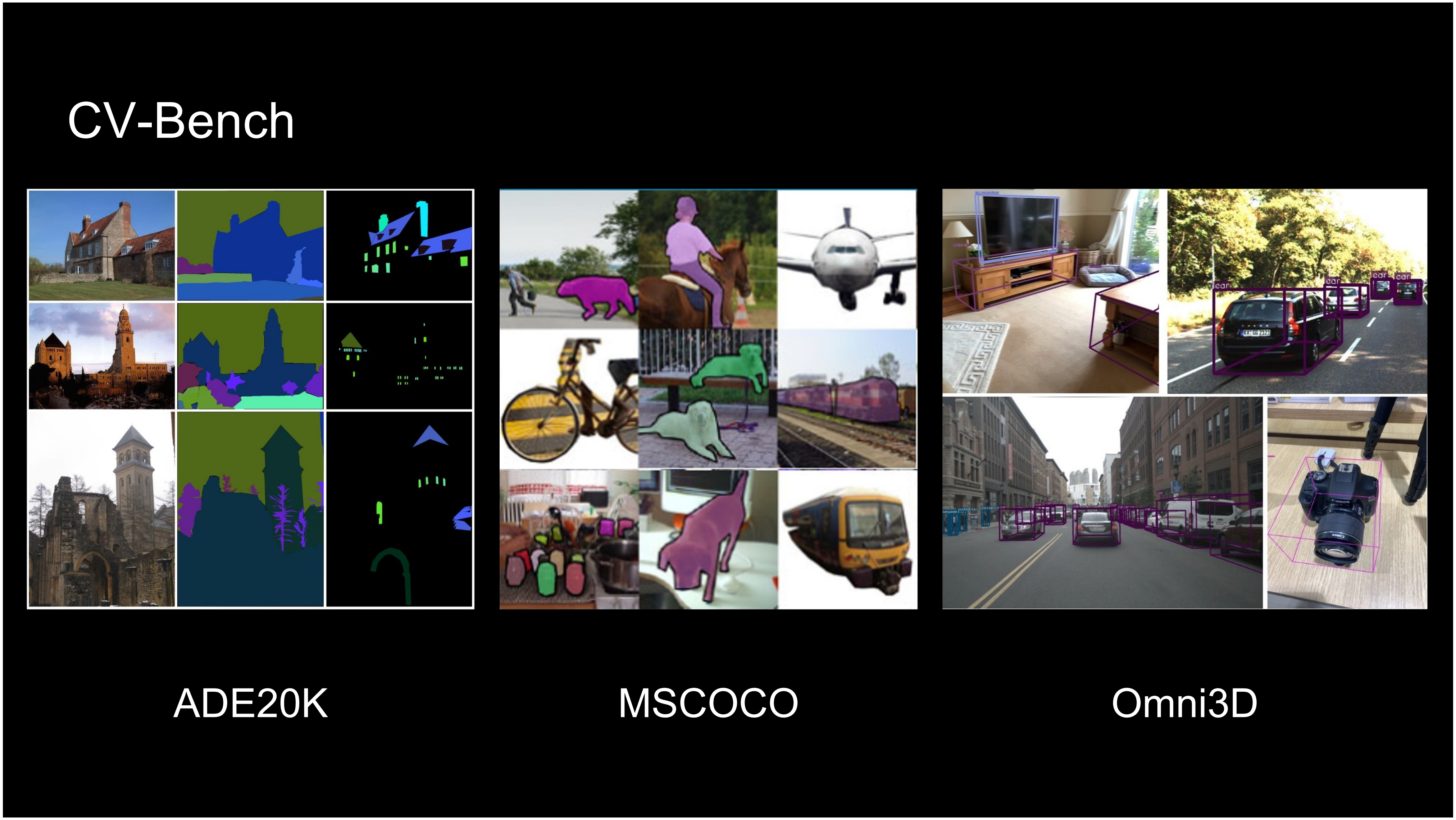
Group Benchmarks by Correlation

Tiny compared to others!



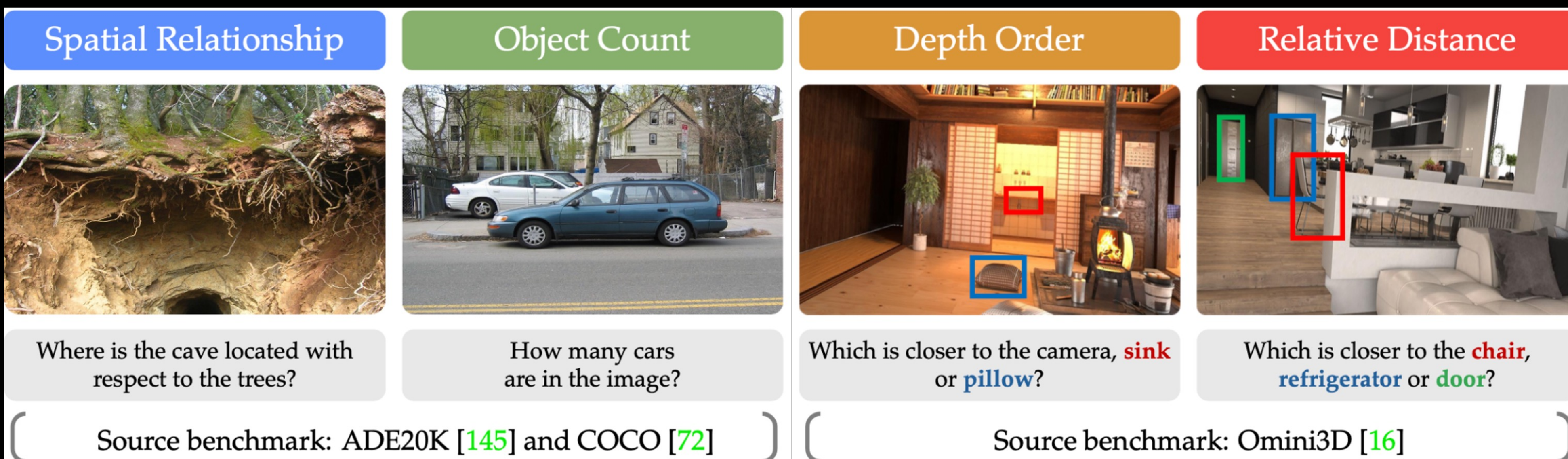
Q: How can we scalably generate ***vision-centric*** MLLM evaluations?

💡 Repurpose existing vision benchmarks!


CV-Bench
ADE20K
MSCOCO
Omni3D

CV-Bench

2D ***3D***

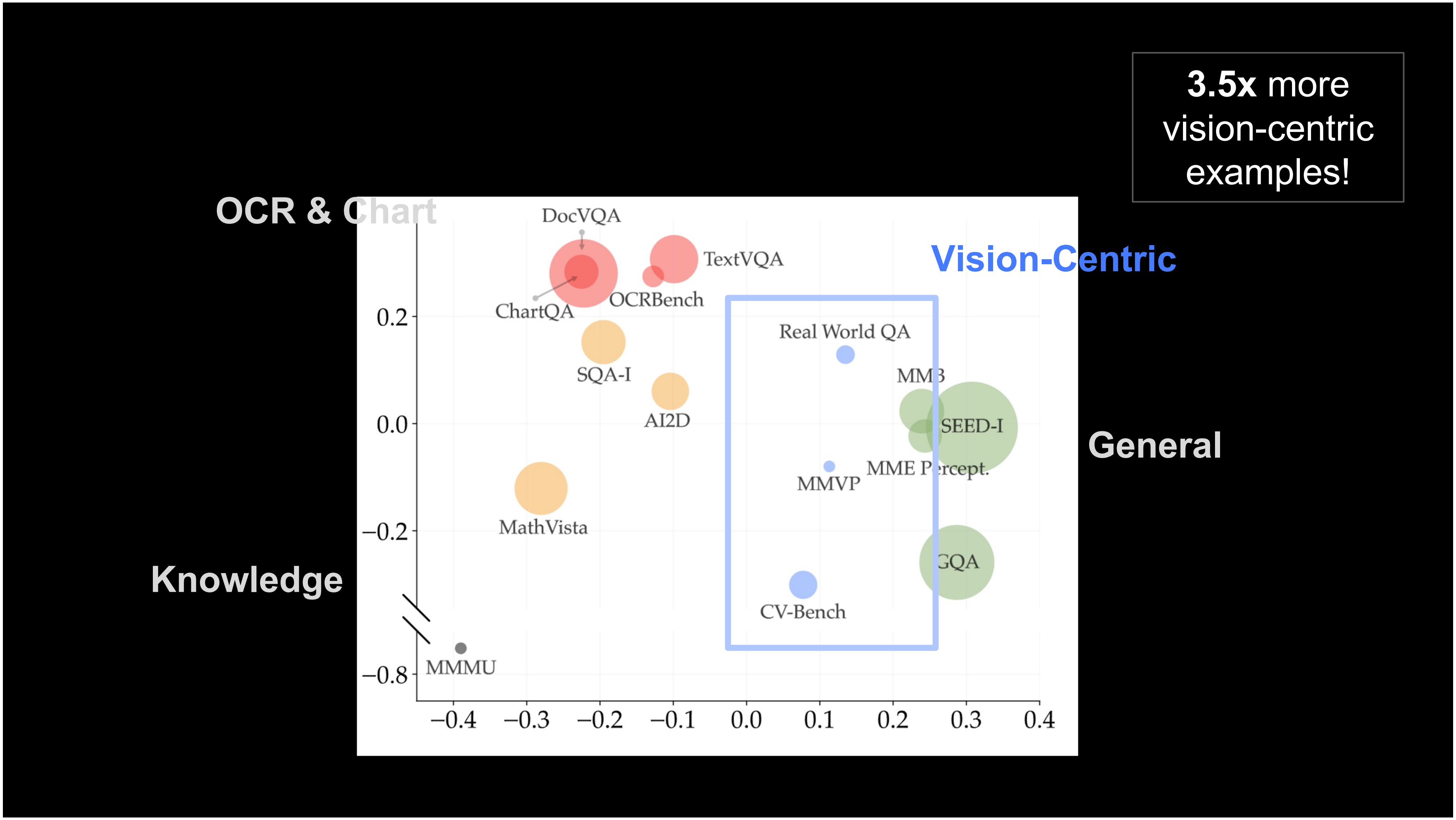


CV-Bench

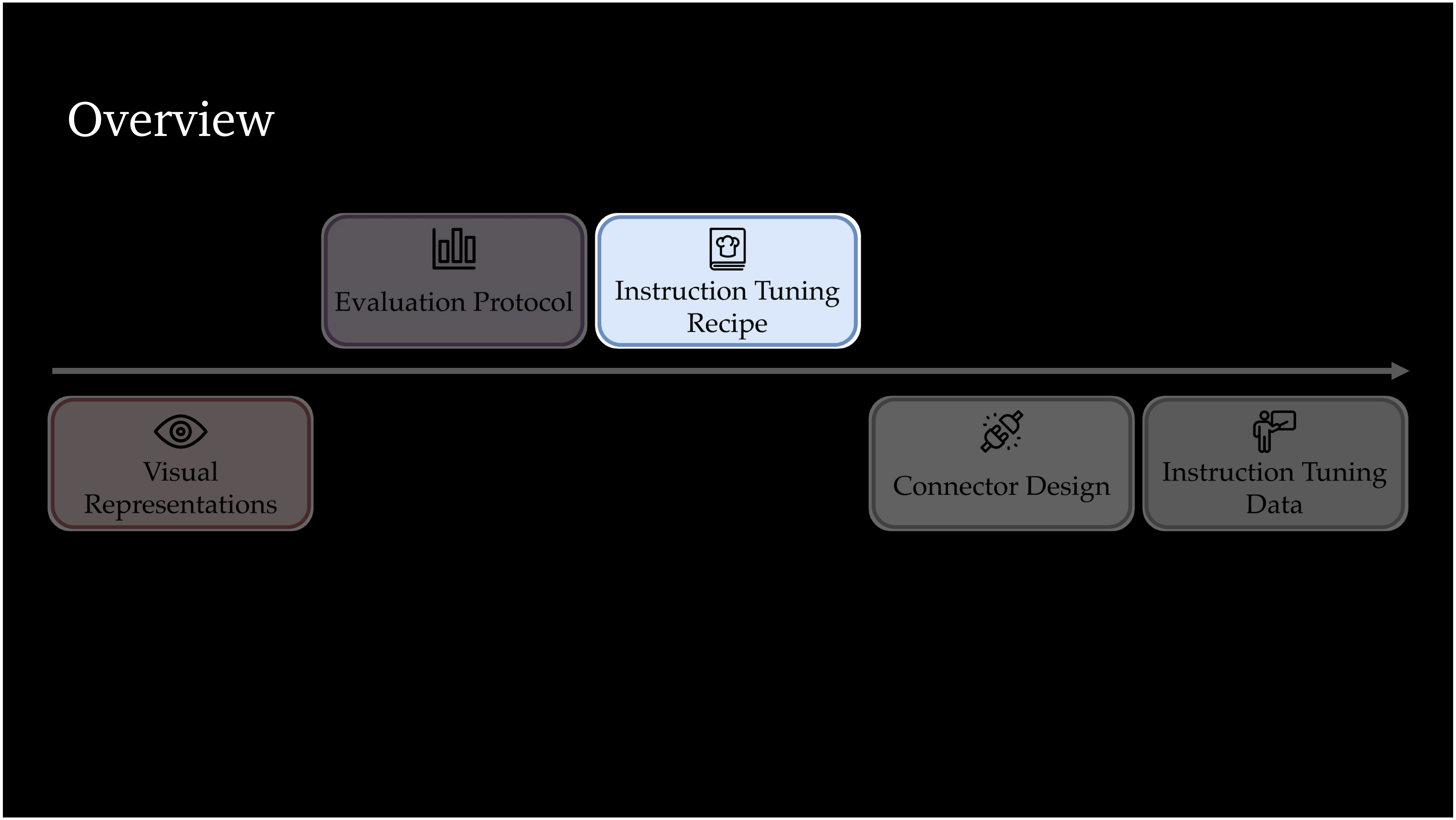
2,638 manually-inspected examples

Type	Task	Description	Sources	# Samples
2D	**Spatial Relationship**	Determine the relative position of an object w.r.t. the anchor object. Consider left-right or top-bottom relationship.	ADE20K COCO	650
	Object Count	Determine the number of instances present in the image.	ADE20K COCO	788
3D	**Depth Order**	Determine which of the two distinct objects is closer to the camera.	Omni3D	600
	Relative Distance	Determine which of the two distinct objects is closer to the anchor object.	Omni3D	600

