

ELV26 Tutorial 5

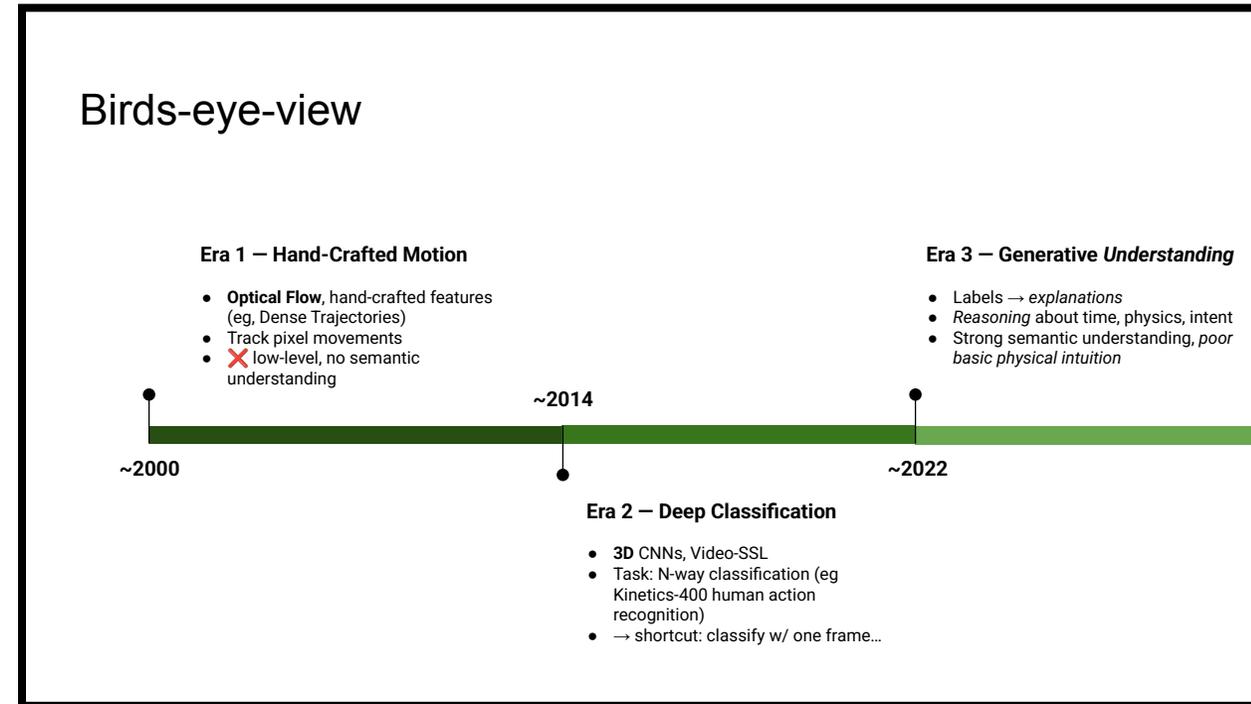
From *Seeing* to Modeling the World

Ellis Brown

2026-03-03

[Recap] Tutorial 3: Video Learning

- **Video-LLMs:** how they work, how to train them, and where they fail
- **Key failure modes:** temporal confusion, spatial grounding, long videos, continual sensing
- **Takeaway:** strong semantic understanding, but *poor basic physical intuition*



Roadmap: my research

I. How Do MLLMs See?

1. [Cambrian-1](#): Vision-centric MLLMs

II. Spatial Intelligence

1. [VSI-Bench](#)*: Benchmarking Spatial Intelligence
2. [SIMS-V](#): Spatial Training via Embodied Simulator
3. [Test-Set Stress Test](#): Debiasing Benchmarks

III. Towards World Models

1. [Cambrian-S](#): Spatial “SuperSensing”
2. [Unified Multimodal Models](#): Unified Multimodal Pretraining

* Not my paper, but from my research group!

I: How Do MLLMs See?

1. [Cambrian-1](#): Vision-centric MLLMs

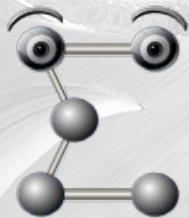
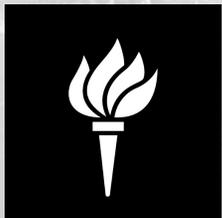
* Equal Contribution

Cambrion-1 🪼

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



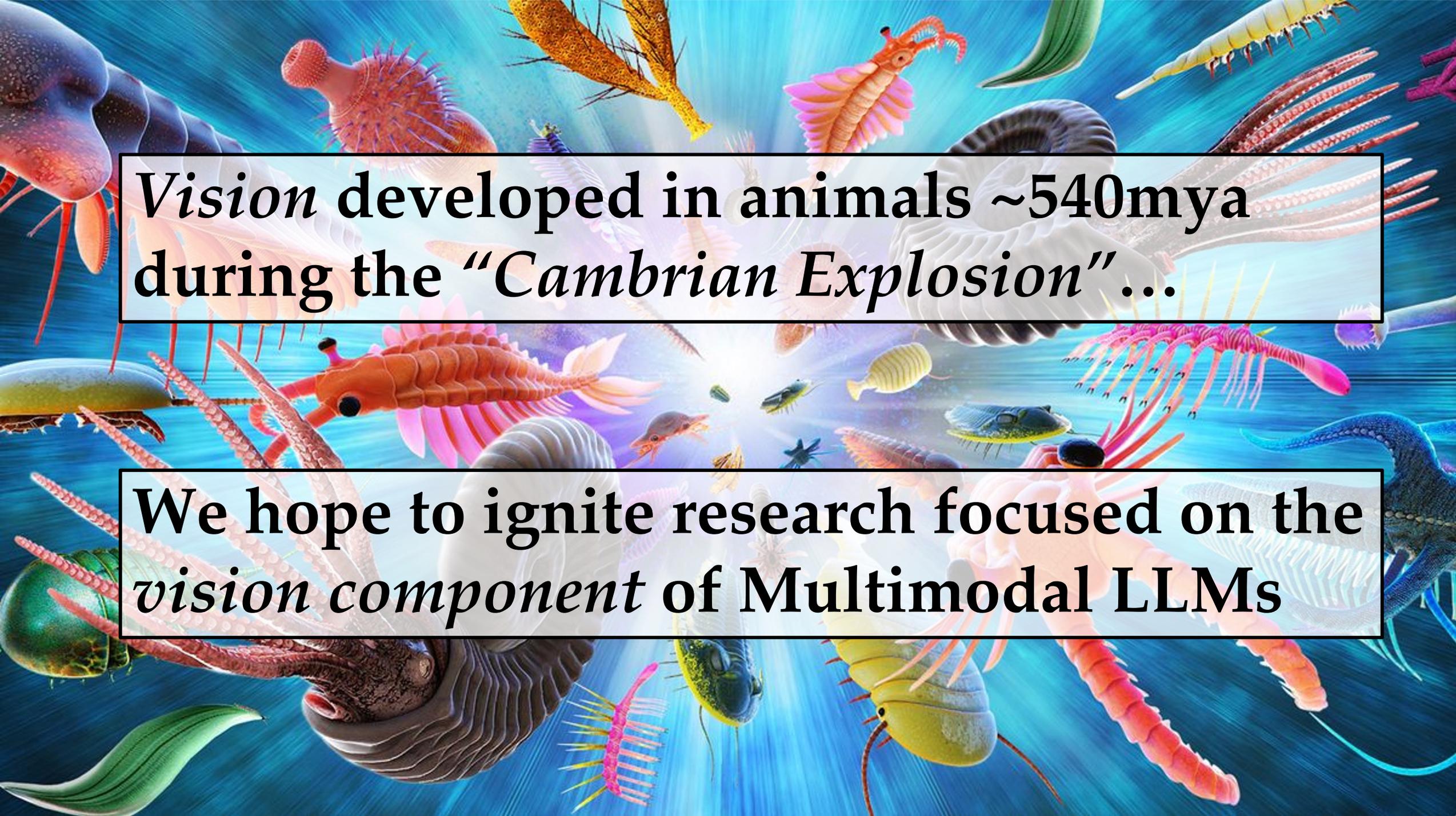
Shengbang Tong*, Ellis Brown*, Penghao Wu*, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, Saining Xie



New York University

NeurIPS 2024





Vision developed in animals ~540mya during the “*Cambrian Explosion*”...

We hope to ignite research focused on the *vision component* of Multimodal LLMs

Why Vision?

Capability	Benchmark	Description	GEMINI 1.0 PRO	GEMINI 1.0 ULTRA	GEMINI 1.5 PRO (Feb 2024)	GEMINI 1.5 FLASH	GEMINI 1.5 PRO (May 2024)
General	MMLU	Representation of questions in 57 subjects (incl. STEM, humanities, and others)	71.8%	83.7%	81.9%	78.9%	85.9%
Code	Natural2Code	Python code generation. Held out dataset HumanEval-like, not leaked on the web	69.6%	74.9%	77.7%	77.2%	82.6%
Math	MATH	Challenging math problems (incl. algebra, geometry, pre-calculus, and others)	32.6%	53.2%	58.5%	54.9%	67.7%
Reasoning	GPQA (main)	Challenging dataset of questions written by domain experts in biology, physics, and chemistry	27.9%	35.7%	41.5%	39.5%	46.2%
	Big-Bench Hard	Diverse set of challenging tasks requiring multi-step reasoning	75.0%	83.6%	84.0%	85.5%	89.2%
Multilingual	WMT23	Language translation	71.7	74.4	75.2	74.1	75.3
Image	MMMU	Multi-discipline college-level reasoning problems	47.9%	59.4%	58.5%	56.1%	62.2%
	MathVista	Mathematical reasoning in visual contexts	46.6%	53.0%	54.7%	58.4%	63.9%
Audio	FLEURS (55 languages)	Automatic speech recognition (based on word error rate, lower is better)	6.4%	6.0%	6.6%	9.8%	6.5%
Video	EgoSchema	Video question answering	55.7%	61.5%	65.1%	65.7%	72.2%

 Really Good Performance,
Especially on Knowledge

Why Vision?

ChatGPT 4o

Peter Tong



From camera's perspective, is this dog facing left or right?

ChatGPT

The dog in the image is facing to the right from the camera's perspective.

ChatGPT 4o

Peter Tong



In this image, how many wheel(s) can you see?

ChatGPT

In the image, two wheels are visible on the car.



Q: Do you see any window in this image?

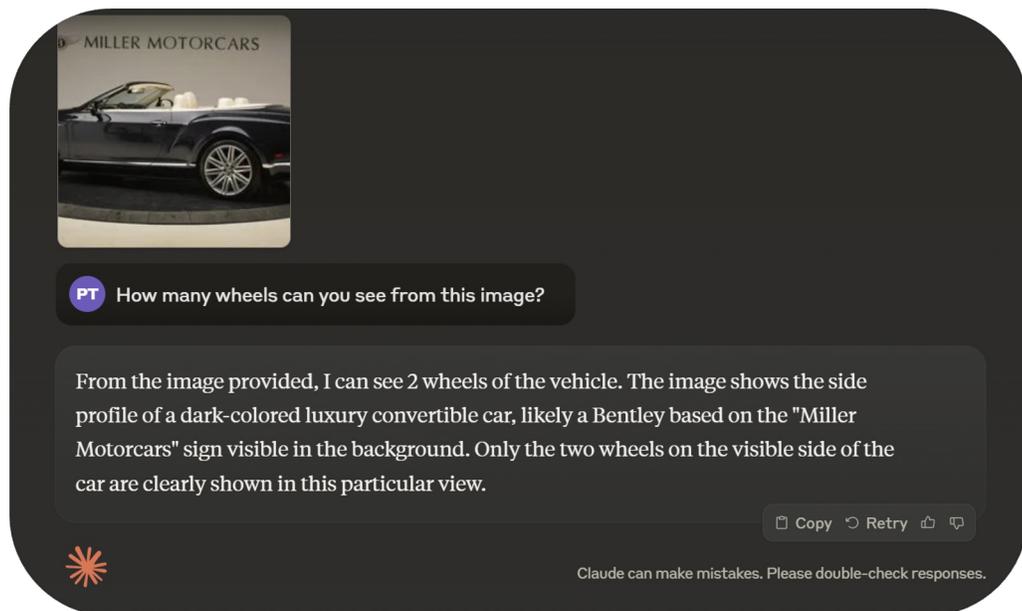
No, there are **no windows visible** in this image. The image shows a traditional red pillar box, which is a type of mailbox found in the United Kingdom.



Q: Is the door of the truck cab open?

No, the door of the truck is **not open** in the image provided.

Why Vision?