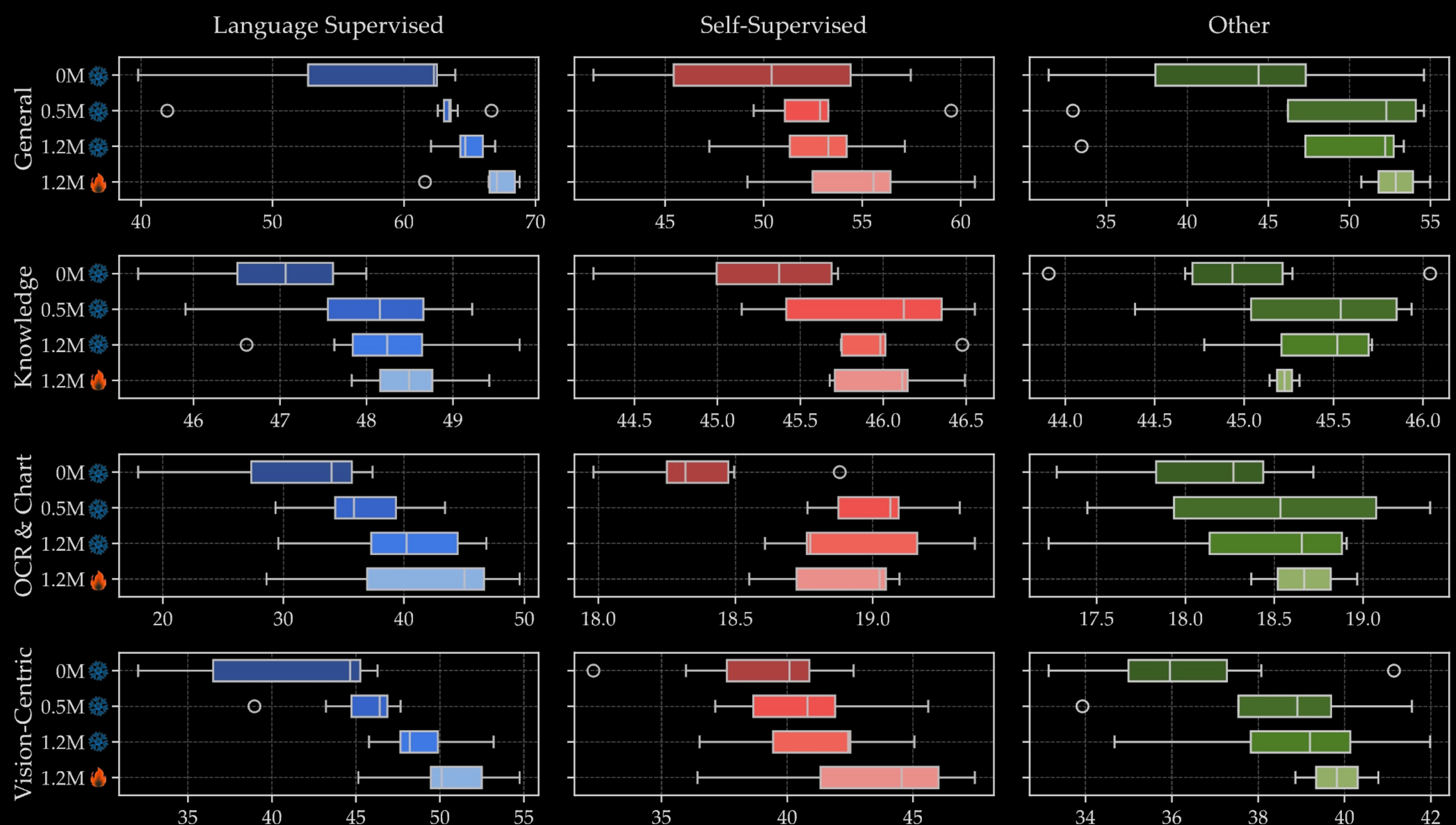
Table 1 | Breakdown of the 2D and 3D tasks evaluated in the Cambrian Vision-Centric Benchmark (CV-Bench). The examples are sourced from ADE20K [145], COCO [72], and Omni3D [16].



3.5x more vision-centric examples!
OCR & Chart
Vision-Centric
General
Knowledge
DocVQA
TextVQA
ChartQA
OCRBench
SQA-I
AI2D
MathVista
MMMU
Real World QA
MMB
SEED-I
MME Percept.
MMVP
GQA
CV-Bench
0.2
0.0
−0.2
−0.8
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4

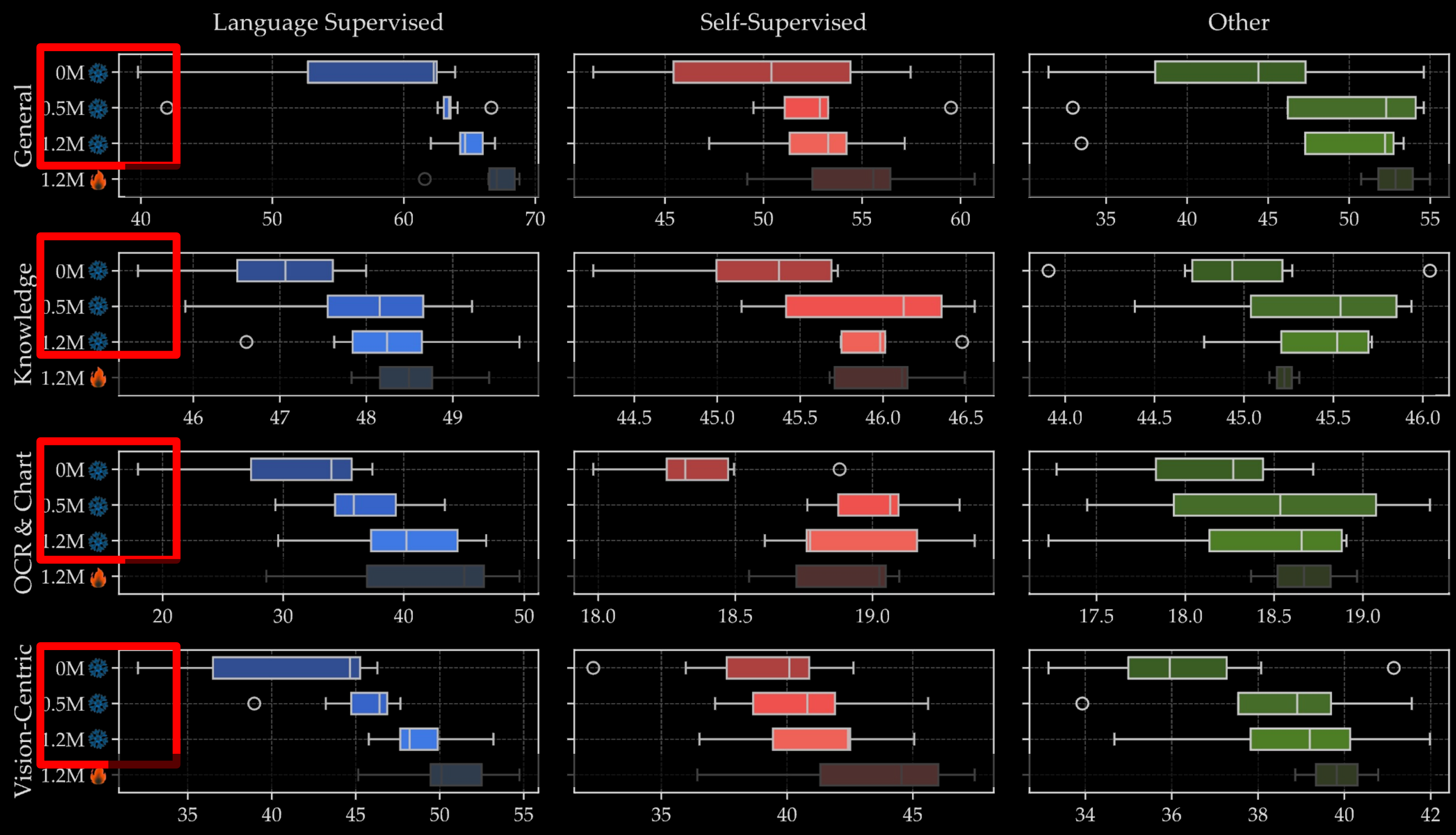

Overview
Evaluation Protocol
Instruction Tuning Recipe
Visual Representations
Connector Design
Instruction Tuning Data

Instruction Tuning Recipe



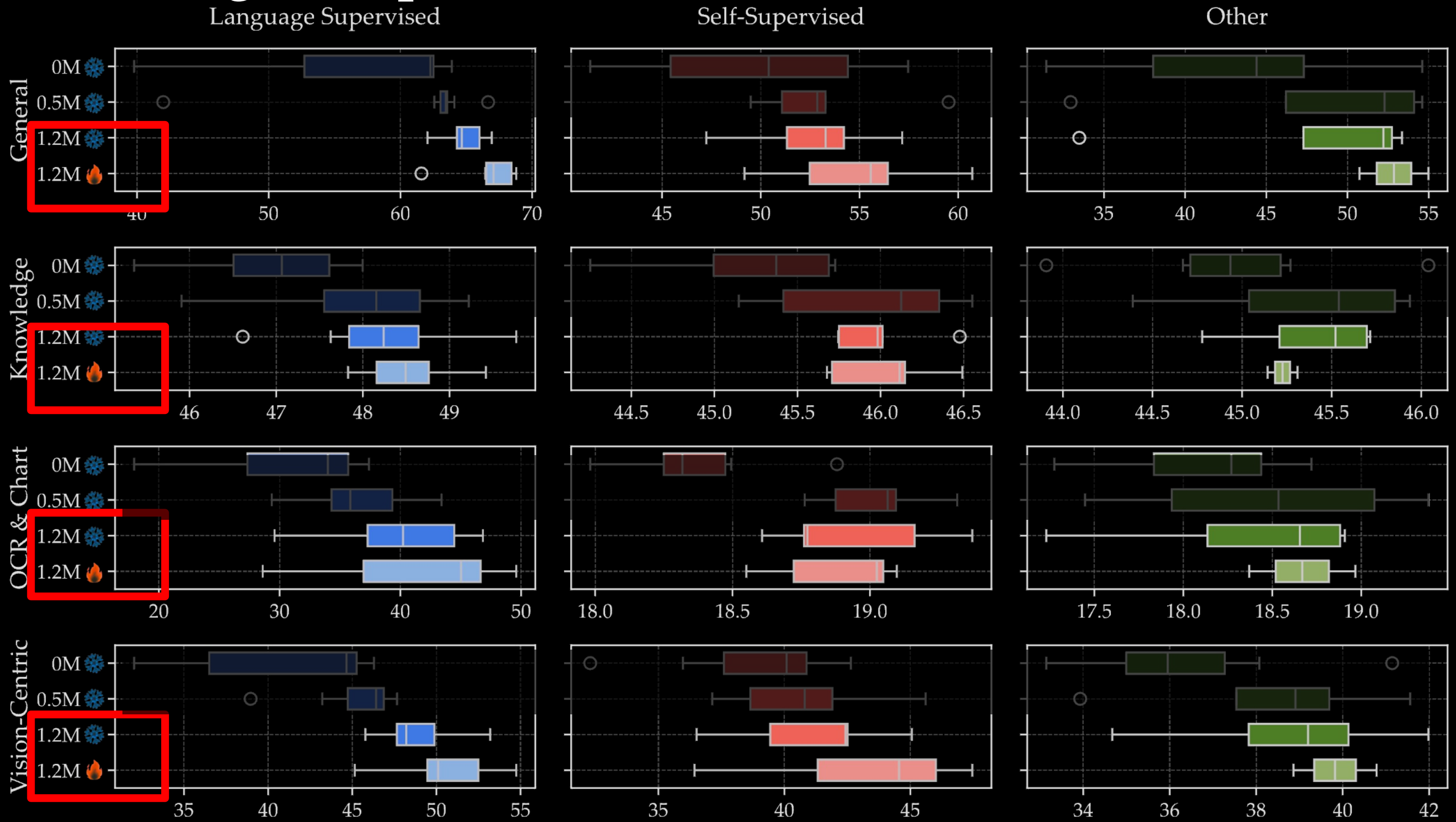
Instruction Tuning Recipe

More Alignment Data helps!

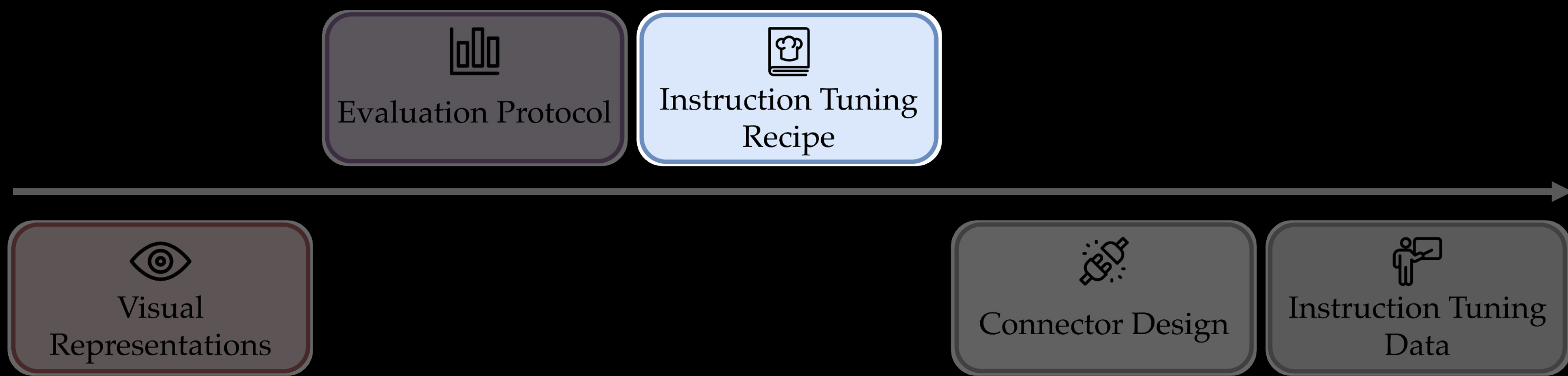


Instruction Tuning Recipe

Unfreezing Vision Encoder Helps 🔥

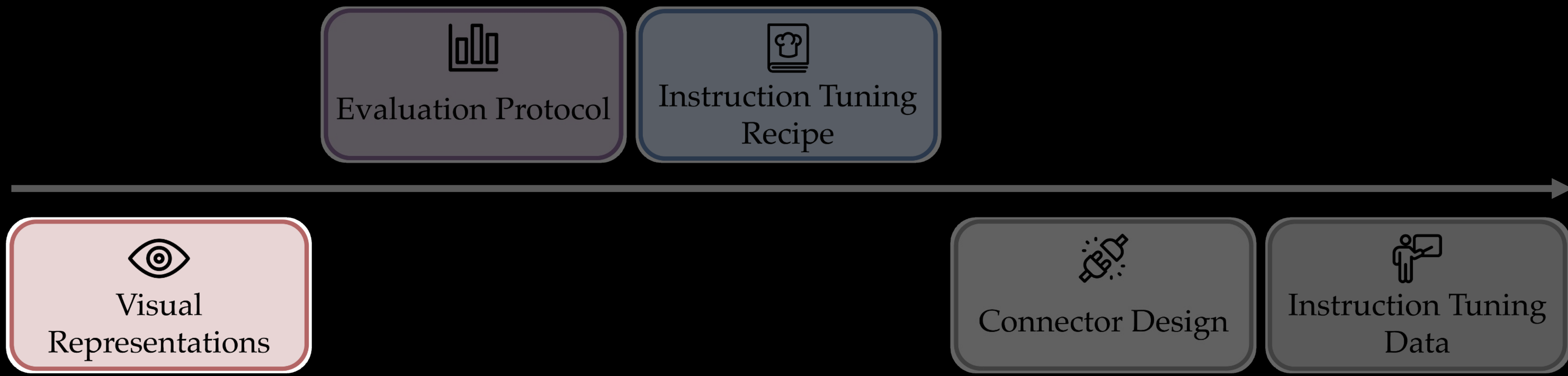


Overview

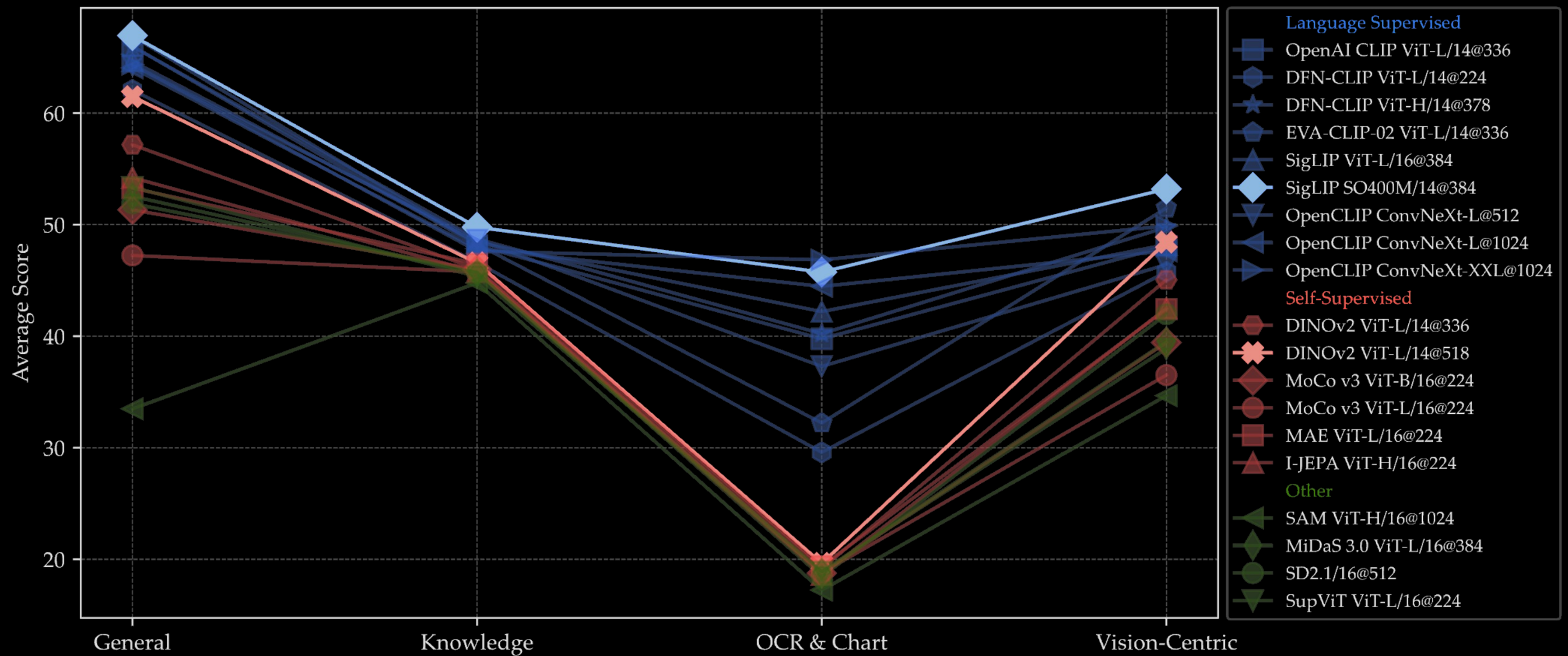


Evaluation Protocol
Instruction Tuning Recipe
Visual Representations
Connector Design
Instruction Tuning Data

Overview

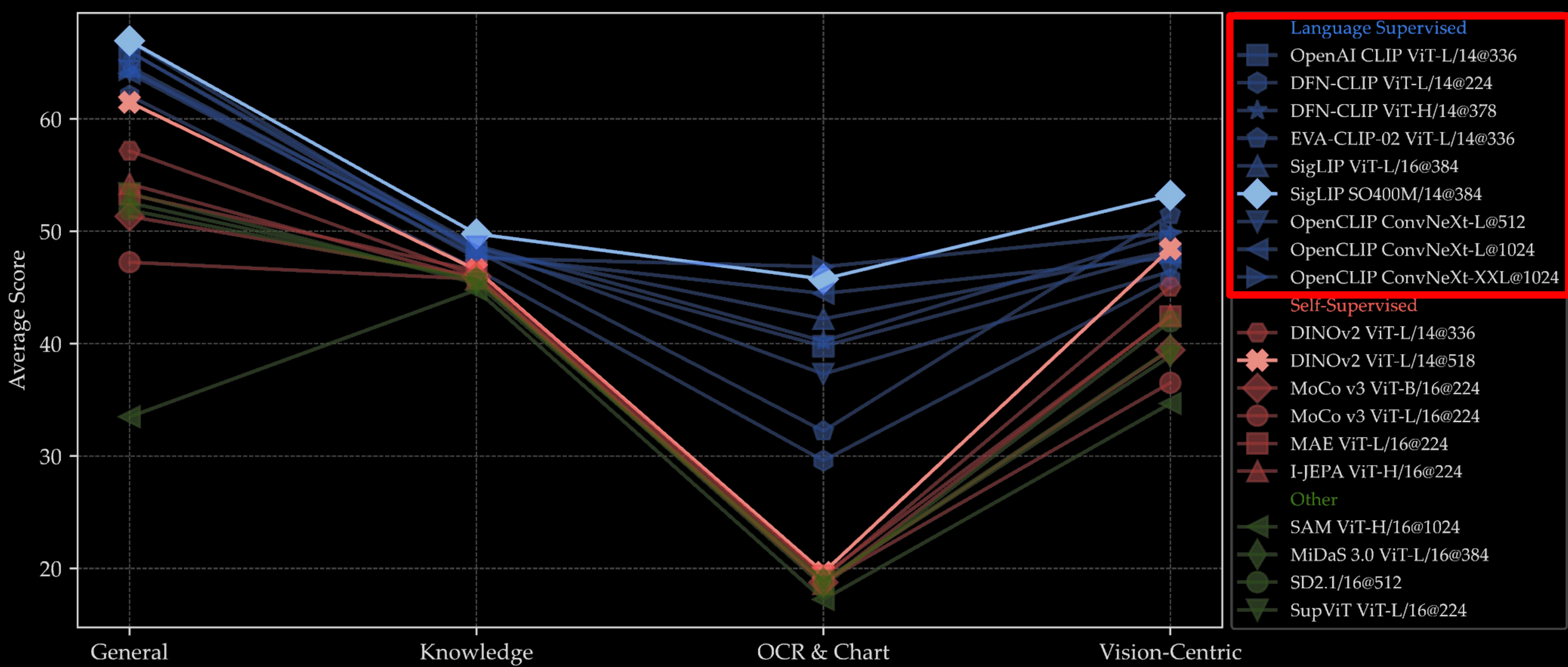

Evaluation Protocol
Instruction Tuning Recipe
Visual Representations
Connector Design
Instruction Tuning Data

Visual Representation



Visual Representation

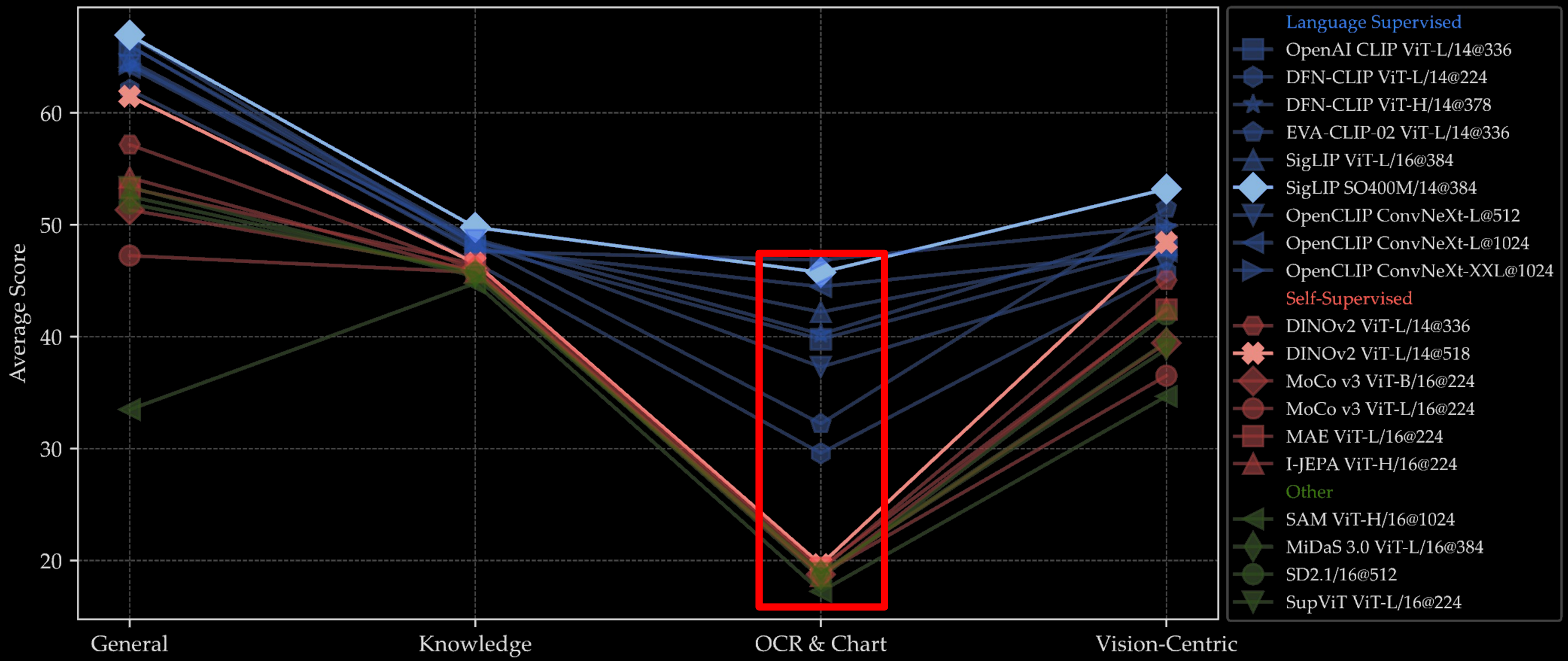
#1 Language Supervised Models are better



Visual Representation

#1 Language Supervised Models are better

#2 Gap is largest in OCR & Chart



Visual Representation

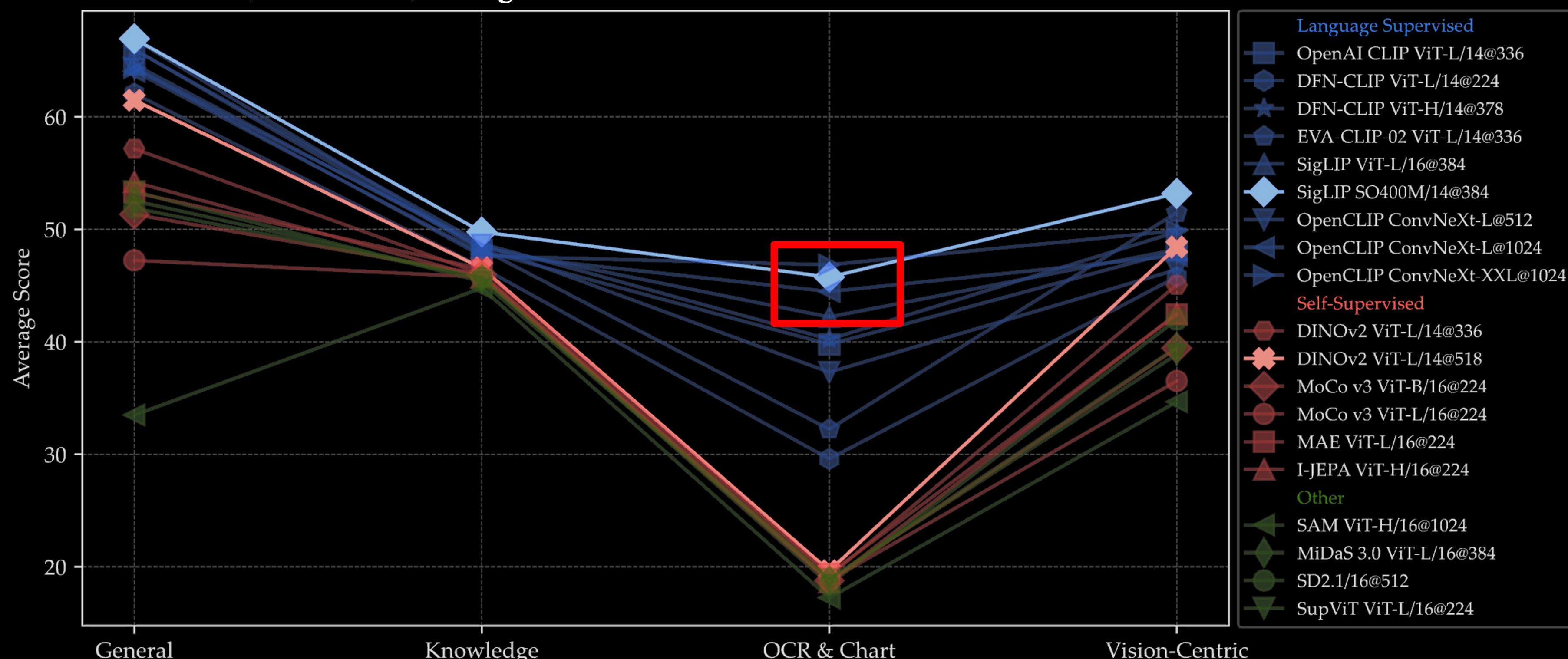
A ConvNet for the 2020s

Zhuang Liu[1,2*] Hanzi Mao[1] Chao-Yuan Wu[1] Christoph Feichtenhofer[1] Trevor Darrell[2] Saining Xie[1†]
[1]Facebook AI Research (FAIR) [2]UC Berkeley
Code: https://github.com/facebookresearch/ConvNeXt