MILLER MOTORCARS

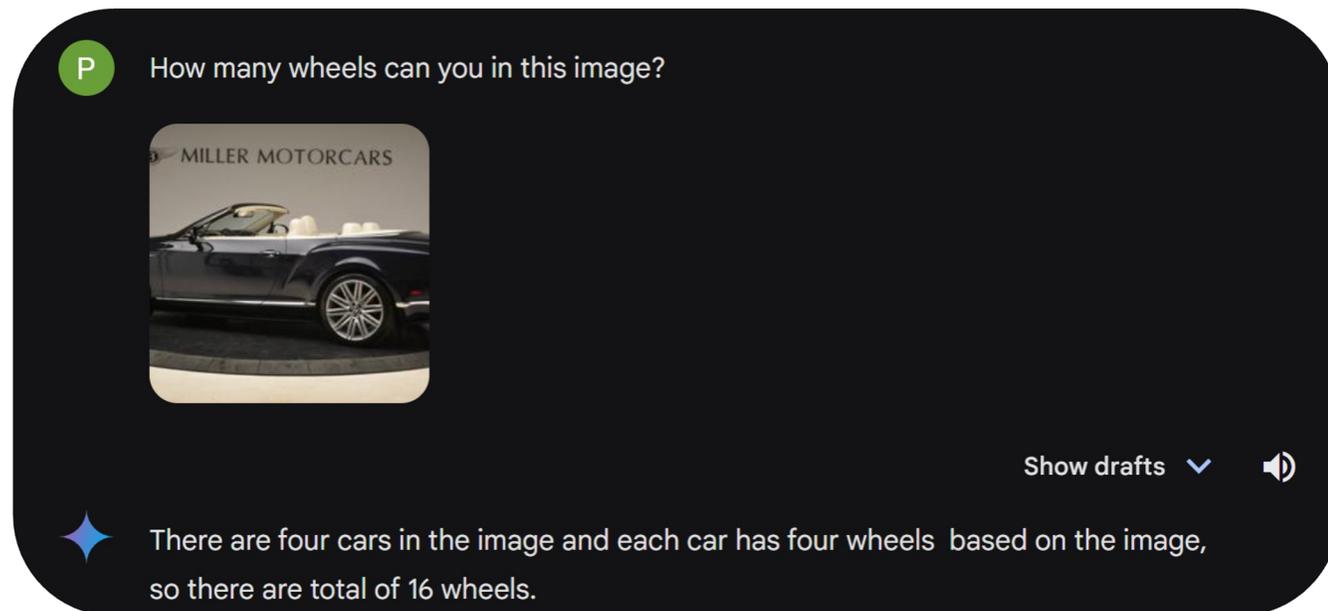


PT How many wheels can you see from this image?

From the image provided, I can see 2 wheels of the vehicle. The image shows the side profile of a dark-colored luxury convertible car, likely a Bentley based on the "Miller Motorcars" sign visible in the background. Only the two wheels on the visible side of the car are clearly shown in this particular view.

Copy Retry Like

Claude can make mistakes. Please double-check responses.



P How many wheels can you in this image?



Show drafts

There are four cars in the image and each car has four wheels based on the image, so there are total of 16 wheels.

Why Vision?



PT From the camera's perspective is the dog facing left or right?

From the camera's perspective, the dog is facing to the right. The dog appears to be a Corgi or Corgi mix, standing in a field full of yellow flowers, likely dandelions. The dog's body is oriented with its head towards the right side of the image, while its tail is visible on the left side.

Copy Retry Like Reply



Claude can make mistakes. Please double-check responses.

P From the camera's perspective, is this dog facing left or right?



Show drafts ▾ 🔊

✦ Based on the image you sent, the dog is facing its left relative to the camera. Its right ear is visible and its left ear is obscured by its head.

Overview: Core Components in MLLMs

Architecture

- LLM
- Vision Backbone
- Vision-Language Connector

Training

- Instruction Tuning Data
- Instruction Tuning Recipe
- Evaluation Protocol



LLM



Visual
Representations



Connector Design



Instruction Tuning
Data



Instruction Tuning
Recipe



Evaluation Protocol

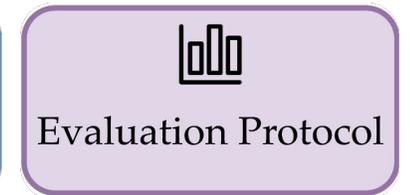
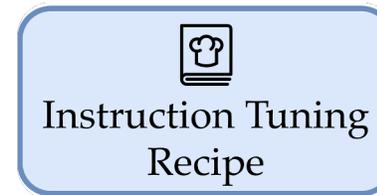
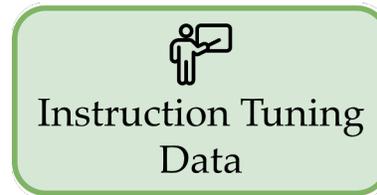
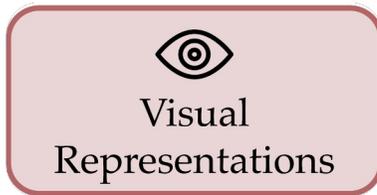
Overview: Core Components in MLLMs

Architecture

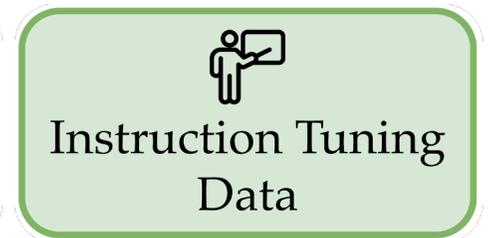
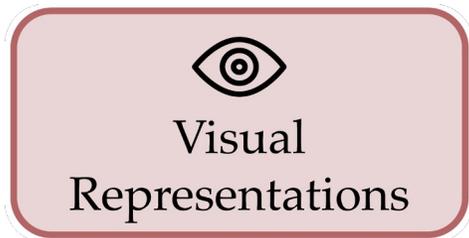
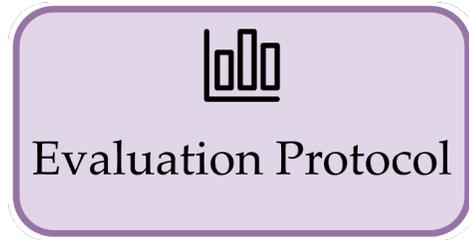
- ~~LLM~~
- Vision Backbone
- Vision-Language Connector

Training

- Instruction Tuning Data
- Instruction Tuning Recipe
- Evaluation Protocol



Overview: Core Components in MLLMs



Overview



Evaluation Protocol



Instruction Tuning
Recipe



Visual
Representations

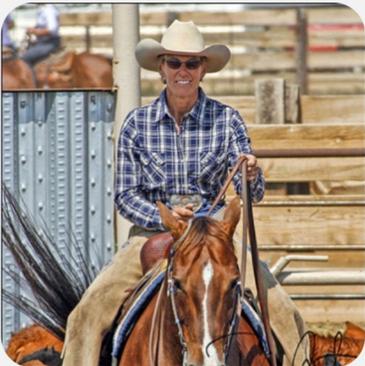


Connector Design



Instruction Tuning
Data

Visual Representations



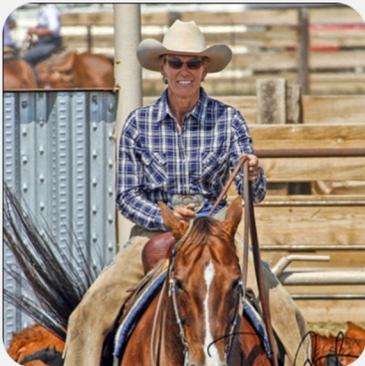
Cowboy Hat
Sorrel
Cowboy Boot
Barrel
Revolver

Class Label Supervised
ImageNet-1K [105]



A cowboy rides a horse at a rodeo.

Language Supervised
CLIP [102]

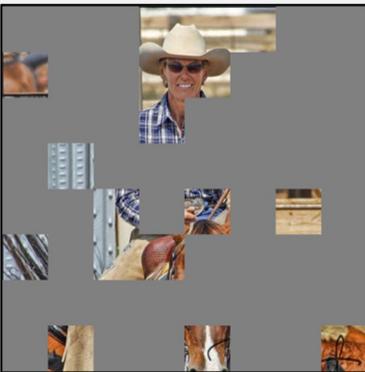


Pos.

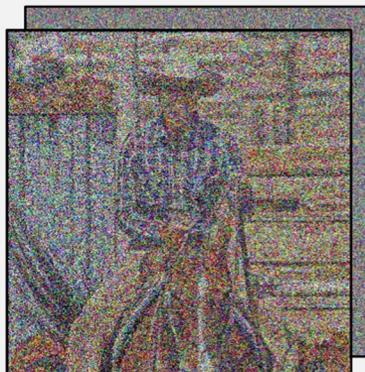


Neg.

SSL-Contrastive
DINOv2 [96]



SSL-Masking
MAE [45]



Diffusion
Stable Diffusion [104]



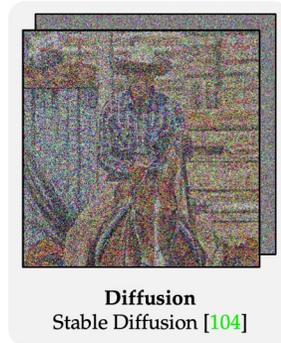
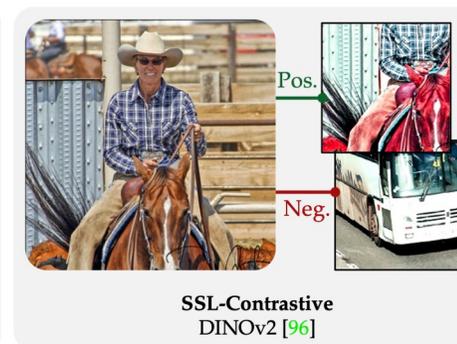
Depth Supervised
MiDaS [13]



Segmentation Supervised
SAM [61]

Visual Representations

Supervision Type	Method	Architecture	Patch Size	Res.	# Tok.
Language-Supervised					
Language	OpenAI CLIP	ViT-L	14	336	576
	DFN-CLIP	ViT-L	14	224	256
	DFN-CLIP	ViT-H	14	378	729
	EVA-CLIP-02	ViT-L	14	336	576
	SigLIP	ViT-L	16	384	576
	SigLIP	ViT-SO400M	14	384	729
	OpenCLIP	ConvNeXT-L	-	512	1576
	OpenCLIP	ConvNeXT-L	-	1024	1576
	OpenCLIP	ConvNeXT-XXL	-	1024	1576
Self-Supervised					
Contrastive	DINOv2	ViT-L	14	336	576
	DINOv2	ViT-L	14	518	1576
	MoCo v3	ViT-B	16	224	196
	MoCo v3	ViT-L	16	224	196
Masked	MAE	ViT-L	16	224	196
	MAE	ViT-H	14	224	256
JEPA	I-JEPA	ViT-H	14	224	256
Other					
Segmentation	SAM	ViT-L	16	1024	1576
	SAM	ViT-L	16	1024	1576
Depth	MiDaS 3.0	ViT-L	16	384	576
	MiDaS 3.1	ViT-L	16	518	1024
Diffusion	Stable Diffusion 2.1	VAE+UNet	16	512	1024
Class Labels	SupViT	ViT-L	16	224	196
	SupViT	ViT-H	14	224	256



How to evaluate different representations?

23 models!

Overview



Evaluation Protocol



Instruction Tuning
Recipe



Visual
Representations

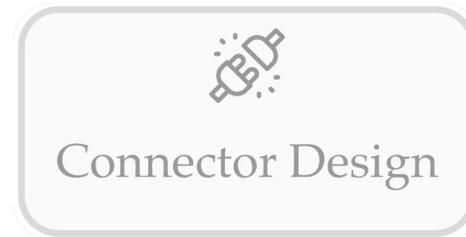
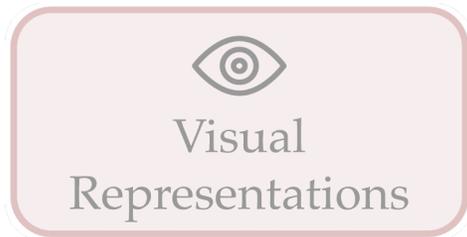
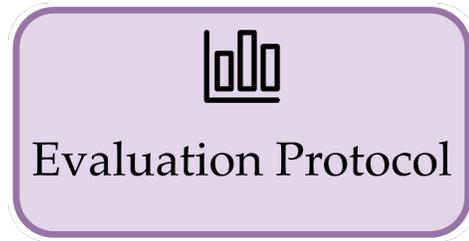


Connector Design



Instruction Tuning
Data

Overview

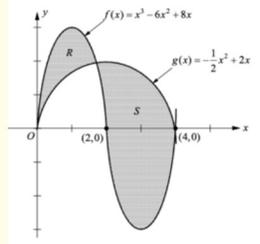


Evaluation Protocol

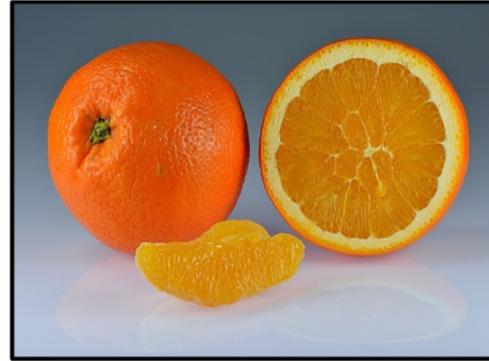
Question: *<image 1>* The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R.