#1 Language Supervised Models are better

#2 Gap is largest in OCR & Chart

#3 ConvNets (ConvNeXt) are good at OCR



Visual Representation

#1 Language Supervised Models are better

#2 Gap is largest in OCR & Chart

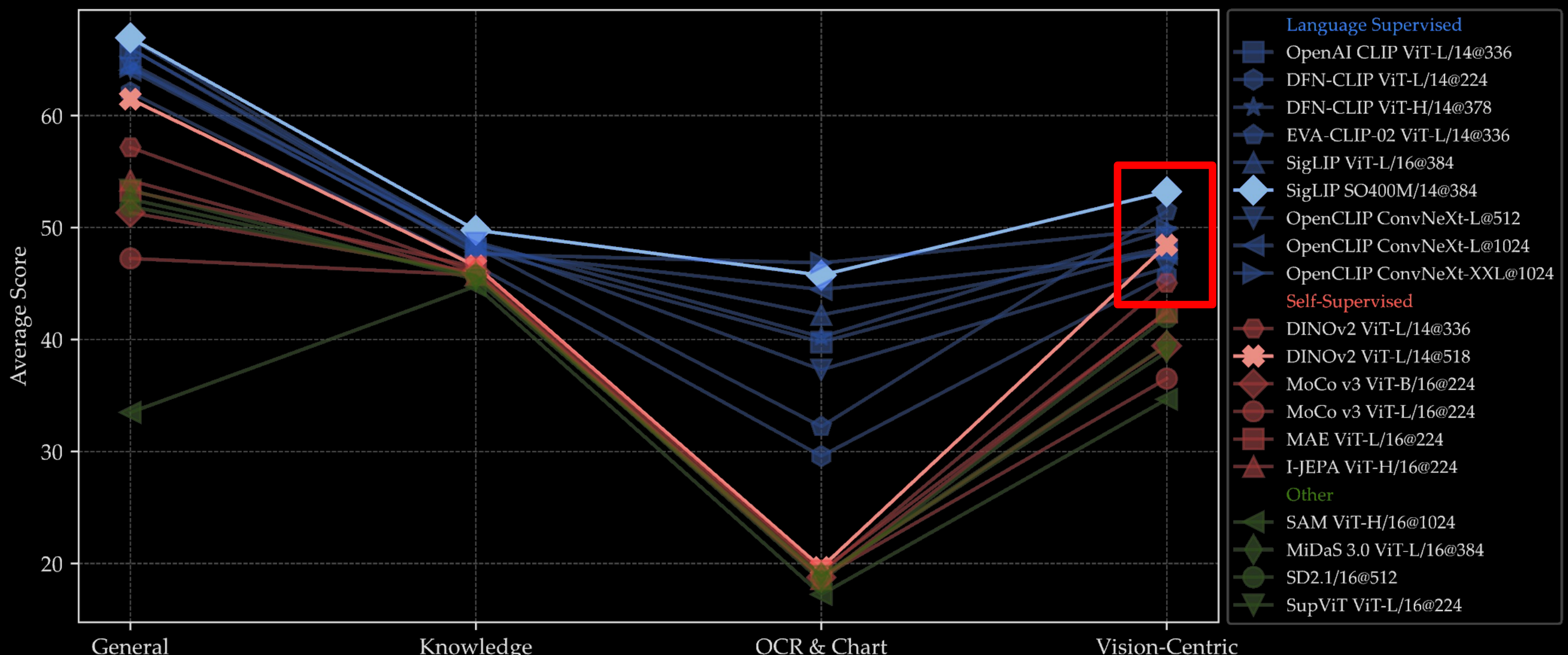
#3 ConvNets are good at OCR

Language Supervised

Model	Architecture	All	G	K	O	V
SigLIP	ViT-SO400M/14@384	1	1	1	2	1
OpenCLIP	ConvNeXt-XXL@1024	2	6	8	1	3
DFN-CLIP	ViT-H/14@378	3	4	2	5	4
OpenCLIP	ConvNeXt-L@1024	4	8	7	3	8
SigLIP	ViT-L/16@384	5	5	4	4	6
OpenAI CLIP	ViT-L/14@336	6	3	6	6	7
EVA-CLIP-02	ViT-L/14@336	7	2	5	8	2
OpenCLIP	ConvNeXt-L@512	8	7	3	7	9
DFN-CLIP	ViT-L/14@224	9	9	9	9	10
DINOv2*	ViT-L/14@518	10	10	10	10	5

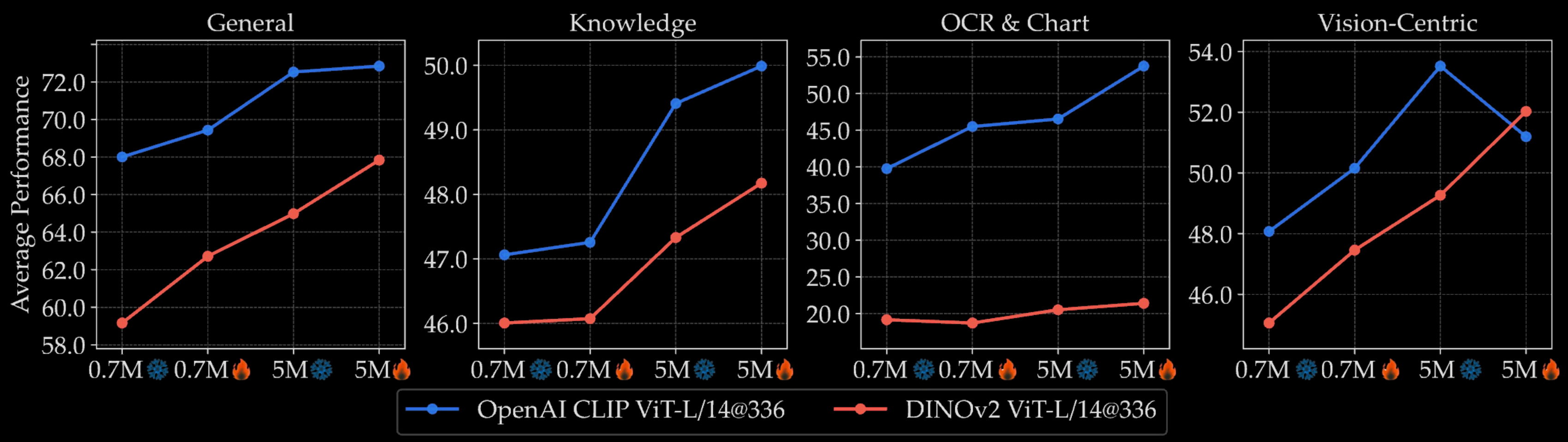
Visual Representation

#1 Language Supervised Models are better
#2 Gap is largest in OCR & Chart
#3 ConvNets are good at OCR
#4 Best SSL model is good at Vision-Centric tasks



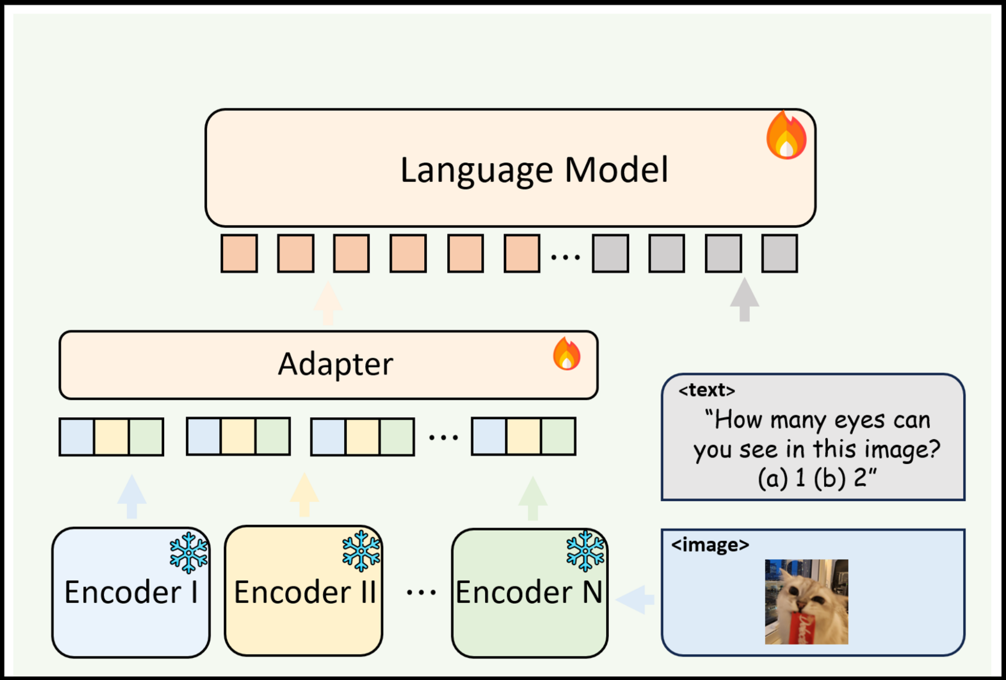
Visual Representation

Training with more data narrows the gap between Language-Supervised and Self-Supervised Models



Visual Representation

Combining different models improves performance



Visual Representation

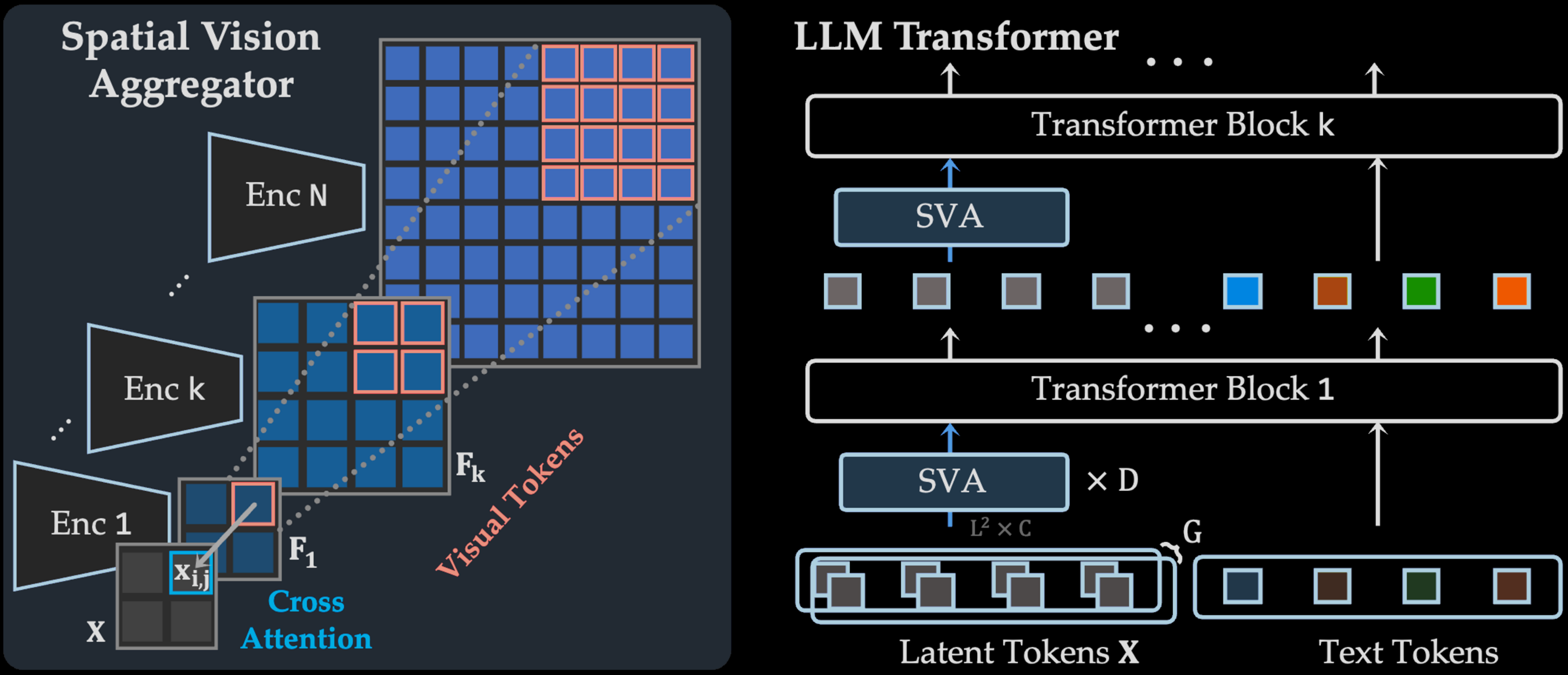
Combining different models improves performance (using a simple linear connector)

Vision Backbone		General				Knowledge				OCR & Chart				Vision-Centric			
Method	Average	MME^{P}	MMB	$SEED^{I}$	GQA	SQA^{I}	$MMMU^{V}$	$MathVista^{M}$	AI2D	ChartQA	OCRBench	TextVQA	DocVQA	MMVP	RealWorldQA	$CV\text{-}Bench^{2D}$	$CV\text{-}Bench^{3D}$
SigLIP+DINOv2	51.61	1,432.02	61.28	65.99	63.30	68.82	35.69	29.40	60.01	43.00	35.70	60.40	37.54	30.00	53.99	55.52	53.58
SigLIP+DINOv2+ConvNext	54.52	1,503.51	63.83	67.97	63.95	70.40	35.99	29.30	60.69	48.20	36.90	64.97	45.53	34.67	58.69	55.74	60.33
SigLIP+DINOv2+ConvNext+CLIP	54.74	1,479.46	63.32	67.63	64.04	71.39	35.49	29.10	59.88	50.24	39.60	64.55	46.12	32.67	58.95	58.54	60.42
SigLIP+ConvNext	54.53	1,494.97	64.60	67.98	63.58	71.05	34.90	29.80	60.85	50.64	38.00	64.53	46.52	32.00	57.91	58.83	56.58
CLIP+ConvNext	54.45	1,511.08	63.83	67.41	63.63	70.80	35.09	30.40	59.91	51.32	35.00	64.45	47.88	33.33	57.25	56.32	59.08
SigLIP+DINOv2+ConvNext	53.78	1,450.64	63.57	67.79	63.63	71.34	34.80	30.20	61.04	49.32	37.70	64.05	45.83	30.00	56.21	58.08	54.33
SigLIP+CLIP+ConvNext	54.53	1,507.28	63.23	68.64	63.63	71.10	35.89	30.90	59.97	52.36	38.50	65.40	47.92	28.67	57.25	57.66	55.92

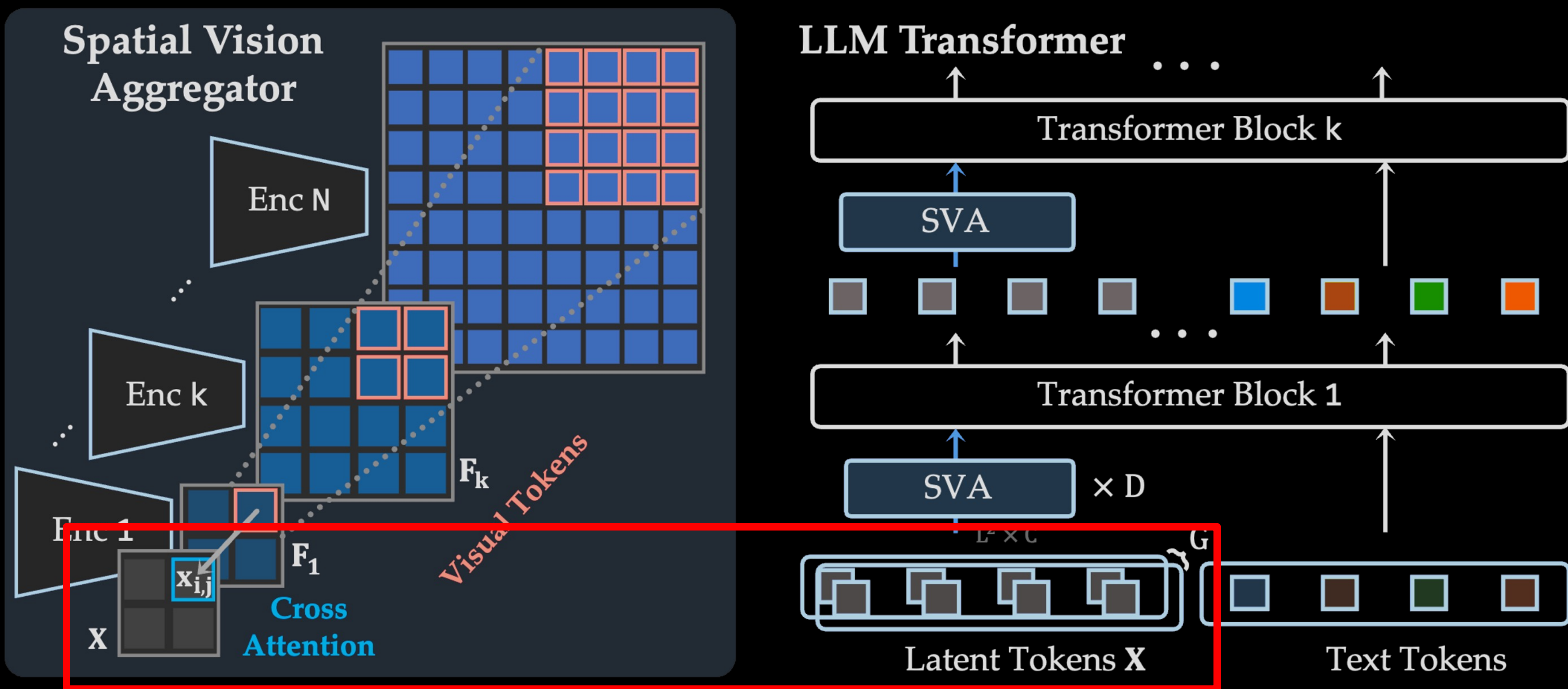
Overview

- Visual Representations
- Evaluation Protocol
- Instruction Tuning Recipe
- Connector Design
- Instruction Tuning Data

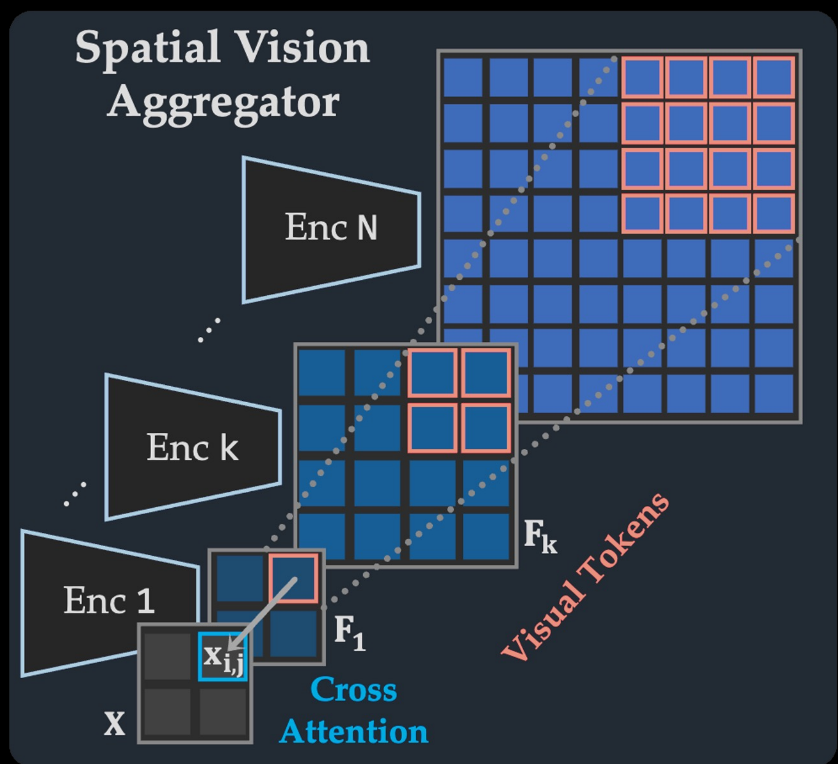
Connector Design - SVA



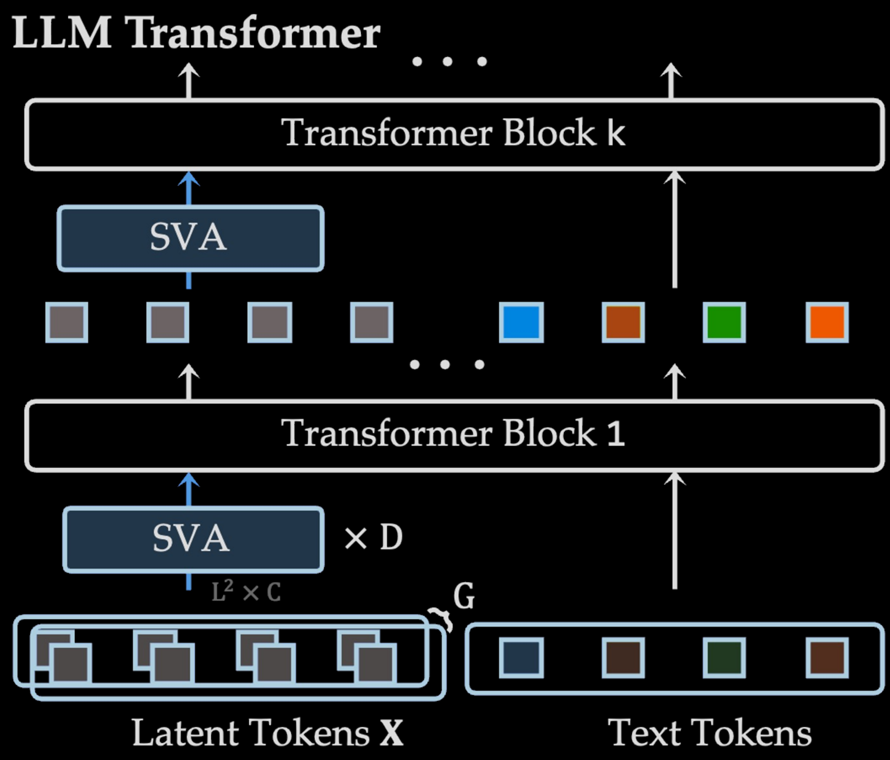
Connector Design - SVA



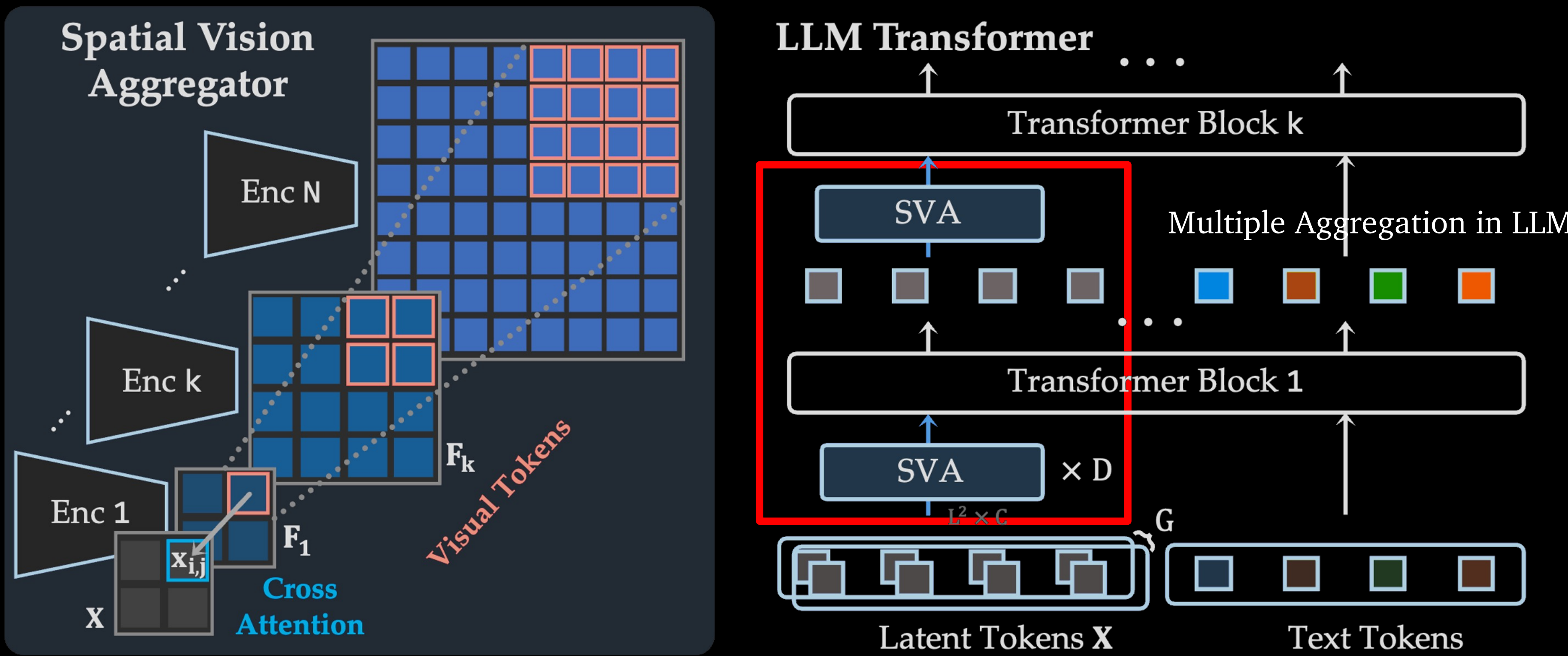
Connector Design - SVA



Spatial Inductive Bias



Connector Design - SVA



Connector Design - SVA

Spatial Inductive Bias is important especially for OCR&Chart and Vision-Centric Domains

Connector	General	Knowledge	OCR & Chart	Vision-Centric
Concat. [117]	67.2	48.9	50.1	52.6
Resampler [51]	63.1	46.5	27.1	42.6
SVA-no-multi-agg	68.0	49.5	55.2	52.6
SVA	**68.5**	**49.7**	**55.5**	**53.2**

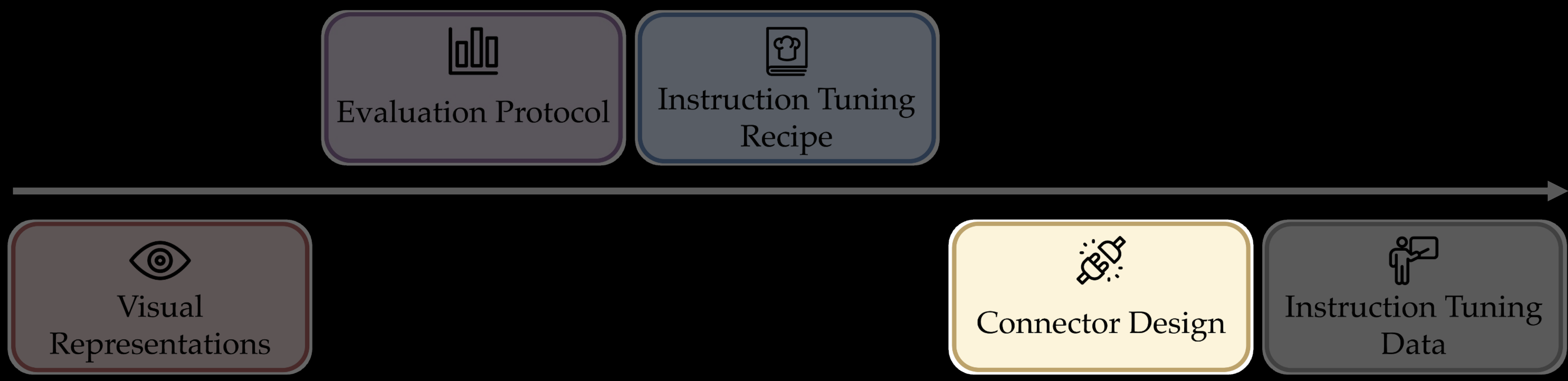
Connector Design - SVA

Spatial Inductive Bias is important especially for OCR&Chart and Vision-Centric Domains