Options:

- (A) $\int_0^{1.5} [f(x) - g(x)] dx$
 (B) $\int_0^{1.5} [g(x) - f(x)] dx$
 (C) $\int_0^2 [f(x) - g(x)] dx$
 (D) $\int_0^2 [g(x) - x(x)] dx$



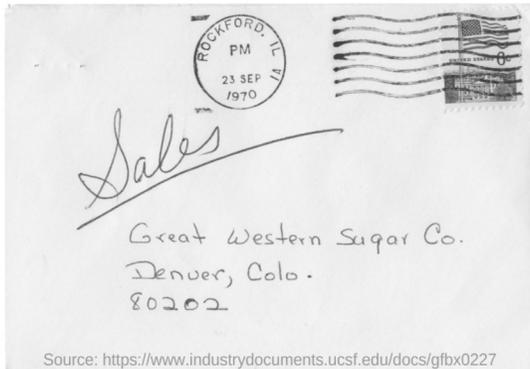
MMMU [Yue, et al. 2024]



Q: what is the color of this object?

- A. Purple
 B. Pink
 C. Gray
 D. Orange
 GT: D

MM-Bench [Liu, et al. 2024]



Q: Mention the ZIP code written?

A: 80202

Q: What date is seen on the seal at the top of the letter?

A: 23 sep 1970

Q: Which company address is mentioned on the letter?

A: Great western sugar Co.

DocVQA [Mathew, et al. 2020]



Where can we go from the current lane? **A. Turn left.** B. Go straight. C. Turn left and go straight. D. Turn right.

RealWorldQA [Grok, et al. 2024]

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



(a) Left

(b) Right

	(b)	(b)	✗
	(a)	(a)	✗
	(b)	(b)	✗
	(a)	(a)	✗

MMVP [Tong, et al. 2024]

and a lot more...

How to systematically interpret the evaluation results & understand model performance?

Benchmark Analysis

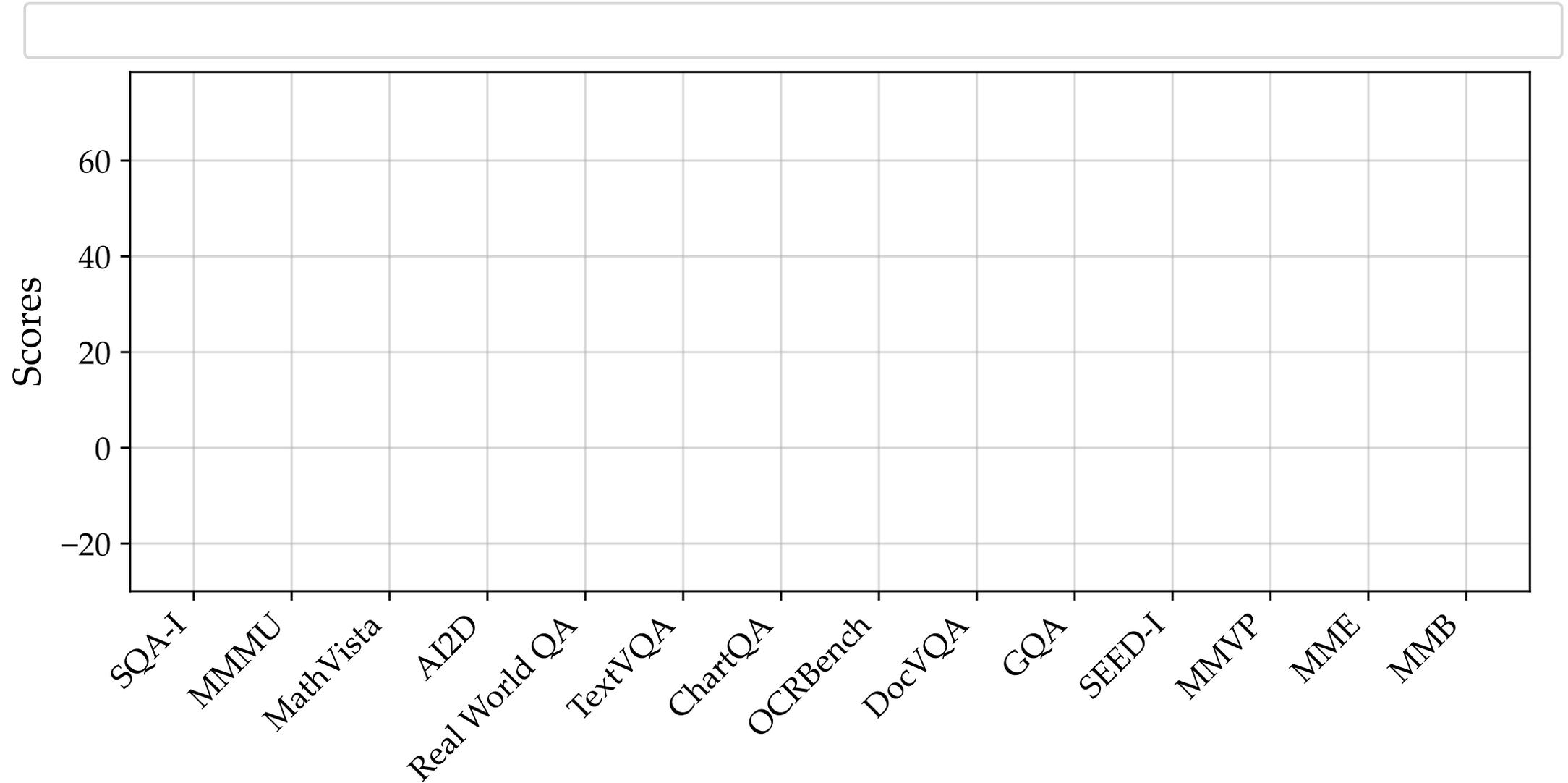
- 1 – Assess the “Multimodality” of the Benchmarks
- 2 – Group Benchmarks into Clusters

Benchmark Analysis

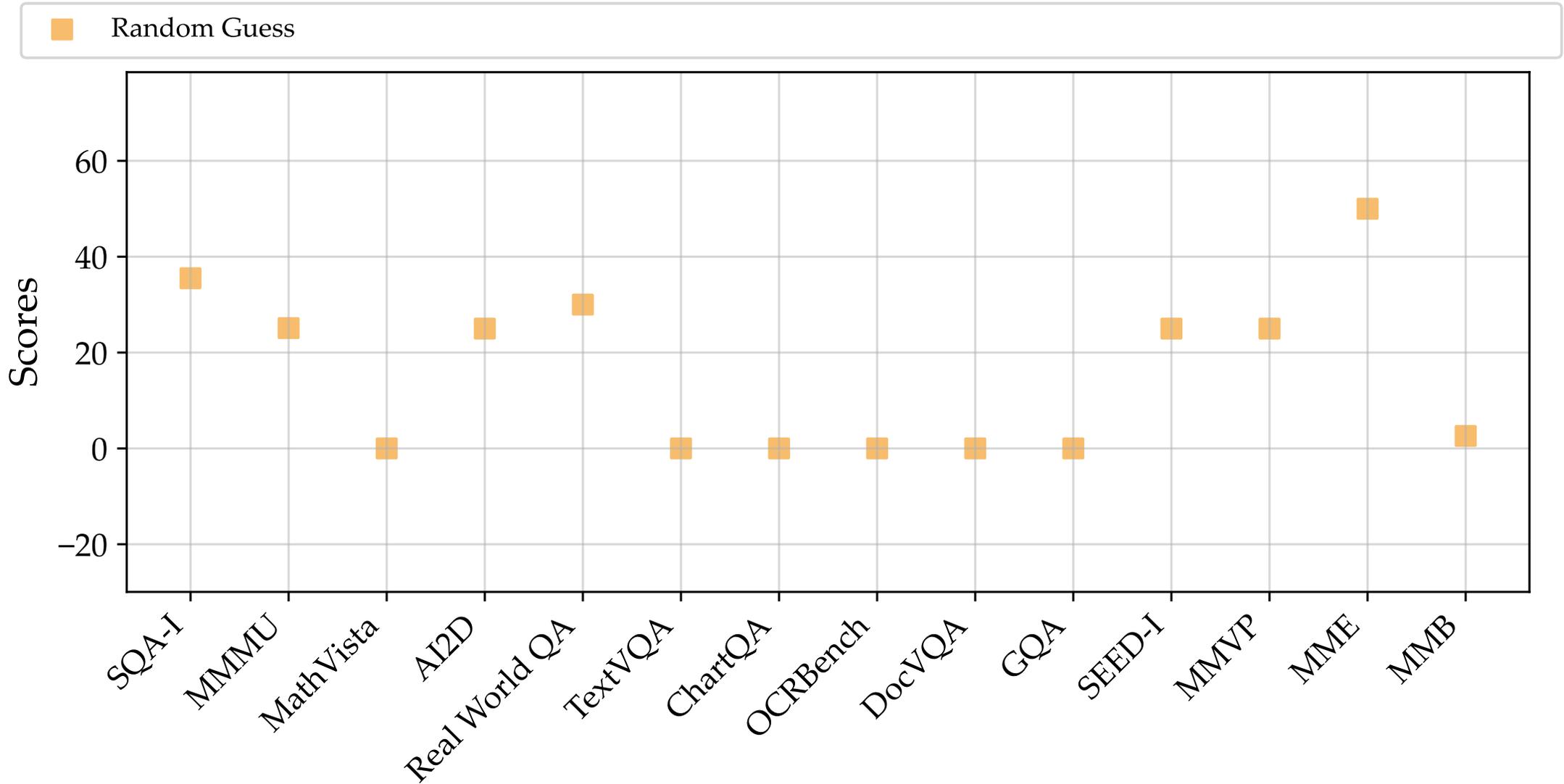
1 – Assess the “Multimodality” of the Benchmarks

2 – Group Benchmarks into Clusters

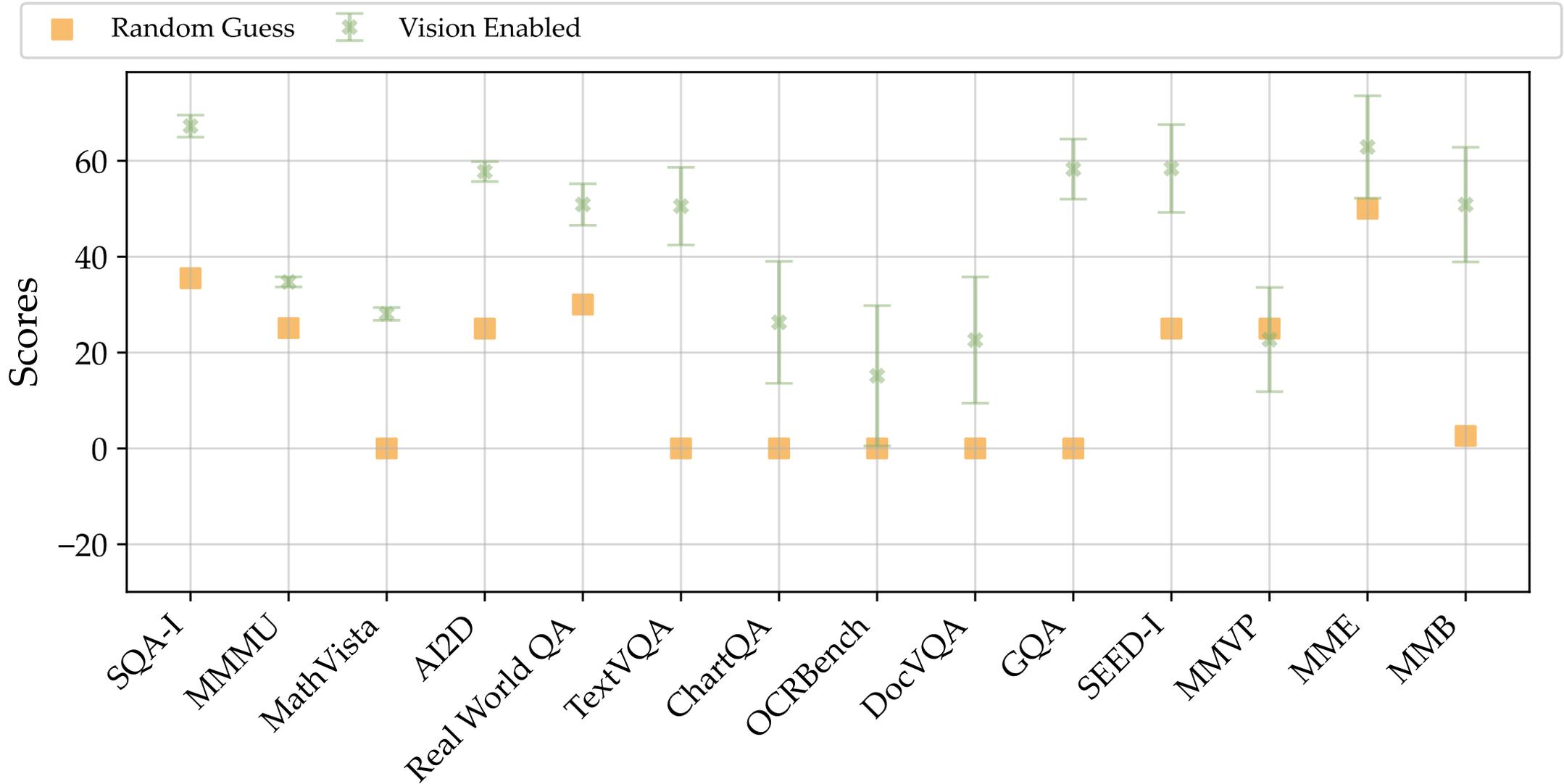
Who's answering the question: the LLM or MLLM?