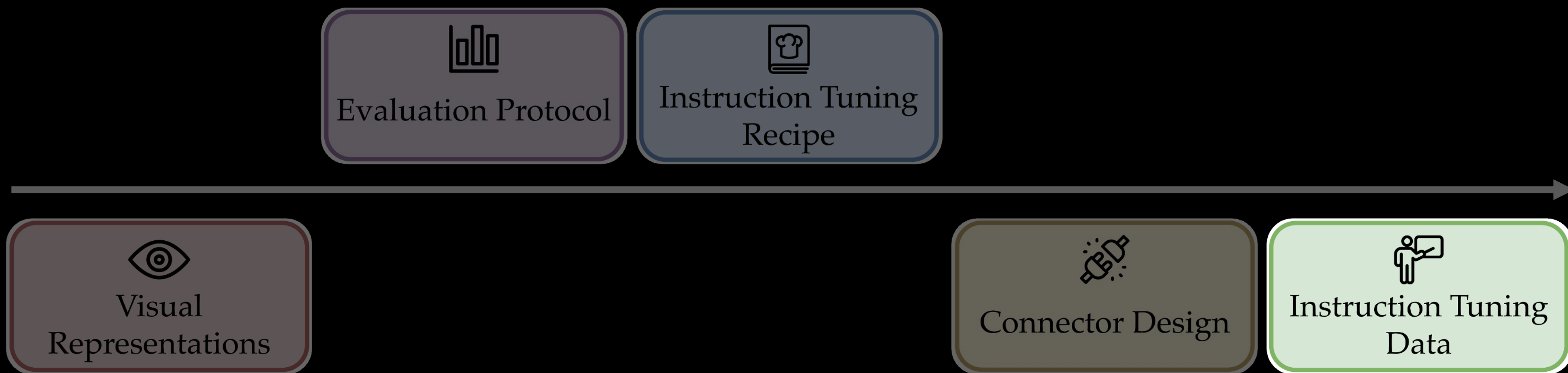
Multiple Aggregation in LLM further improves performance

Connector	General	Knowledge	OCR & Chart	Vision-Centric
Concat. [117]	67.2	48.9	50.1	52.6
Resampler [51]	63.1	46.5	27.1	42.6
SVA-no-multi-agg	68.0	49.5	55.2	52.6
SVA	**68.5**	**49.7**	**55.5**	**53.2**

Overview


Evaluation Protocol
Instruction Tuning Recipe
Visual Representations
Connector Design
Instruction Tuning Data

Overview


Evaluation Protocol
Instruction Tuning Recipe
Visual Representations
Connector Design
Instruction Tuning Data

Instruction Tuning Data

Collect all-potential Instruction Tuning Data

OCR (27.6%)

- Filtered DVQA (1550.0 K)
- DVQA [54] (775.0 K)
- SynthDog [60] (500.0 K)
- ArxivQA [69] (100.0 K)
- OCRVQA [93] (80.0 K)
- ScreenQA [49] (79.0 K)
- WIkiSQL [144] (74.0 K)
- Low-Level Vision [22] (50.0 K)
- DocVQA [90] (39.0 K)
- WTQ [99] (38.0 K)
- ChartQA [89] (28.0 K)
- IconQA [82] (27.0 K)
- Chart2Text [55] (26.0 K)
- TabMWP [81] (23.0 K)
- TextCaps [111] (22.0 K)
- LLAVAR [140] (20.0 K)
- ST-VQA [15] (17.0 K)
- AI2D [58] (15.0 K)
- RenderedText [125] (10.0 K)
- VisText [115] (9.0 K)
- FinQA [26] (6.0 K)
- InfoVQA [14] (2.0 K)
- TAT-QA [148] (2.0 K)
- HiTab [27] (2.0 K)

General (33.3%)

- ALLaVA [20] (700.0 K)
- Q-Instruct [126] (400.0 K)
- LNQA [101] (302.0 K)
- LVIS-Instruct4V [122] (220.0 K)
- LLaVA150K [75] (150.0 K)
- VisualGenome [62] (86.0 K)
- VQAv2 [43] (83.0 K)
- GPT4V Rewritten (77.0 K)
- GQA [50] (72.0 K)
- A-OKVQA [108] (50.0 K)
- AlfWorld [137] (45.0 K)
- ShareGPT [22] (40.0 K)
- RefCOCO [131] (30.0 K)
- VizWiz [44] (20.0 K)
- Visual7W [149] (14.0 K)
- LAION GPT-4V [63] (11.0 K)
- IDK [17] (11.0 K)
- OKVQA [88] (9.0 K)
- HatefulMemes [59] (8.0 K)
- OODVQA [120] (8.0 K)
- SketchyVQA [120] (8.0 K)
- Visualmrc [114] (3.0 K)

Language (23.8%)

- OpenOrca [71] (994.0 K)
- MathInstruct [133] (262.0 K)
- OrcaMath [92] (200.0 K)
- WizardCoder [86] (143.0 K)
- OpenCodeInterpreter [143] (66.0 K)
- Dolly [30] (11.0 K)

Counting (8.5%)

- Filtered CLEVR (350.0 K)
- CLEVR [52] (350.0 K)
- TallyQA [1] (250.0 K)

Code (0.8%)

- Filtered WebSight (790.0 K)
- WebSight [64] (10.0 K)
- DaTikz [12] (47.0 K)
- Design2Code [110] (0.5 K)

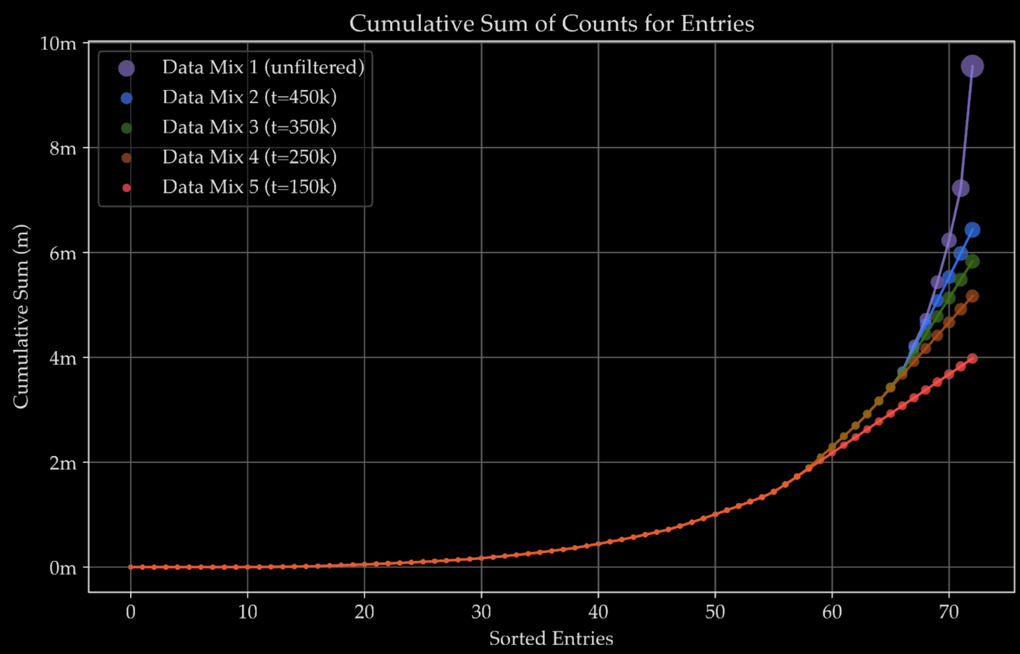
Math (3.2%)

- Geo170K [37] (170.0 K)
- RAVEN [139] (42.0 K)
- GeomVerse [57] (9.0 K)
- MathVision [123] (3.0 K)
- Inter-GPS [83] (1.0 K)
- TQA [4] (1.0 K)

Science (2.9%)

- Data Engine (161.0 K)
- PathVQA [46] (32.0 K)
- ScienceQA [84] (12.0 K)

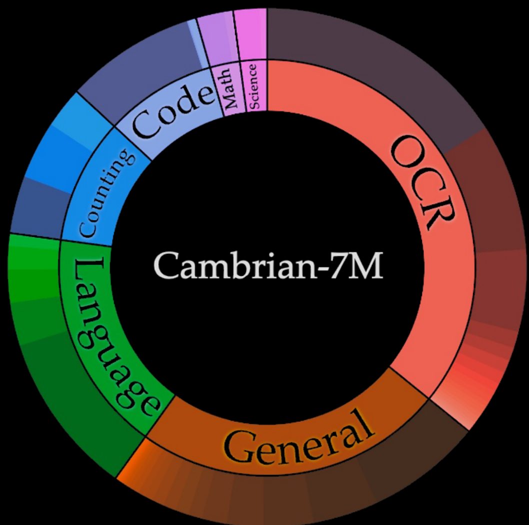
Data Balancing (Filtering)



	Average	General	Knowledge	OCR & Chart	Vision-Centric
150k	53.7	68.0	51.3	45.2	50.5
250k	**54.3**	**68.1**	51.5	45.3	52.2
350k	54.3	67.4	51.4	**46.0**	**52.3**
450k	54.2	68.0	**52.2**	45.5	50.7

Data Mixing Ratio

Effect of Data Curation



Data quality matters

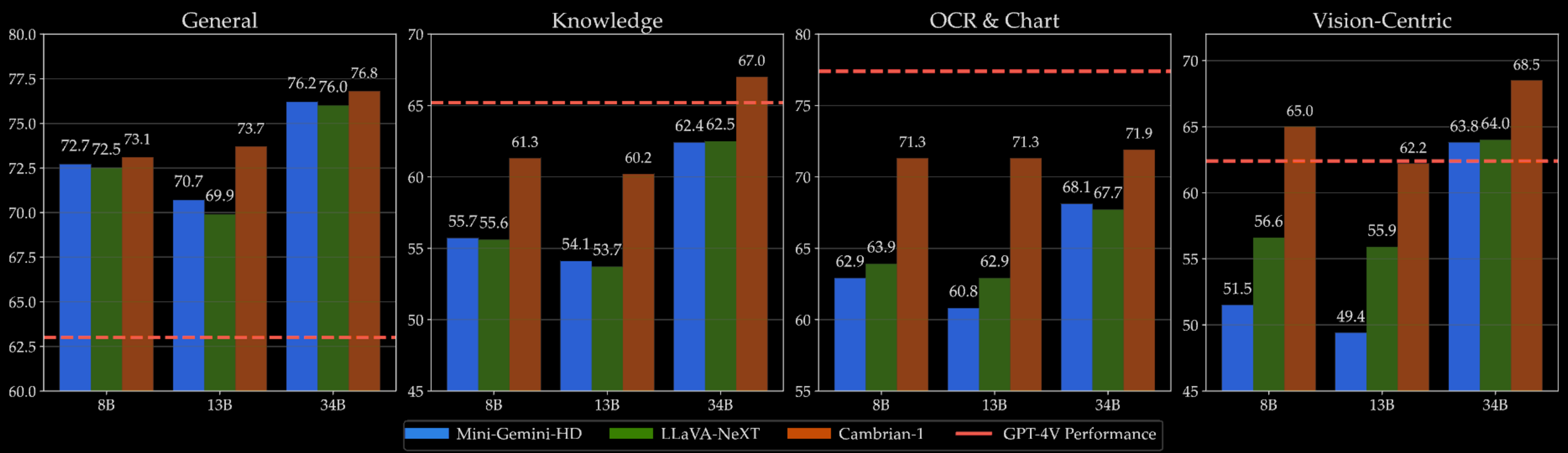
	Average	General	Knowledge	OCR & Chart	Vision-Centric
LLaVA-665K	40.7	64.7	45.2	20.8	32.0
Cambrian-10M	54.8	68.7	51.6	**47.3**	51.4
Cambrian-7M	**55.9**	**69.6**	**52.6**	**47.3**	**54.1**

Cambrian-1 Models

"SOTA" Performance

Model		General					Knowledge					OCR & Chart					Vision-Centric				
Method	# Vis Tok.	Avg	MME^{P}	MMB	SEED^{I}	GQA	Avg	SQA^{I}	MMMU^{V}	$\text{MathVista}^{\text{M}}$	AI2D	Avg	ChartQA	OCRBench	TextVQA	DocVQA	Avg	MMVP	RealworldQA	$\text{CV-Bench}^{\text{2D}}$	$\text{CV-Bench}^{\text{3D}}$
GPT-4V	UNK.	63.0	1409.4	75.8	69.1	36.8	65.2	75.7	56.8	49.9	78.2	77.4	78.5	64.5	78.0	88.4	62.4	50.0	61.4	64.3	73.8
Gemini-1.0 Pro	UNK.	-	1496.6	73.6	70.7	-	-	79.5	47.9	45.2	-	-	-	65.9	-	-	-	-	-	-	-
Gemini-1.5 Pro	UNK.	-	-	-	-	-	-	-	58.5	52.1	80.3	-	81.3	-	73.5	86.5	-	-	67.5	-	-
Grok-1.5	UNK.	-	-	-	-	-	-	-	53.6	52.8	88.3	-	76.1	-	78.1	85.6	-	-	68.7	-	-
MM-1-8B	144	-	1529.3	72.3	69.9	-	-	72.6	37.0	35.9	-	-	-	-	-	-	-	-	-	-	-
MM-1-30B	144	-	1637.6	75.1	72.1	-	-	81.0	44.7	39.4	-	-	-	-	-	-	-	-	-	-	-
Base LLM: Llama-3-Ins-8B																					
Mini-Gemini-HD-8B	2880	72.7	**1606.0**	72.7	73.2	64.5	55.7	75.1	37.3	37.0	**73.5**	62.9	59.1	47.7	70.2	74.6	51.5	18.7	62.1	62.2	63.0
LLaVA-NeXT-8B	2880	72.5	1603.7	72.1	72.7	**65.2**	55.6	72.8	41.7	36.3	71.6	63.9	69.5	49.0	64.6	72.6	56.6	38.7	60.1	62.2	65.3
Cambrian-1-8B	576	**73.1**	1,547.1	**75.9**	**74.7**	64.6	**61.3**	**80.4**	**42.7**	**49.0**	73.0	**71.3**	**73.3**	**62.4**	**71.7**	**77.8**	**65.0**	**51.3**	**64.2**	**72.3**	**72.0**
Base LLM: Vicuna-1.5-13B																					
Mini-Gemini-HD-13B	2880	70.7	1597.0	68.6	70.6	63.7	54.1	71.9	37.3	37.0	70.1	60.8	56.6	46.6	70.2	69.8	49.4	19.3	57.5	53.6	67.3
LLaVA-NeXT-13B	2880	69.9	1575.0	70.0	65.6	**65.4**	53.7	73.5	36.2	35.1	70.0	62.9	62.2	51.4	67.1	70.9	55.9	36.0	59.1	62.7	65.7
Cambrian-1-13B	576	**73.7**	**1,610.4**	**75.7**	**74.4**	64.3	**60.2**	**79.3**	**40.0**	**48.0**	**73.6**	**71.3**	**73.8**	**61.9**	**72.8**	**76.8**	**62.2**	**41.3**	**63.0**	**72.5**	**71.8**
Base LLM: Hermes2-Yi-34B																					
Mini-Gemini-HD-34B	2880	76.2	1659.0	80.6	75.3	65.8	62.4	77.7	48.0	43.4	**80.5**	68.1	67.6	51.8	74.1	**78.9**	63.8	37.3	67.2	71.5	79.2
LLaVA-NeXT-34B	2880	76.0	1633.2	79.3	**75.9**	**67.1**	62.5	81.8	46.7	46.5	74.9	67.7	68.7	54.5	69.5	78.1	64.0	47.3	61.0	73.0	74.8
Cambrian-1-34B	576	**76.8**	**1689.3**	**81.4**	75.3	65.8	**67.0**	**85.6**	**49.7**	**53.2**	79.7	**71.9**	**75.6**	**60.0**	**76.7**	75.5	**68.5**	**52.7**	**67.8**	**74.0**	**79.7**