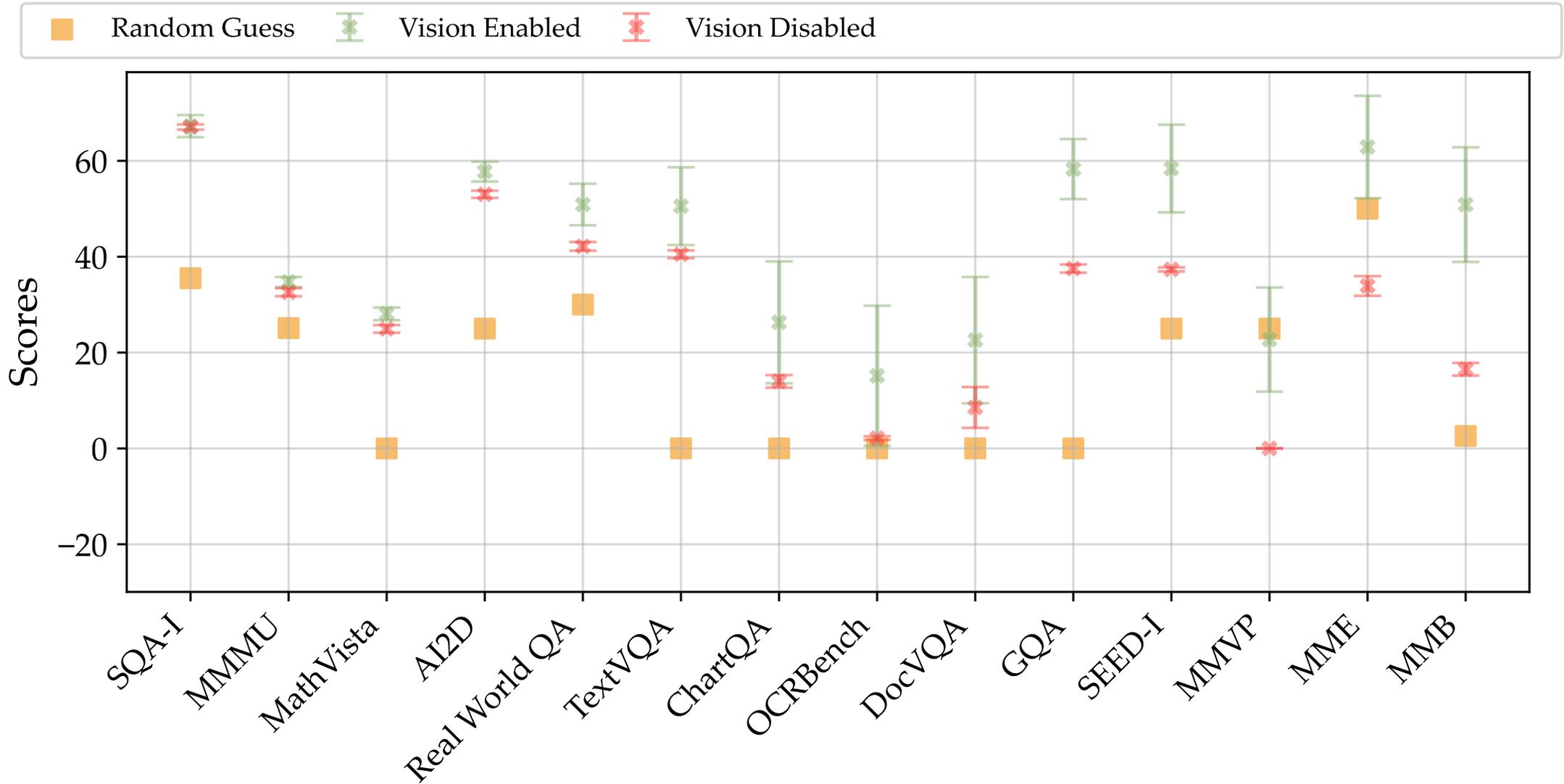
Who's answering the question: the LLM or MLLM?



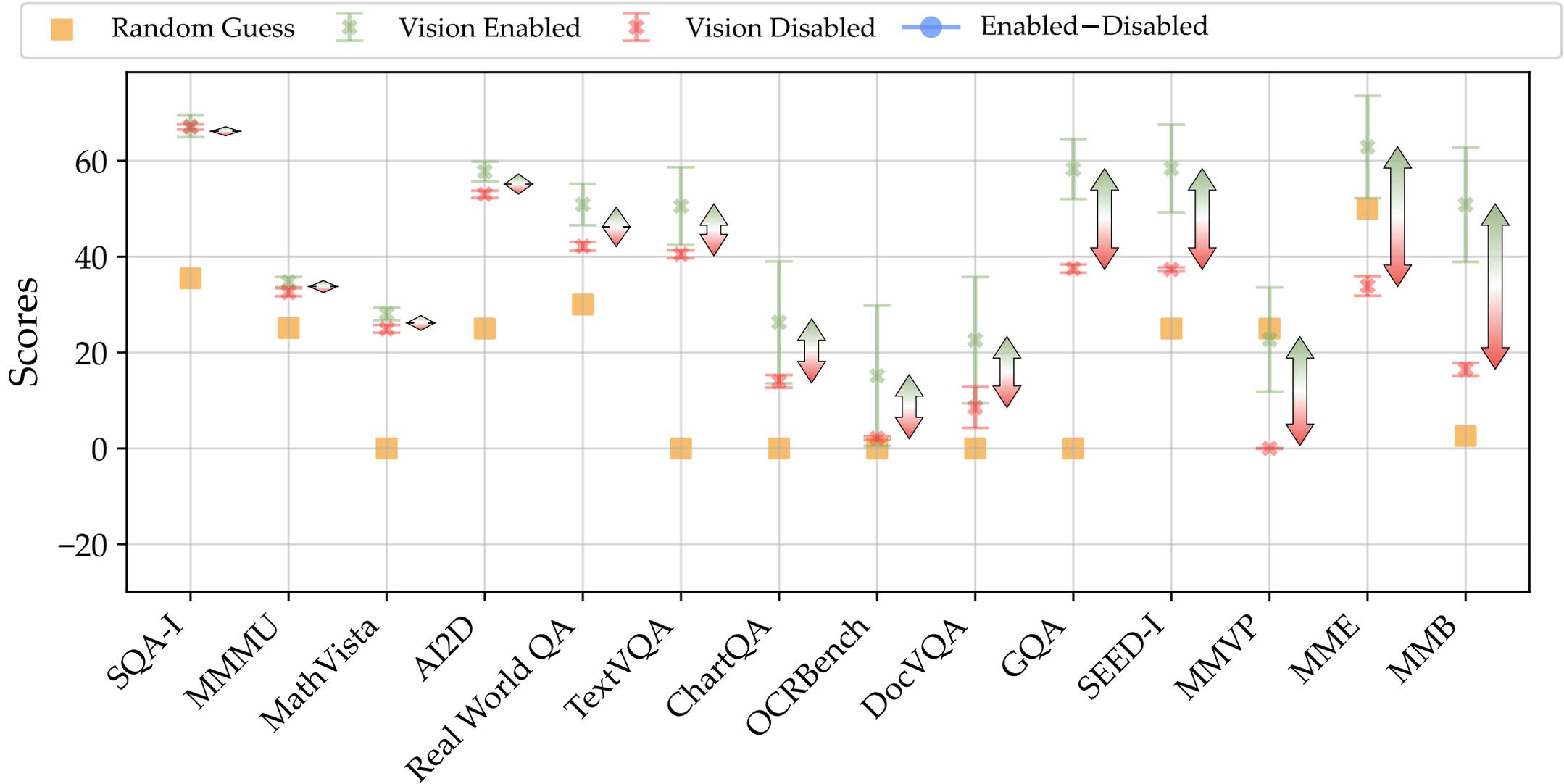
Who's answering the question: the LLM or MLLM?



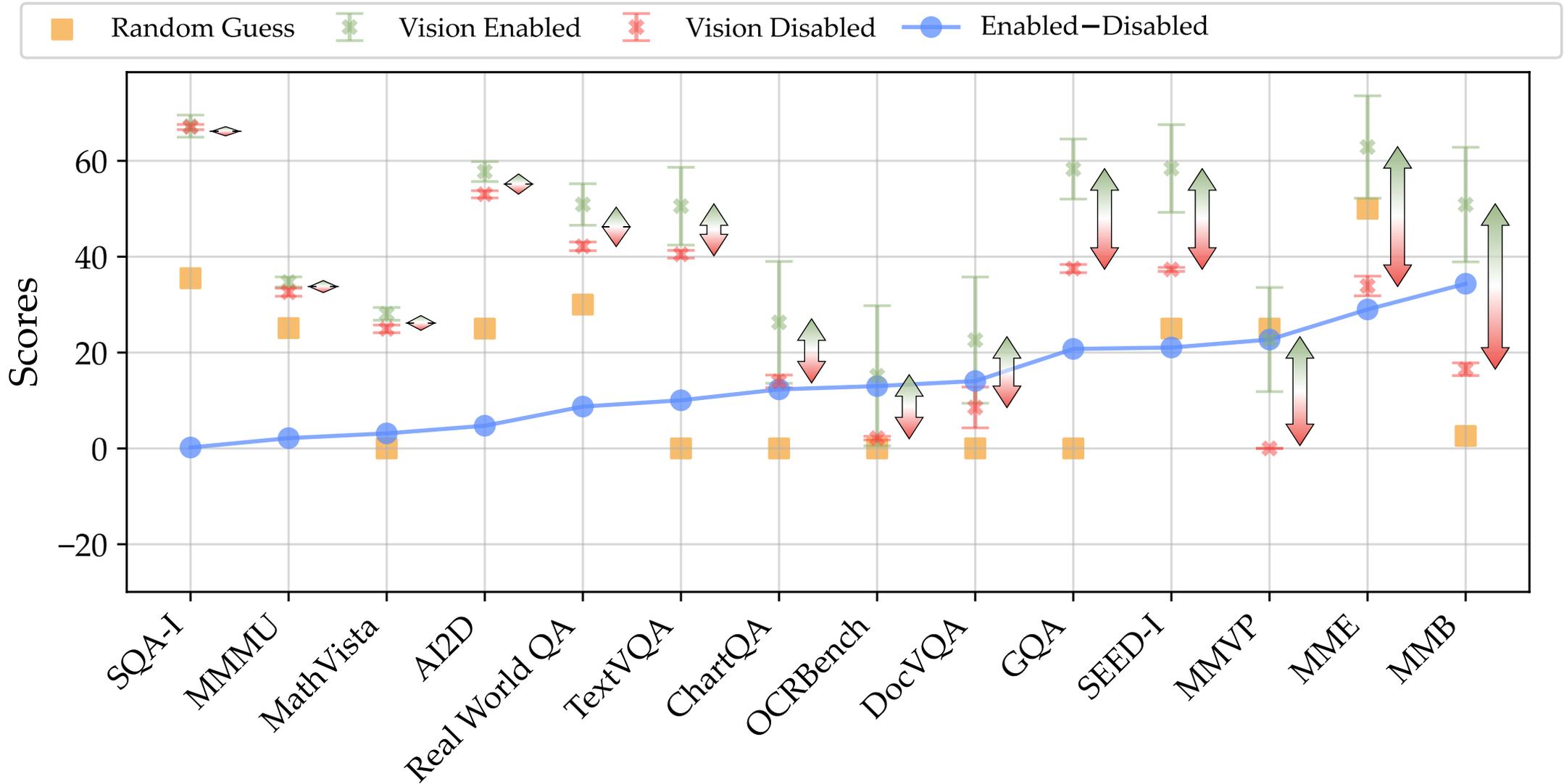
Who's answering the question: the LLM or MLLM?



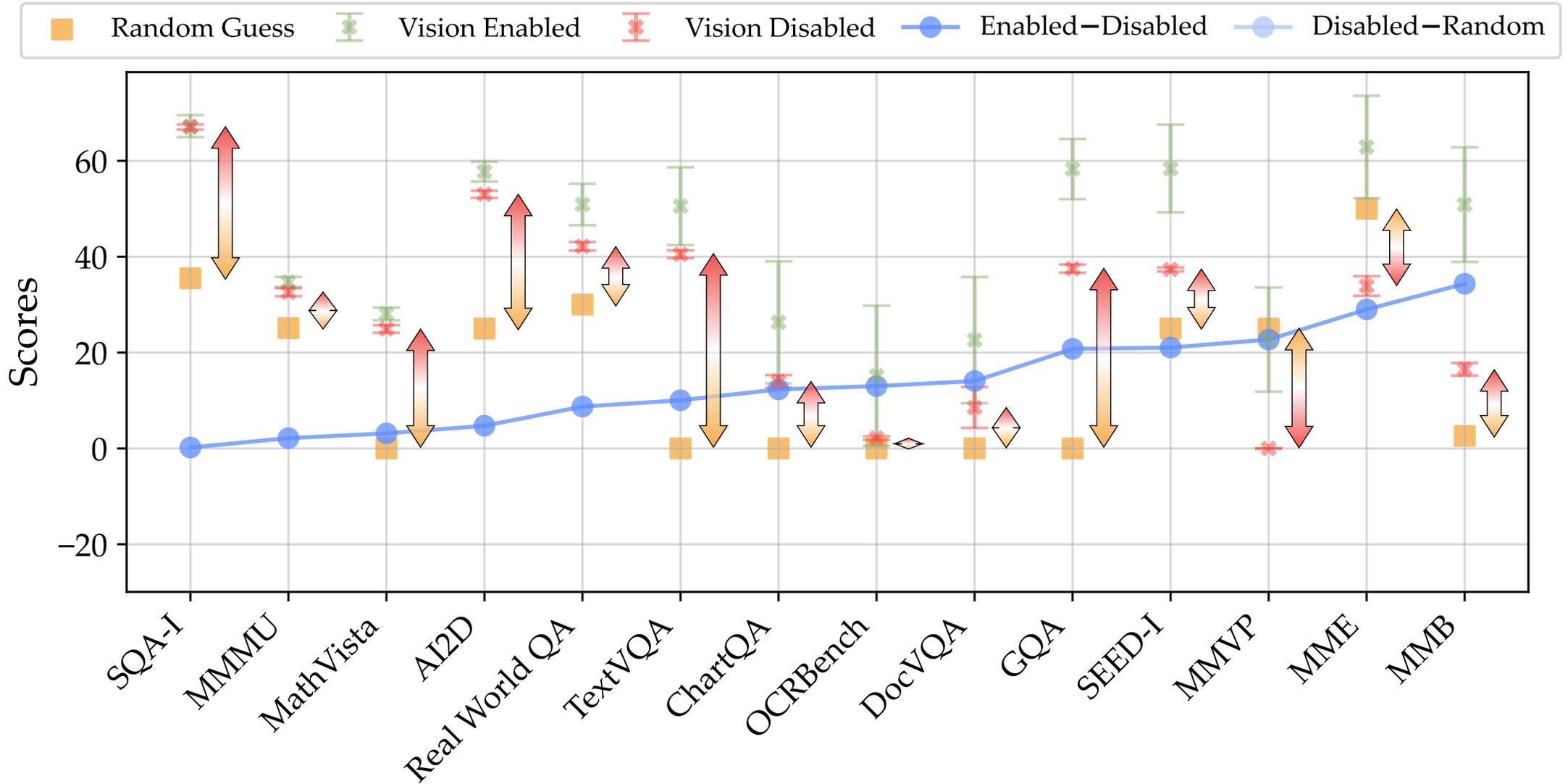
Who's answering the question: the LLM or MLLM?



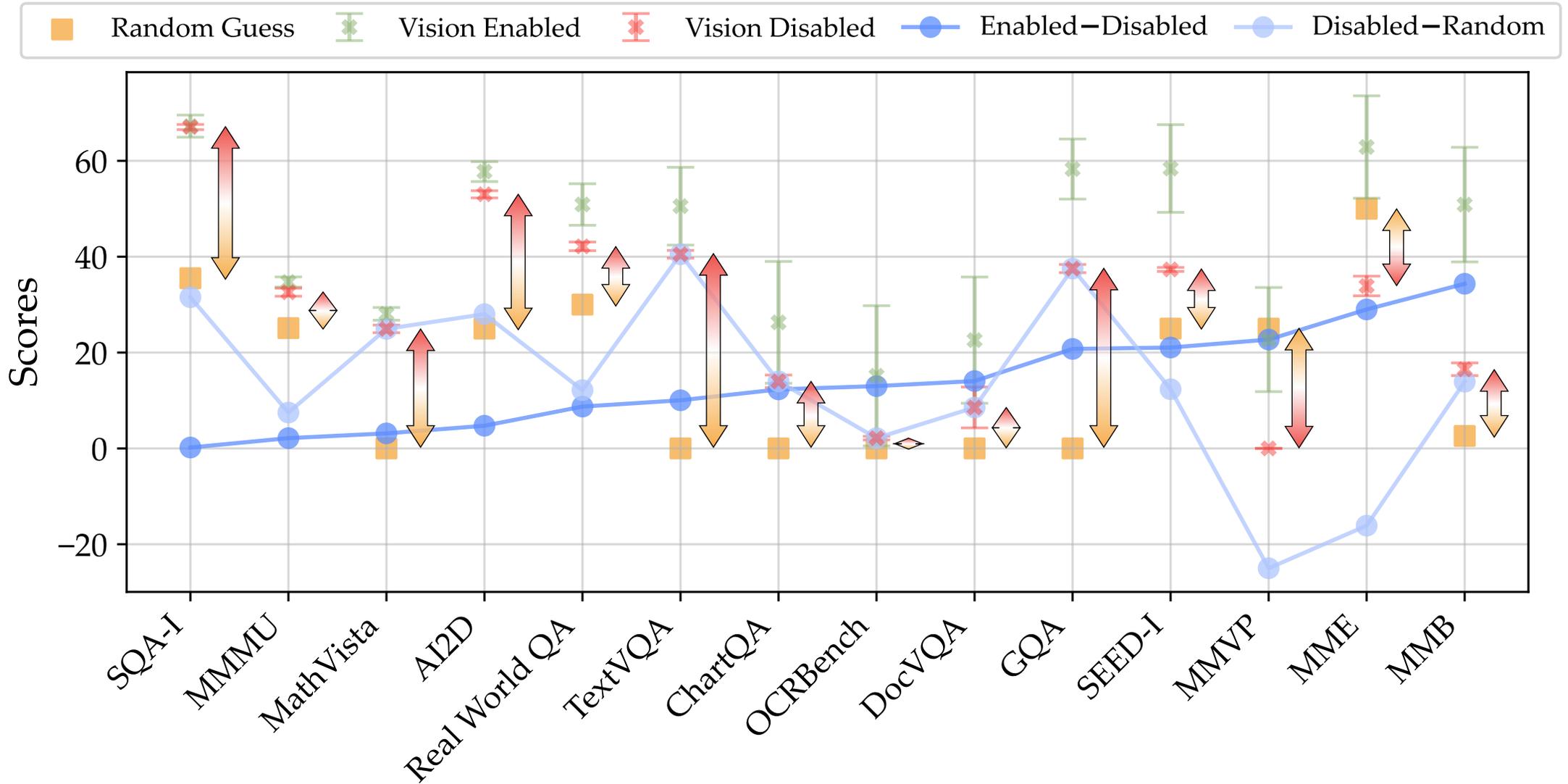
Who's answering the question: the LLM or MLLM?



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



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



Benchmark Analysis

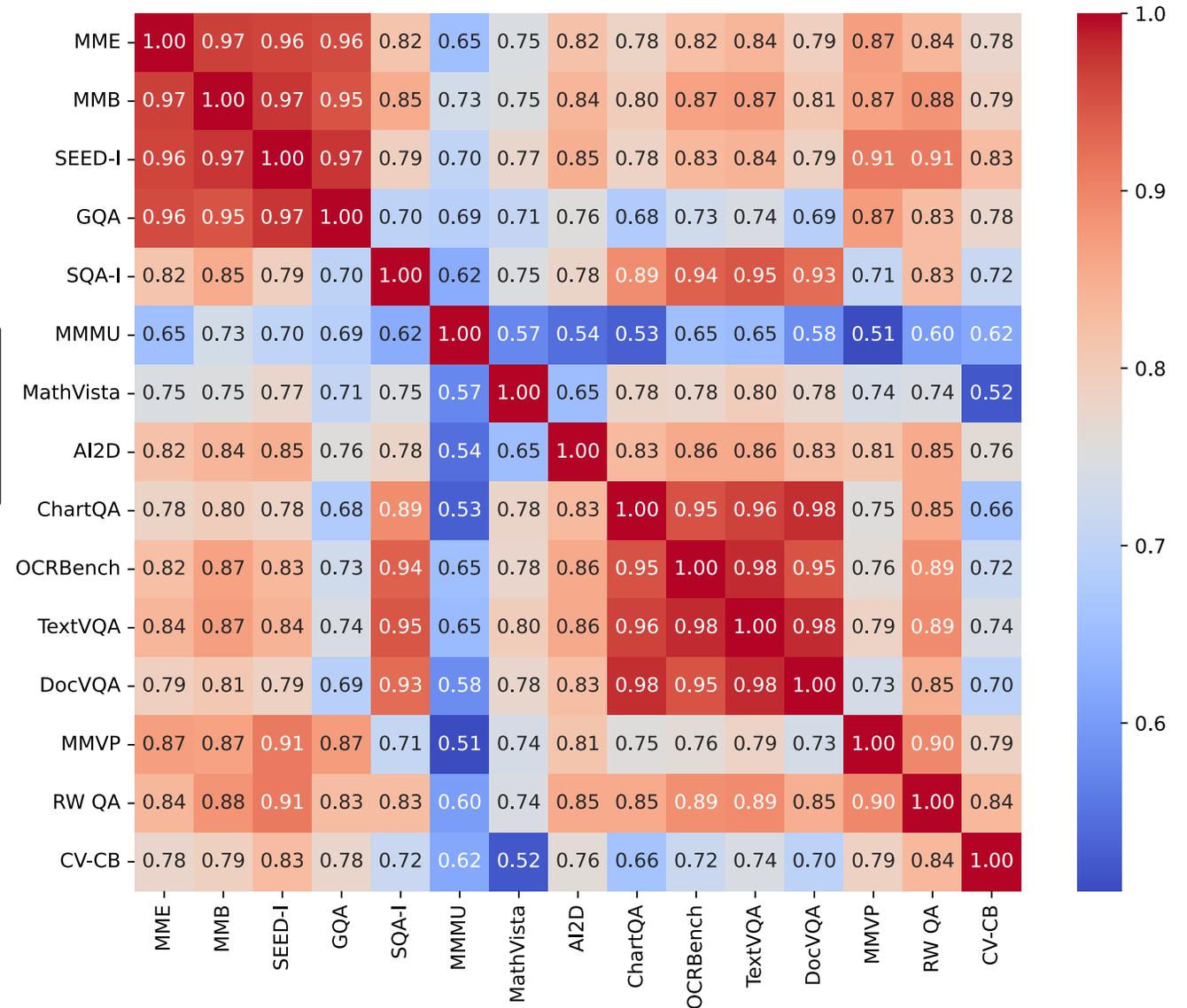
1 – Assess the “Multimodality” of the Benchmarks