"SOTA" Performance



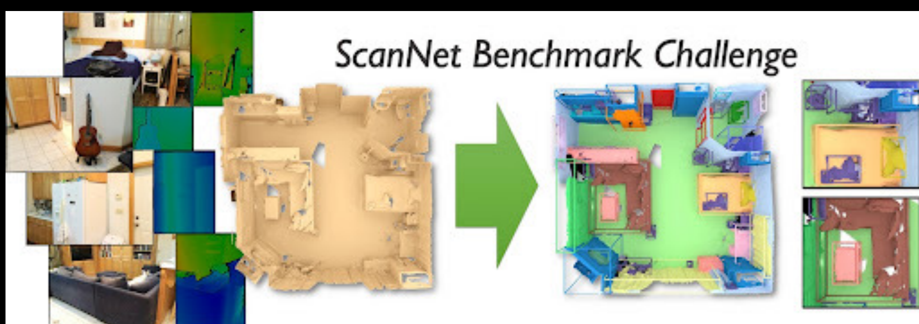
Key Question #4:

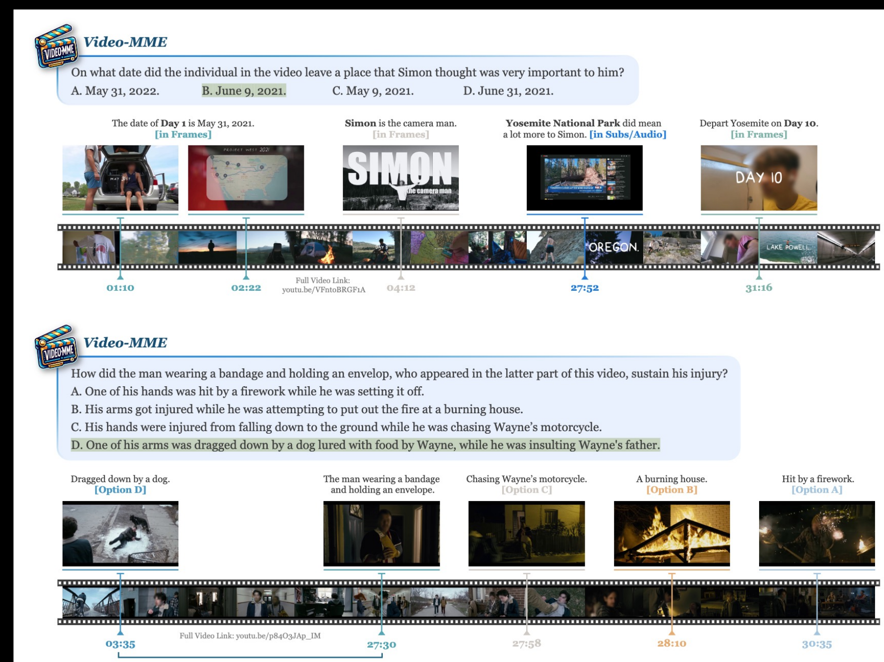
Can multimodal LLMs *think in space*?

In computer vision…

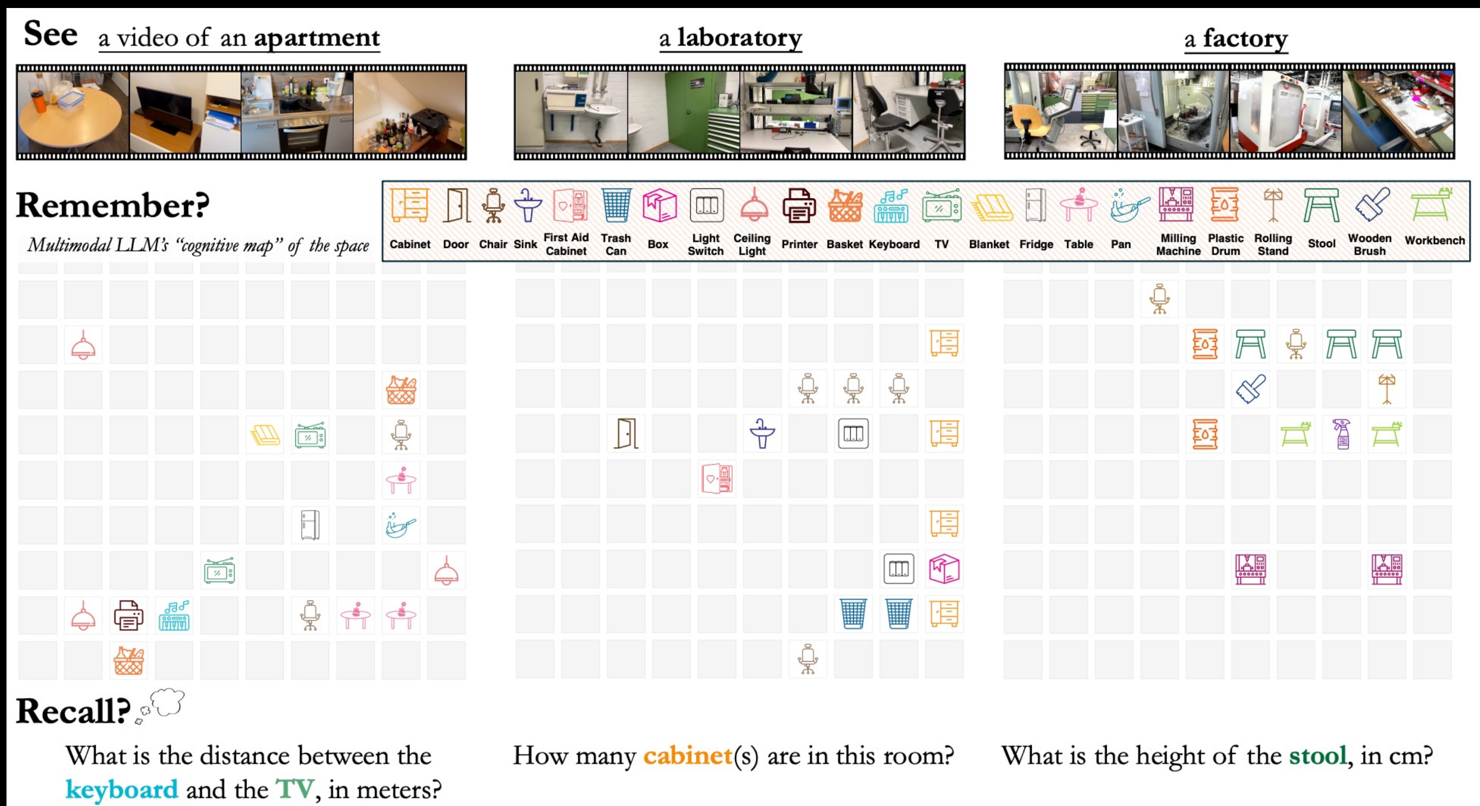
We study space, but not thinking…

We study thinking, but not in space





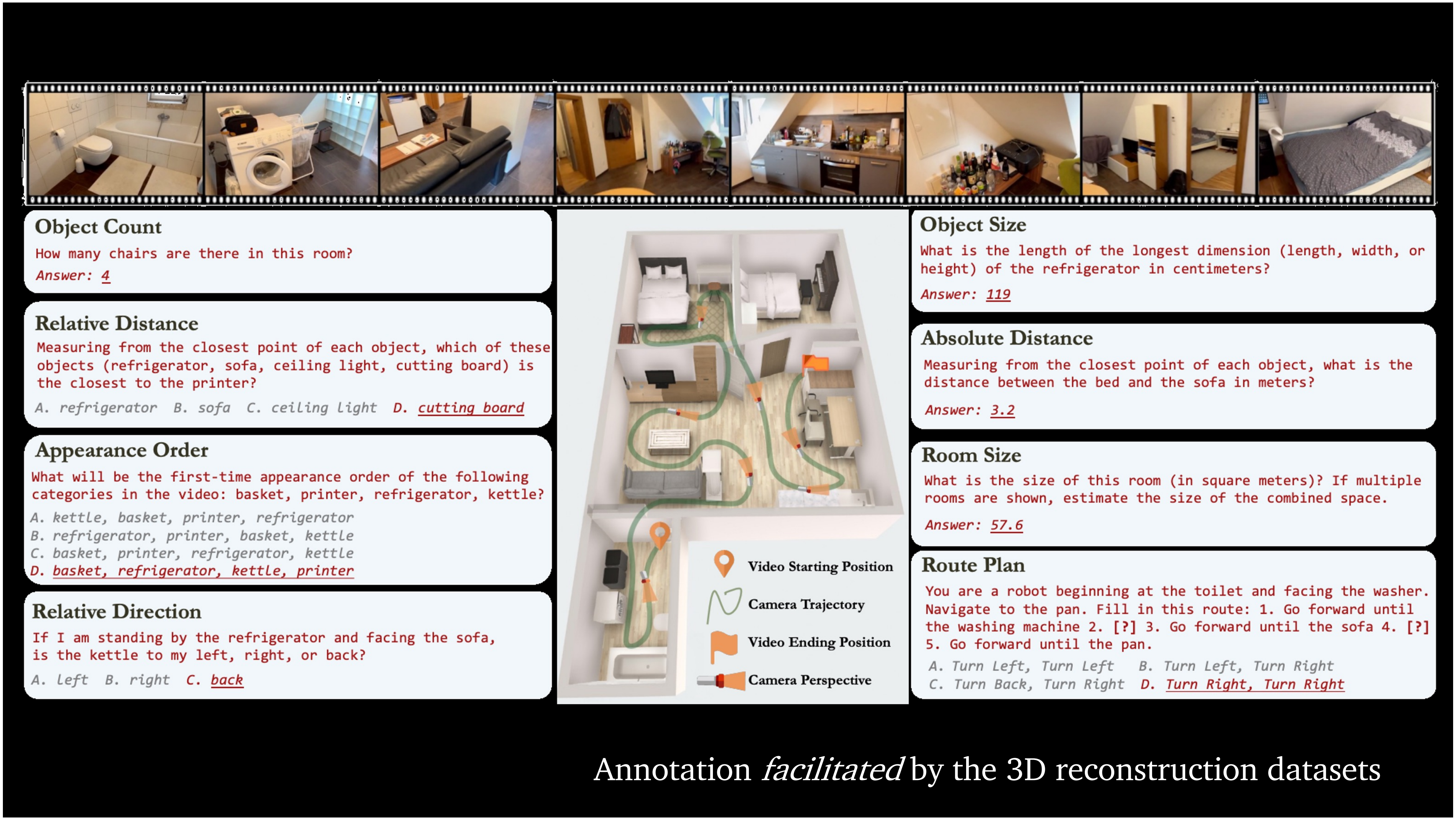
Thinking in Space: How Multimodal LLMs See, Remember and Recall Spaces



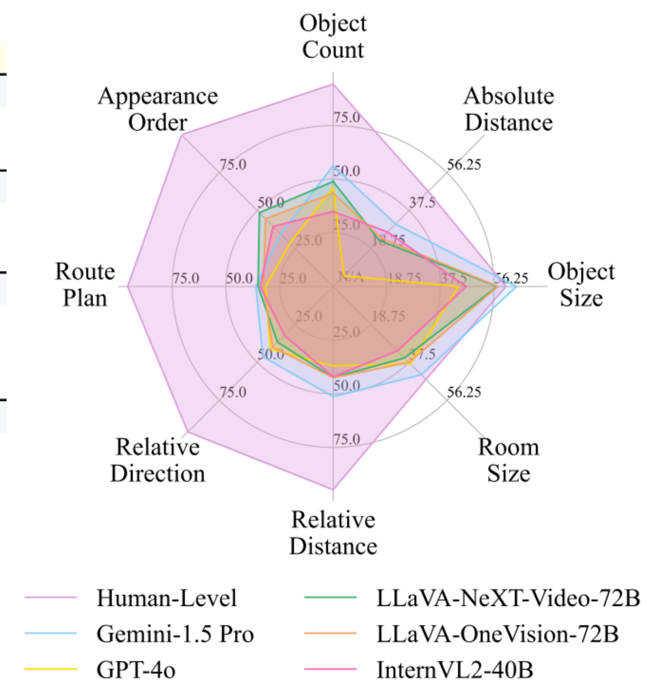
With Jihan Yang, Shusheng Yang, Anjali Gupta, Rilyn Han, and Fei-Fei Li