2 – Group Benchmarks into Clusters

Group Benchmarks

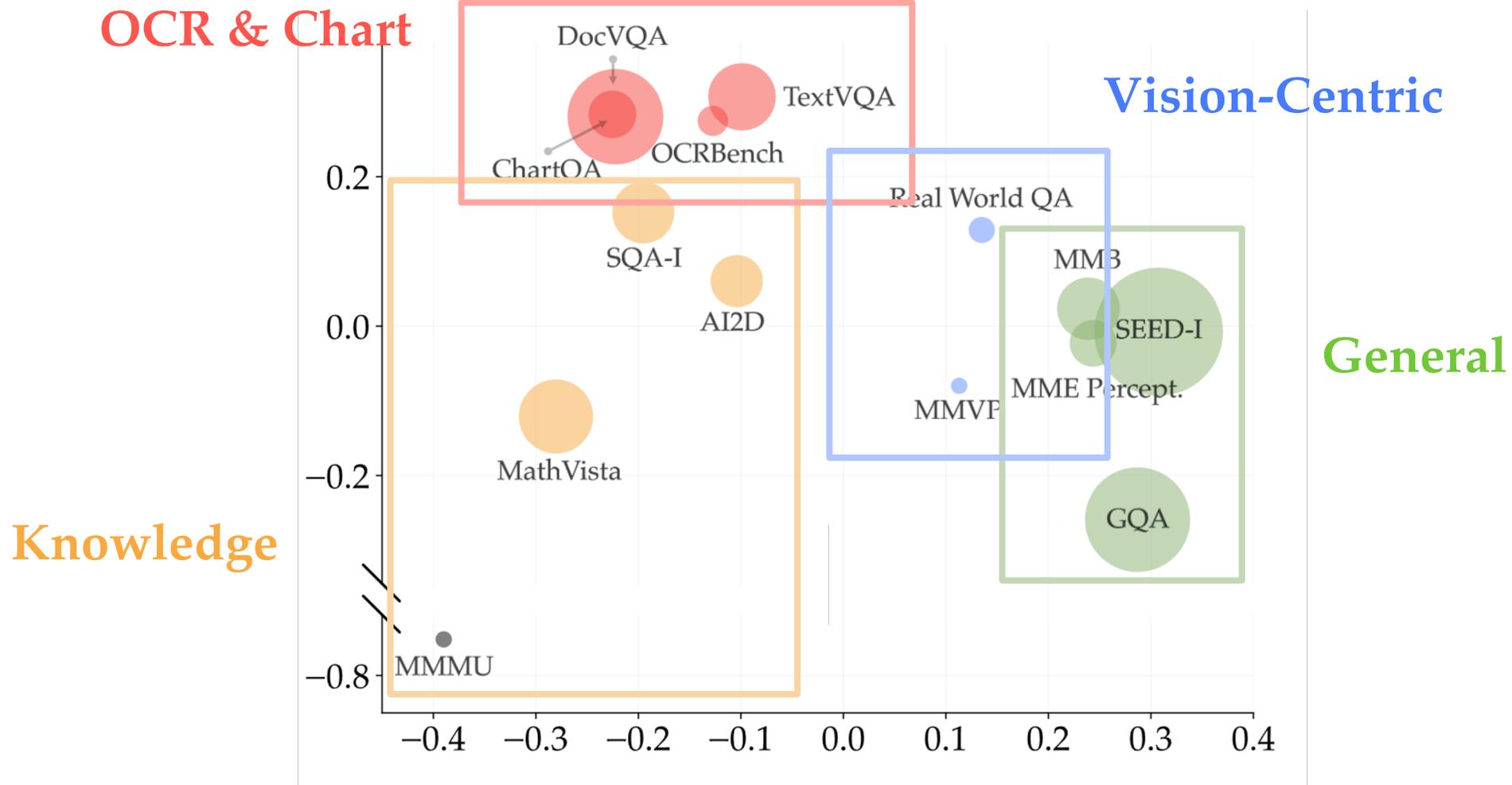


Similar benchmarks \Rightarrow correlated performance across visual encoders

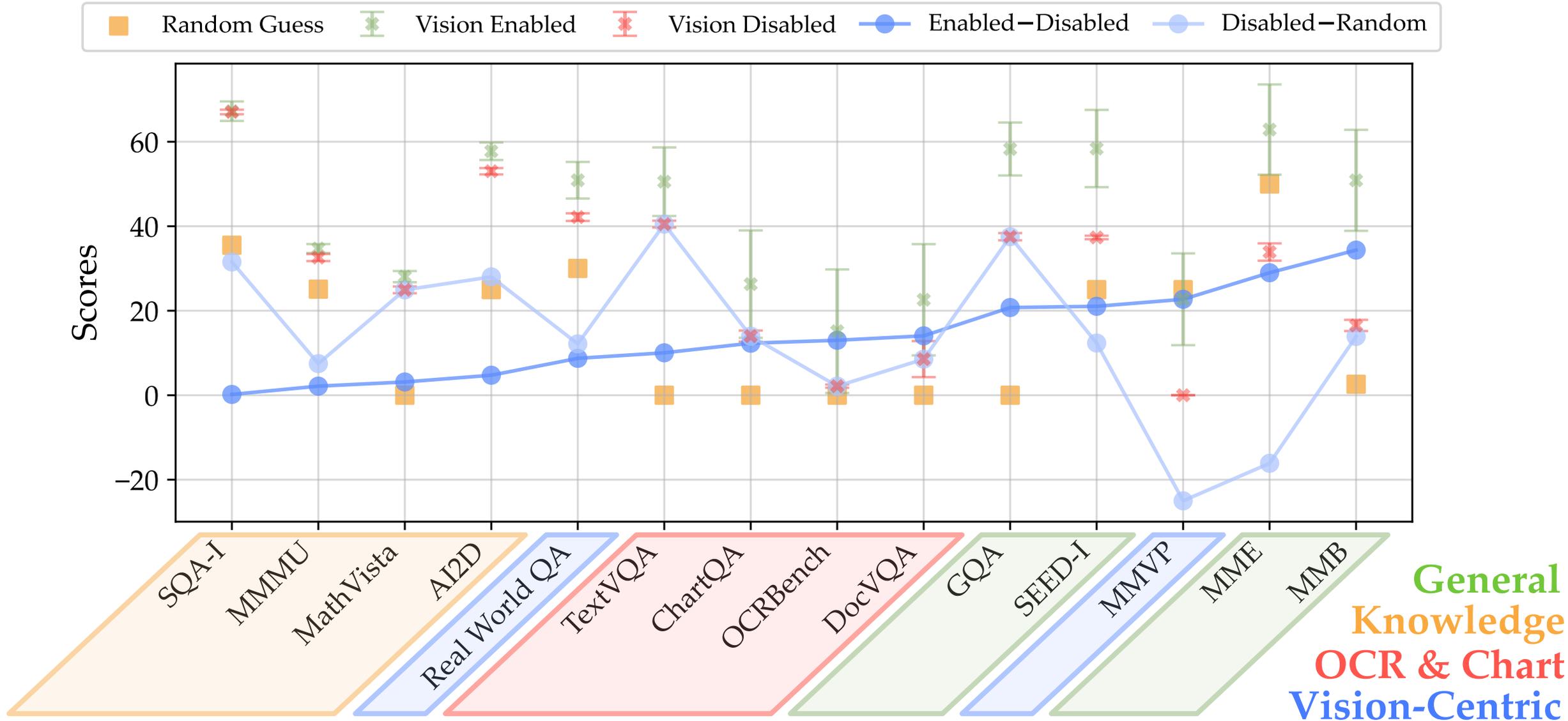


Group Benchmarks

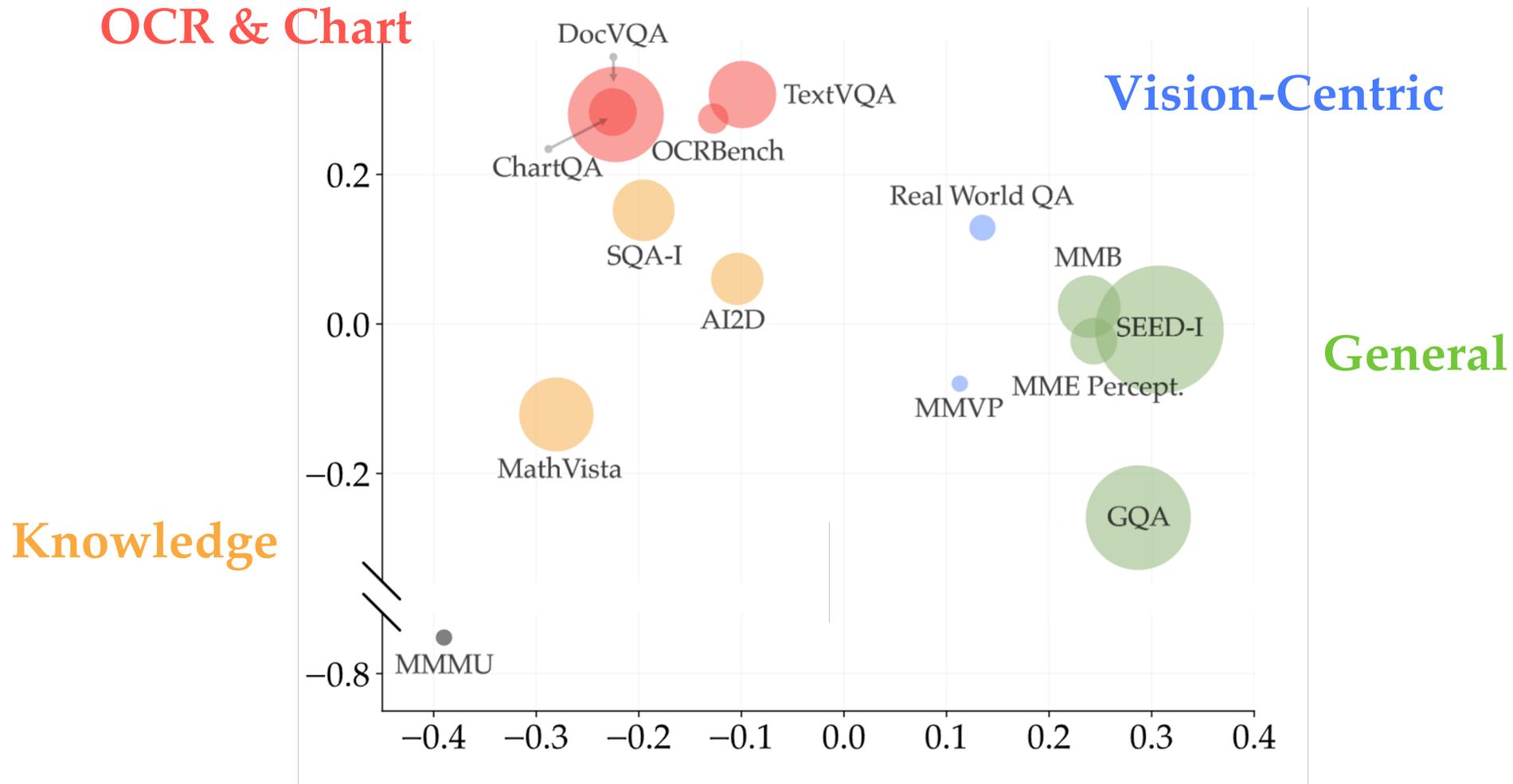
💡 Similar benchmarks \Rightarrow correlated performance across visual encoders



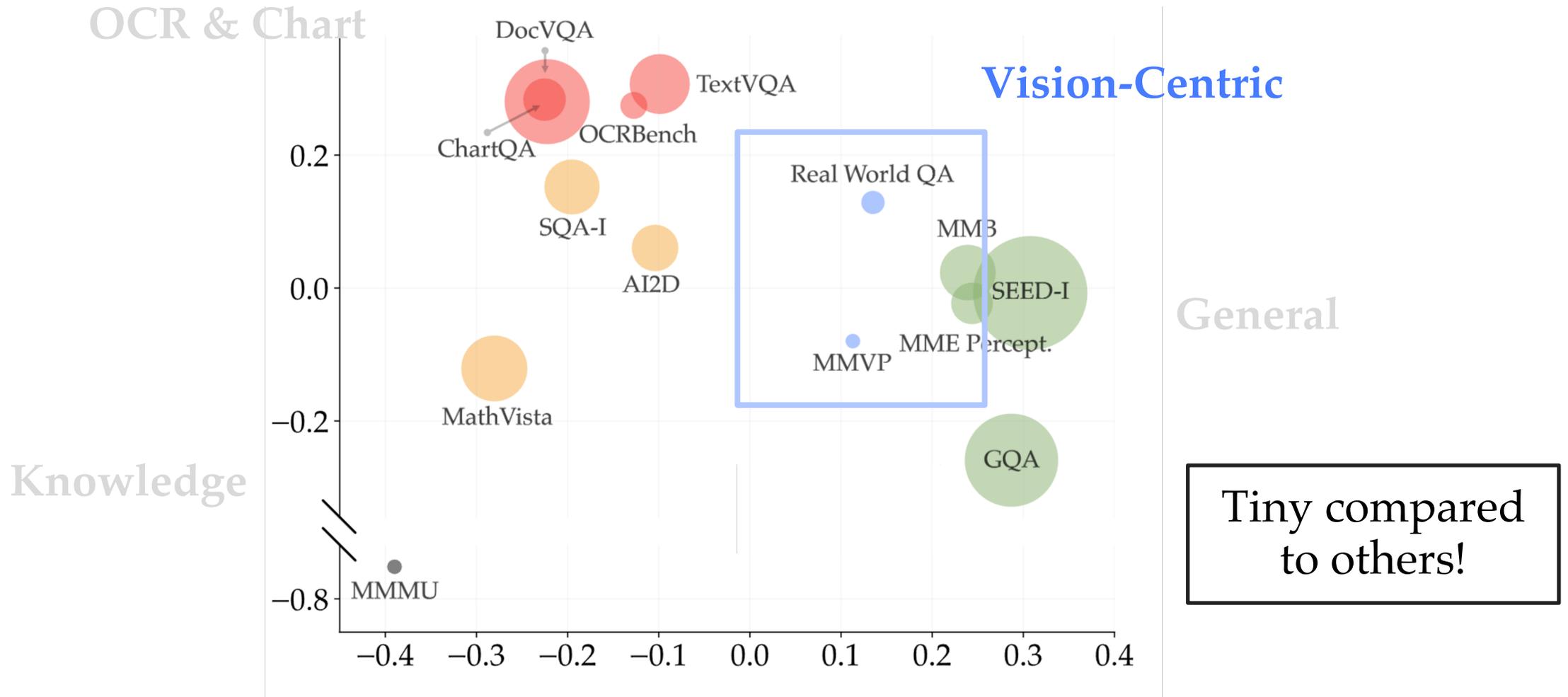
Who's answering the question: the LLM or MLLM?



Group Benchmarks



Group Benchmarks



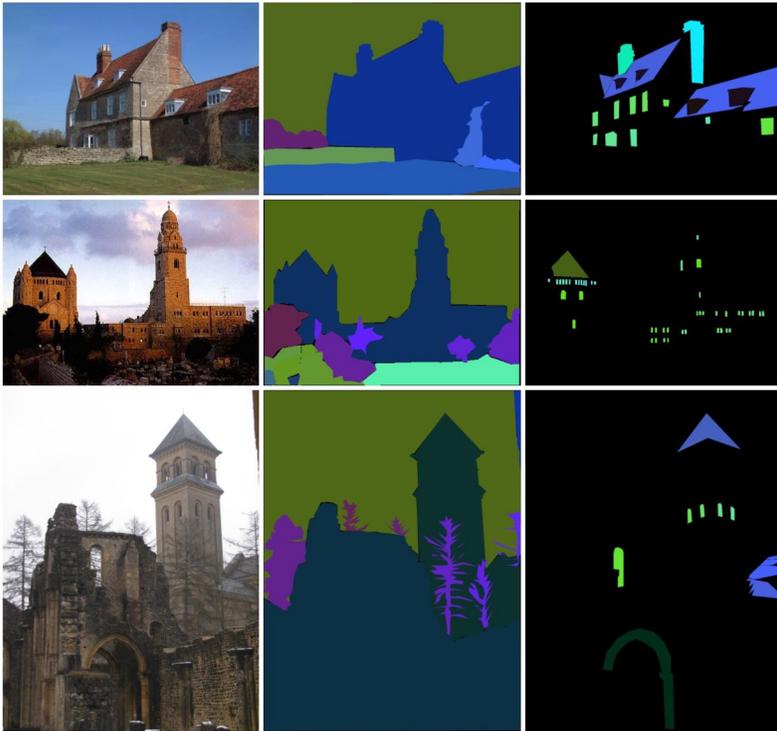
Q: How can we scalably generate *vision-centric* MLLM evaluations?