Apartment

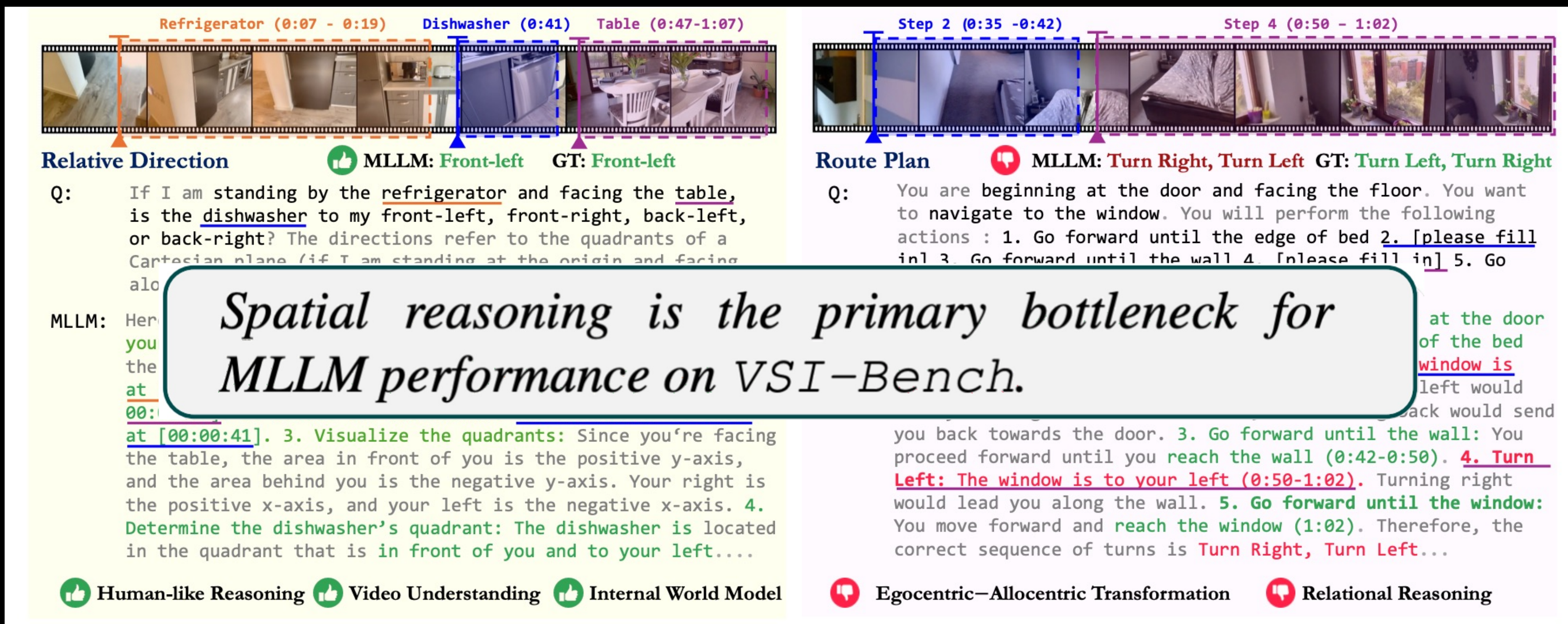
Lab


Object Count
How many chairs are there in this room?
Answer: 4
Relative Distance
Measuring from the closest point of each object, which of these objects (refrigerator, sofa, ceiling light, cutting board) is the closest to the printer?
A. refrigerator B. sofa C. ceiling light D. cutting board
Appearance Order
What will be the first-time appearance order of the following categories in the video: basket, printer, refrigerator, kettle?
A. kettle, basket, printer, refrigerator
B. refrigerator, printer, basket, kettle
C. basket, printer, refrigerator, kettle
D. basket, refrigerator, kettle, printer
Relative Direction
If I am standing by the refrigerator and facing the sofa, is the kettle to my left, right, or back?
A. left B. right C. back
Video Starting Position
Camera Trajectory
Video Ending Position
Camera Perspective
Object Size
What is the length of the longest dimension (length, width, or height) of the refrigerator in centimeters?
Answer: 119
Absolute Distance
Measuring from the closest point of each object, what is the distance between the bed and the sofa in meters?
Answer: 3.2
Room Size
What is the size of this room (in square meters)? If multiple rooms are shown, estimate the size of the combined space.
Answer: 57.6
Route Plan
You are a robot beginning at the toilet and facing the washer. Navigate to the pan. Fill in this route: 1. Go forward until the washing machine 2. [?] 3. Go forward until the sofa 4. [?] 5. Go forward until the pan.
A. Turn Left, Turn Left B. Turn Left, Turn Right
C. Turn Back, Turn Right D. Turn Right, Turn Right
Annotation *facilitated* by the 3D reconstruction datasets

Methods	Rank	Avg.	Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
			Numerical Answer				Multiple-Choice Answer			
Baseline										
Chance Level (Random)	-	-	-	-	-	-	25.0	36.1	28.3	25.0
Chance Level (Frequency)	-	34.0	62.1	32.0	29.9	33.1	25.1	47.9	28.4	25.2
`VSI-Bench` *(tiny) Perf.*										
†Human Level	-	79.2	94.3	47.0	60.4	45.9	94.7	95.8	95.8	100.0
†Gemini-1.5 Pro	-	46.0	47.0	43.2	36.2	42.4	58.2	60.0	43.8	50.0
Proprietary Models (API)										
GPT-4o	3	34.0	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5
Gemini-1.5 Flash	2	42.1	49.8	30.8	53.5	54.4	37.7	41.0	31.5	37.8
Gemini-1.5 Pro	1	45.4	56.2	30.9	64.1	43.6	51.3	46.3	36.0	34.6
Open-source Models										
InternVL2-2B	11	27.4	21.8	24.9	22.0	35.0	33.8	44.2	30.5	7.1
InternVL2-8B	5	34.6	23.1	28.7	48.2	39.8	36.7	30.7	29.9	39.6
InternVL2-40B	3	36.0	34.9	26.9	46.5	31.8	42.1	32.2	34.0	39.6
LongVILA-8B	12	21.6	29.1	9.1	16.7	0.0	29.6	30.7	32.5	25.5
VILA-1.5-8B	9	28.9	17.4	21.8	50.3	18.8	32.1	34.8	31.0	24.8
VILA-1.5-40B	7	31.2	22.4	24.8	48.7	22.7	40.5	25.7	31.5	32.9
LongVA-7B	8	29.2	38.0	16.6	38.9	22.2	33.1	43.3	25.4	15.7
LLaVA-NeXT-Video-7B	4	35.6	48.5	14.0	47.8	24.2	43.5	42.4	34.0	30.6
LLaVA-NeXT-Video-72B	1	40.9	48.9	22.8	57.4	35.3	42.4	36.7	35.0	48.6
LLaVA-OneVision-0.5B	10	28.0	46.1	28.4	15.4	28.3	28.9	36.9	34.5	5.8
LLaVA-OneVision-7B	6	32.4	47.7	20.2	47.4	12.3	42.5	35.2	29.4	24.4
LLaVA-OneVision-72B	2	40.2	43.5	23.9	57.6	37.5	42.5	39.9	32.5	44.6

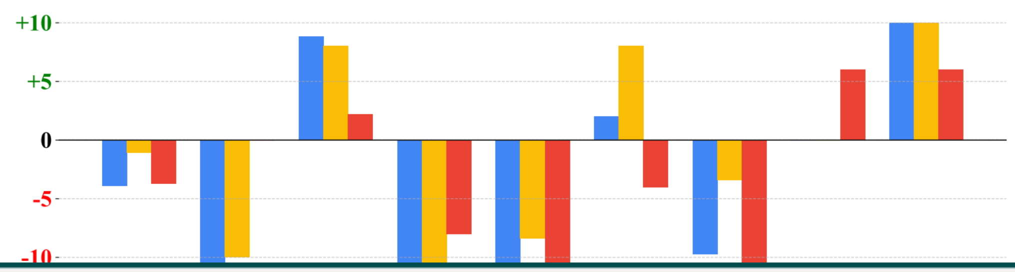


Error Analysis through LLM self-explanation



Spatial reasoning is the primary bottleneck for MLLM performance on `VSI-Bench`.

Language Prompting Techniques



> ***Linguistic prompting techniques, although effective in language reasoning tasks, are primarily harmful for spatial reasoning.***

Tree-of-Thought compared to the baseline. All three prevailing prompting techniques fail on average on our benchmark, and, in some cases, task performance becomes *much worse* after applying them. This implies that `VSI-Bench` cannot be solved by solely improving linguistic capabilities.

Note: works on general video understanding tasks

A *Local* "World Model"

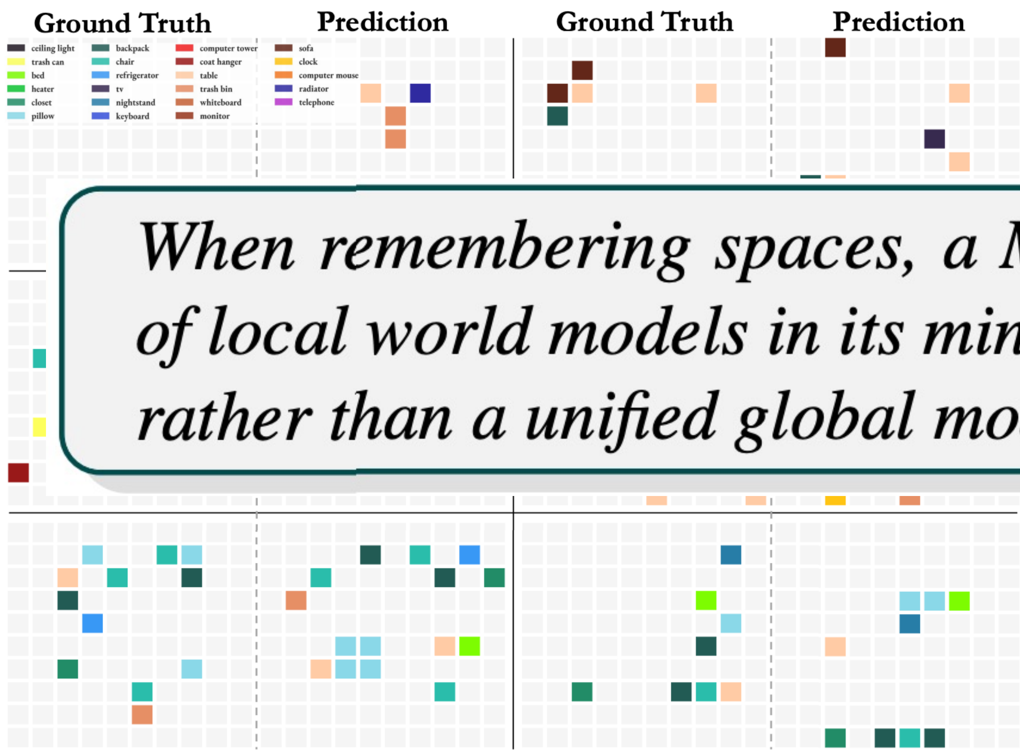


Figure 10. Visualization of cognitive maps from MLLM and GT.

When remembering spaces, a MLLM forms a series of local world models in its mind from a given video, rather than a unified global model.

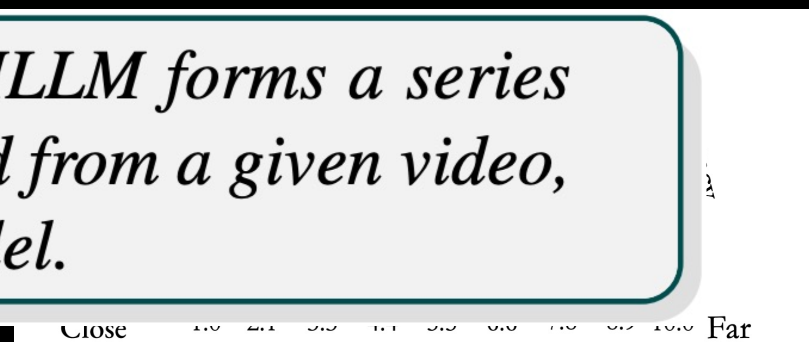
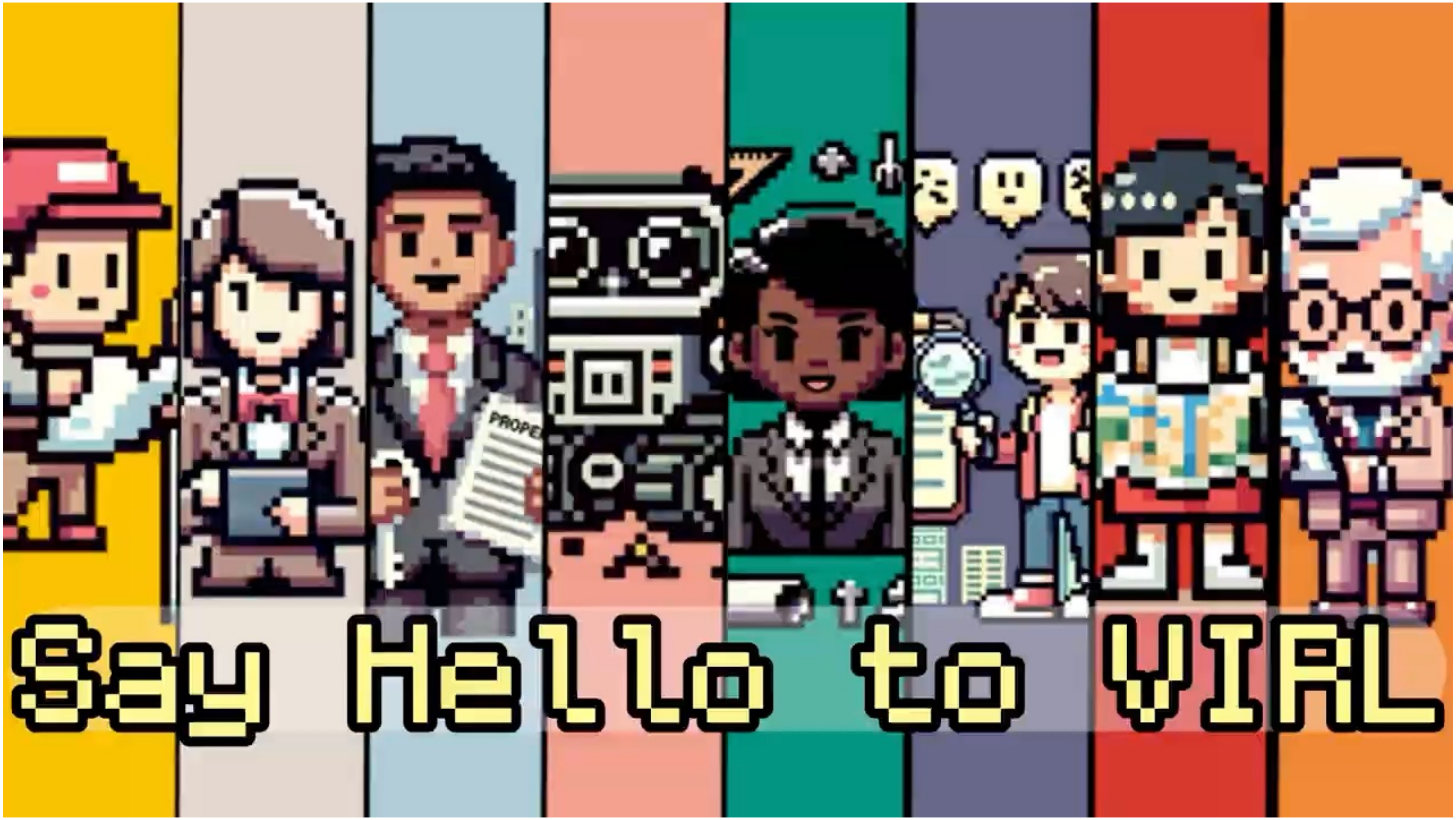


Figure 11. **Locality of the MLLM's predicted cognitive maps.** The MLLM's map-distance accuracy decreases dramatically with increasing object distance.

Key Question #5:

How to ground multimodal agents in real life?


Say Hello to VIRL

Thank You

Improved vision is not just about seeing farther, but about understanding more deeply.