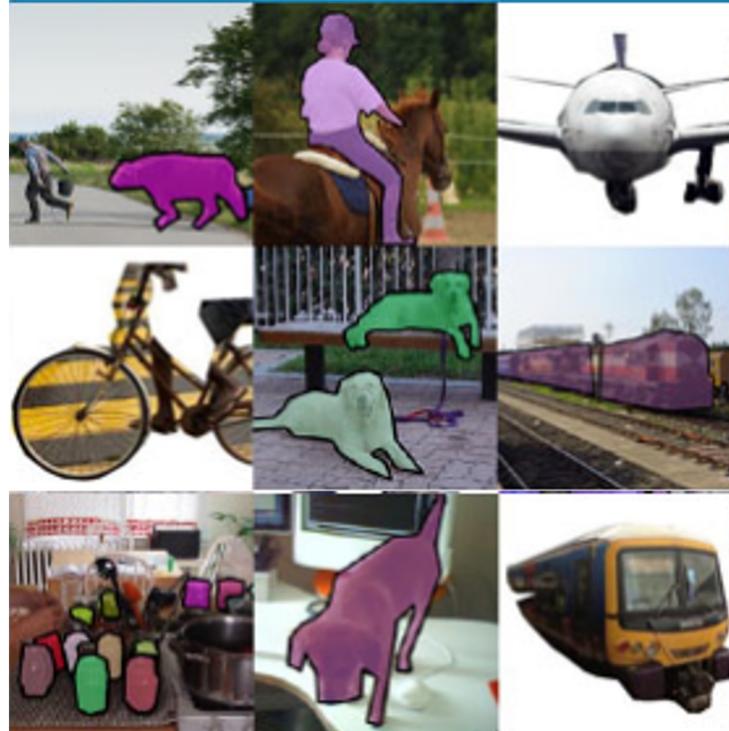
Repurpose existing vision datasets!

Dense GT Annos

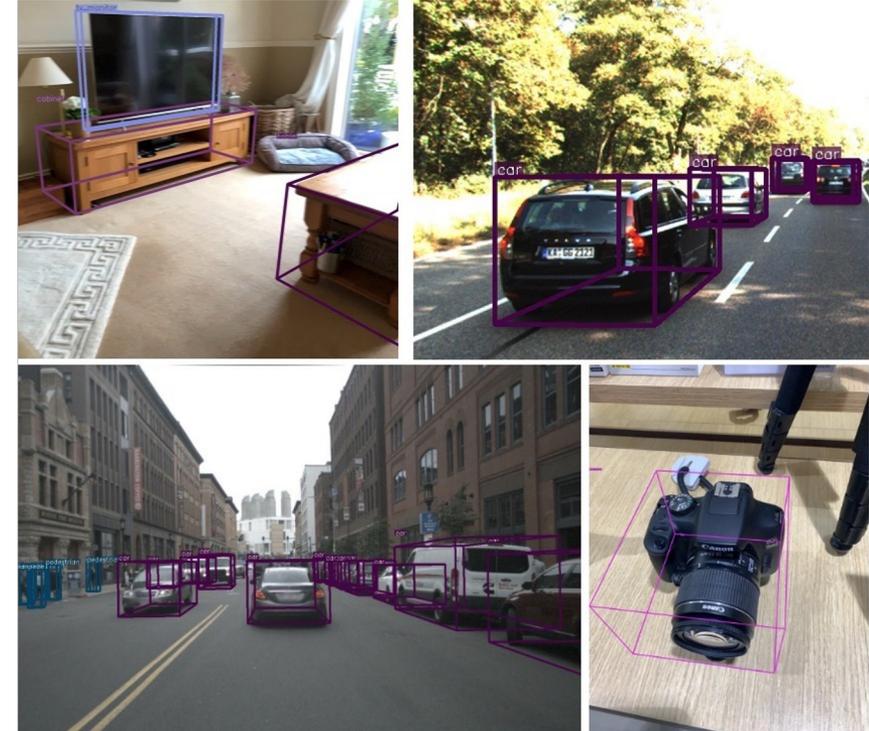
💡 Repurpose existing vision datasets!



ADE20K



MSCOCO



Omni3D

CV-Bench

💡 Repurpose existing vision datasets!

2D

3D

Spatial Relationship



Where is the cave located with respect to the trees?

Object Count



How many cars are in the image?

Depth Order



Which is closer to the camera, **sink** or **pillow**?

Relative Distance



Which is closer to the **chair**, **refrigerator** or **door**?

Source benchmark: ADE20K [145] and COCO [72]

Source benchmark: Omini3D [16]

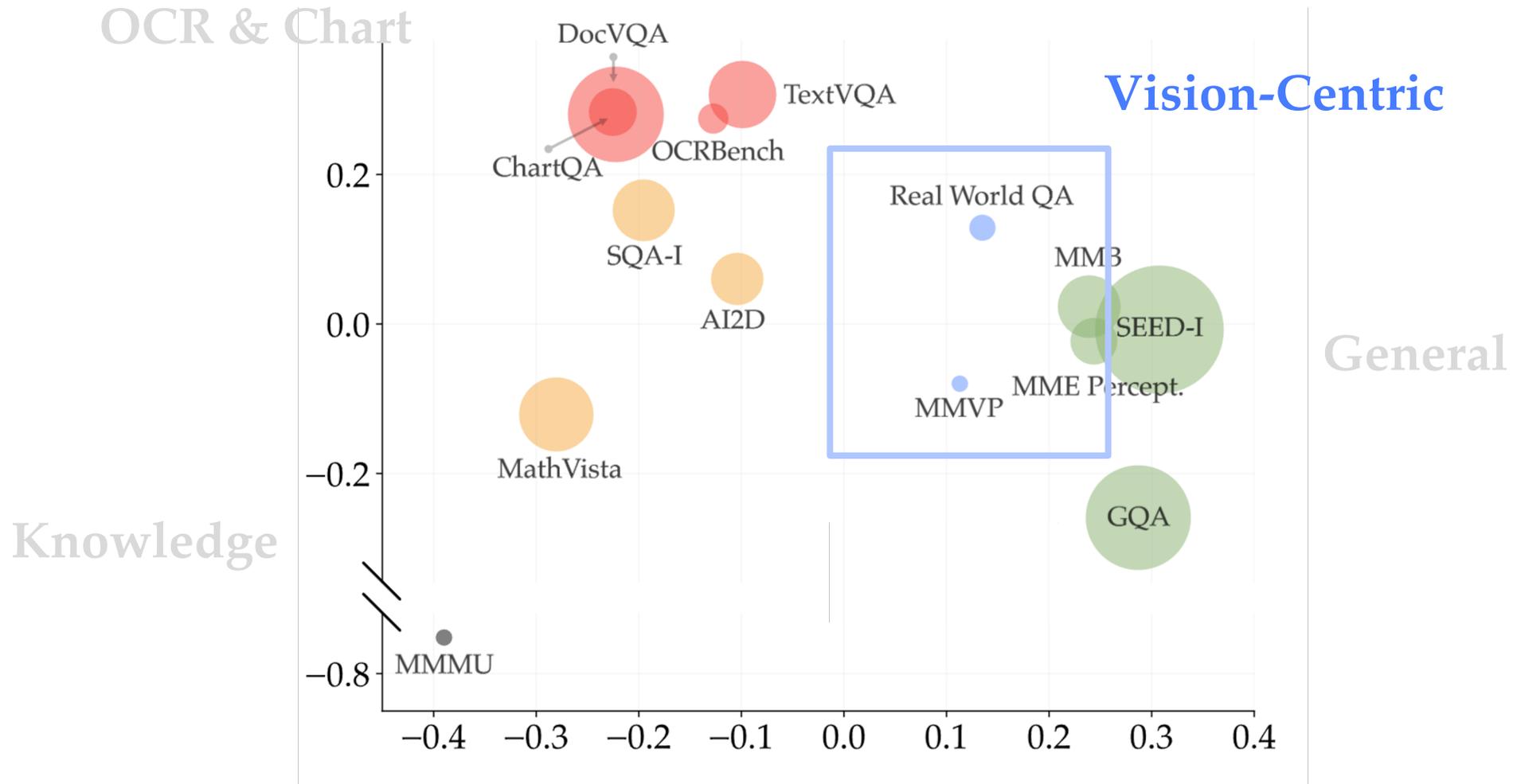
CV-Bench

1. Programmatically construct VQA questions using dataset GT annos
2. Manually filter all examples

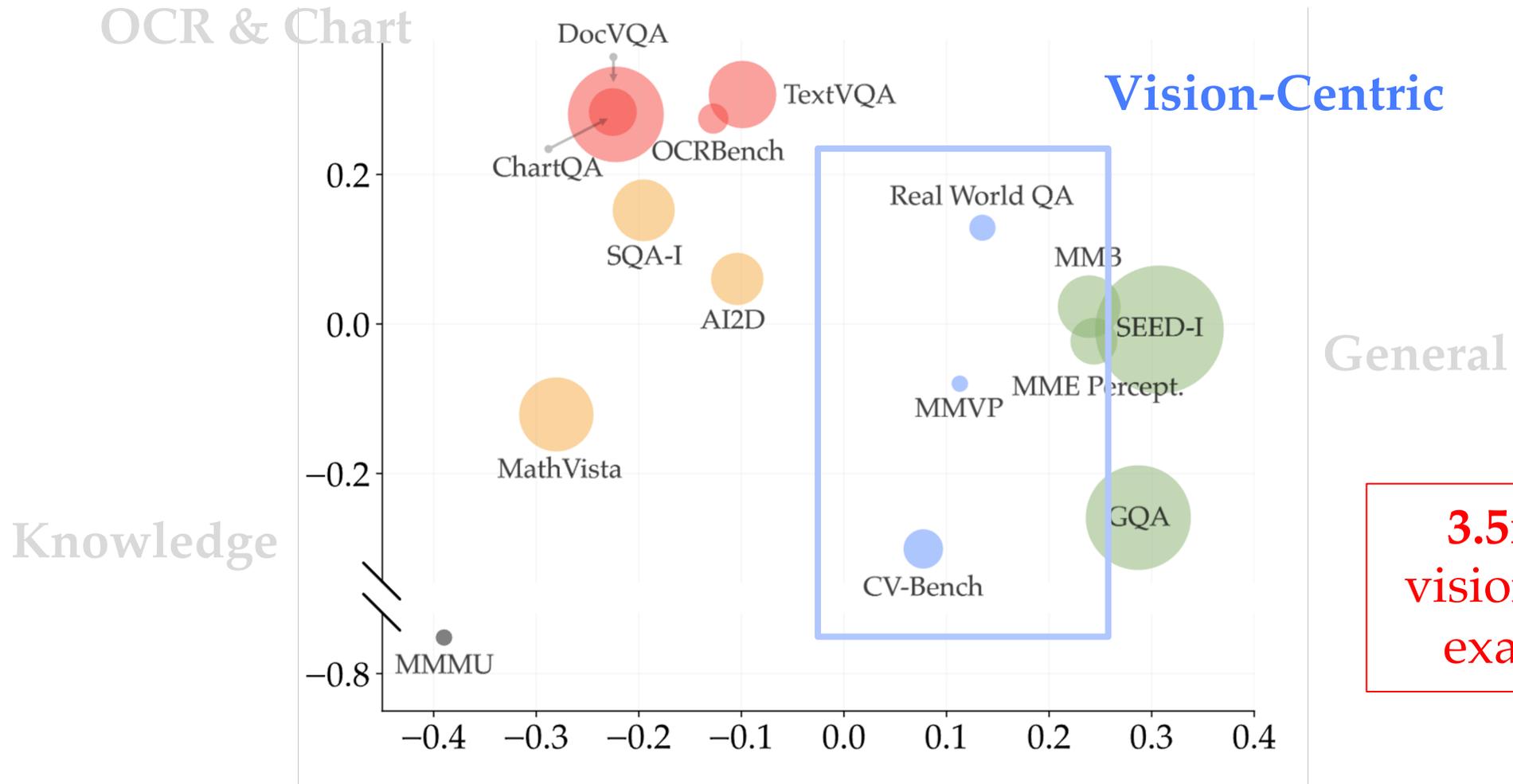
Type	Task	Description	Sources	# Samples
2D	Spatial Relationship	Determine the relative position of an object w.r.t. the anchor object. Consider left-right or top-bottom relationship.	ADE20K COCO	650
	Object Count	Determine the number of instances present in the image.	ADE20K COCO	788
3D	Depth Order	Determine 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

2,638 manually-inspected examples

Group Benchmarks



Group Benchmarks



Overview



Evaluation Protocol



Instruction Tuning
Recipe



Visual
Representations

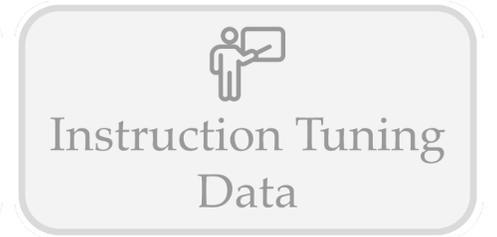
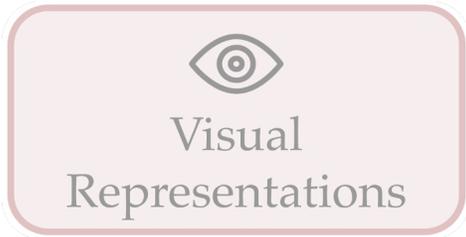
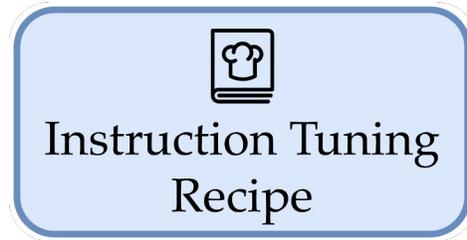


Connector Design

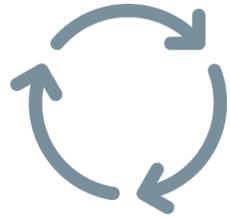


Instruction Tuning
Data

Overview



Instruction Tuning Recipes



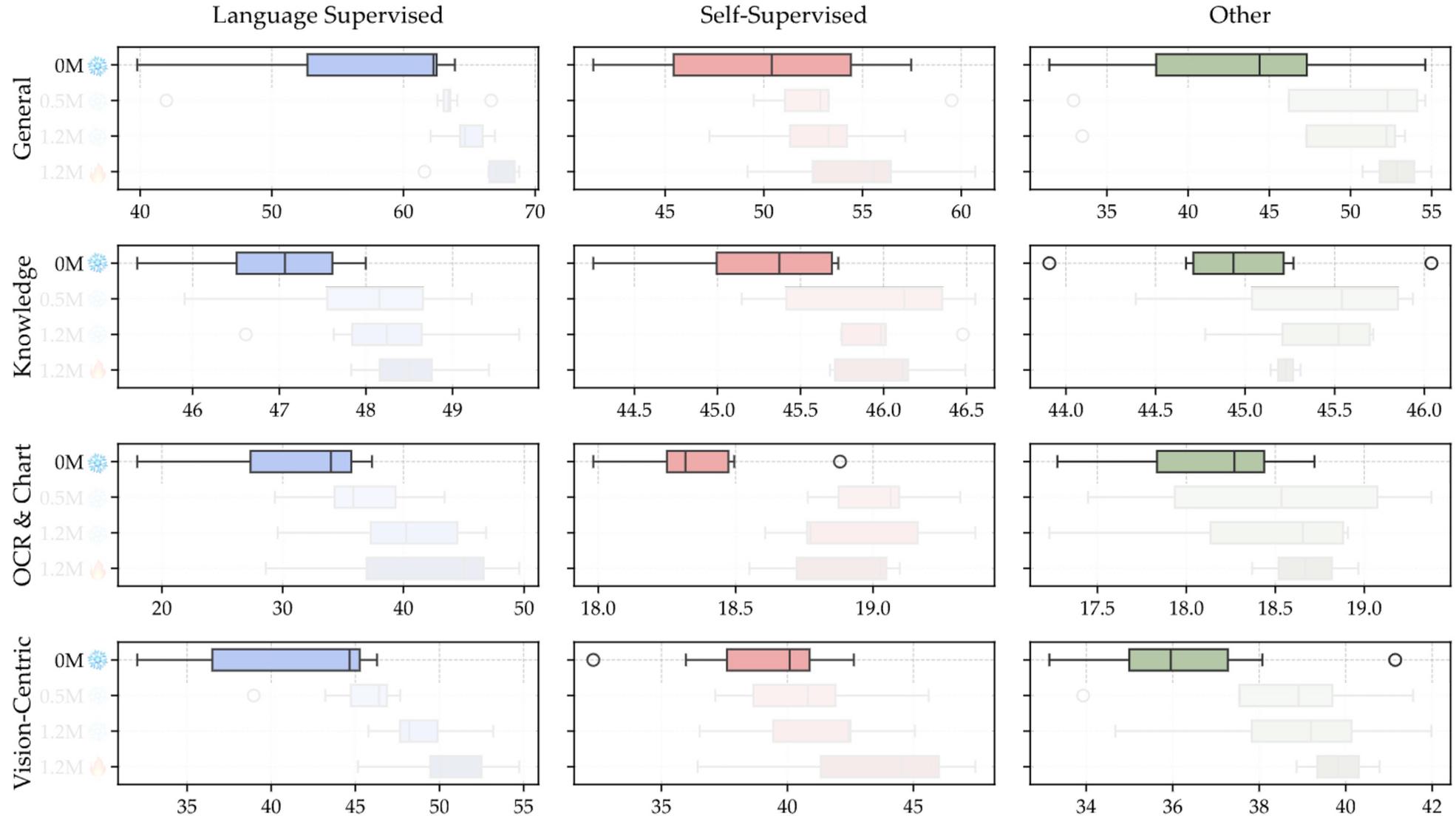
1-Stage or 2-Stage Tuning



Freeze / Unfreeze Vision
Backbone?

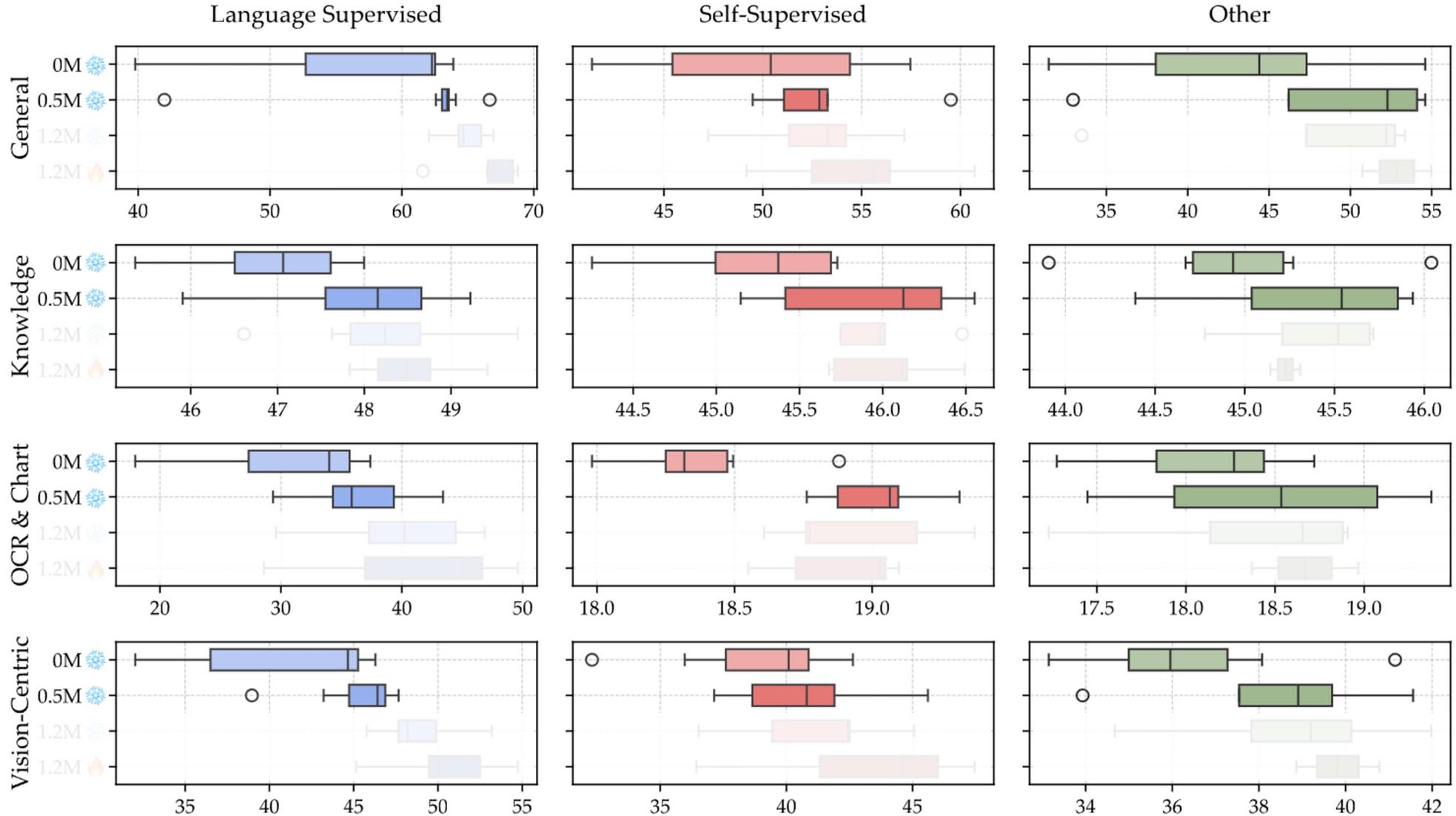
Instruction Tuning Recipes

More
alignment
data

Instruction Tuning Recipes

More
alignment
data

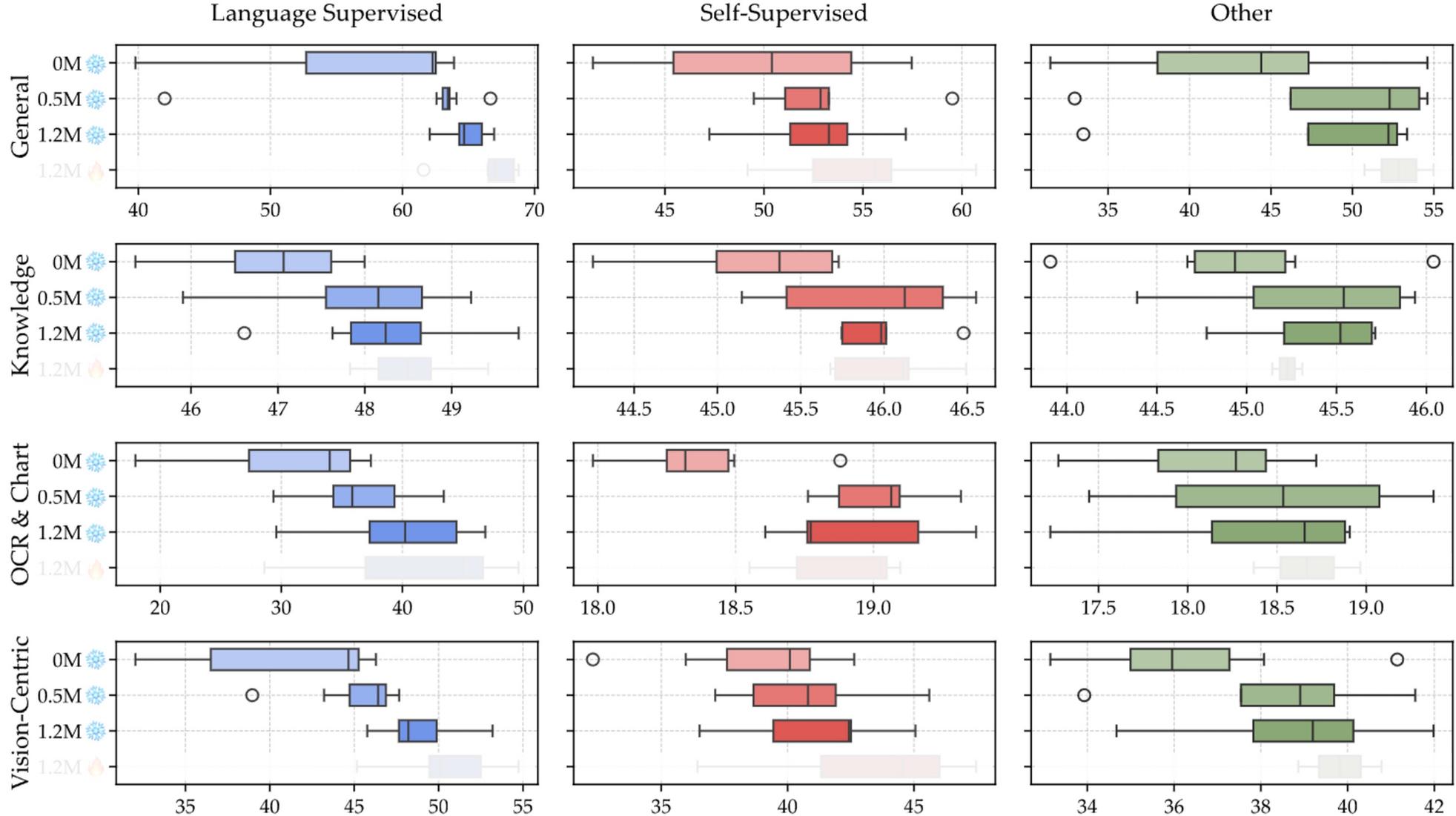





2-stage with more alignment data is best!

Instruction Tuning Recipes

More alignment data





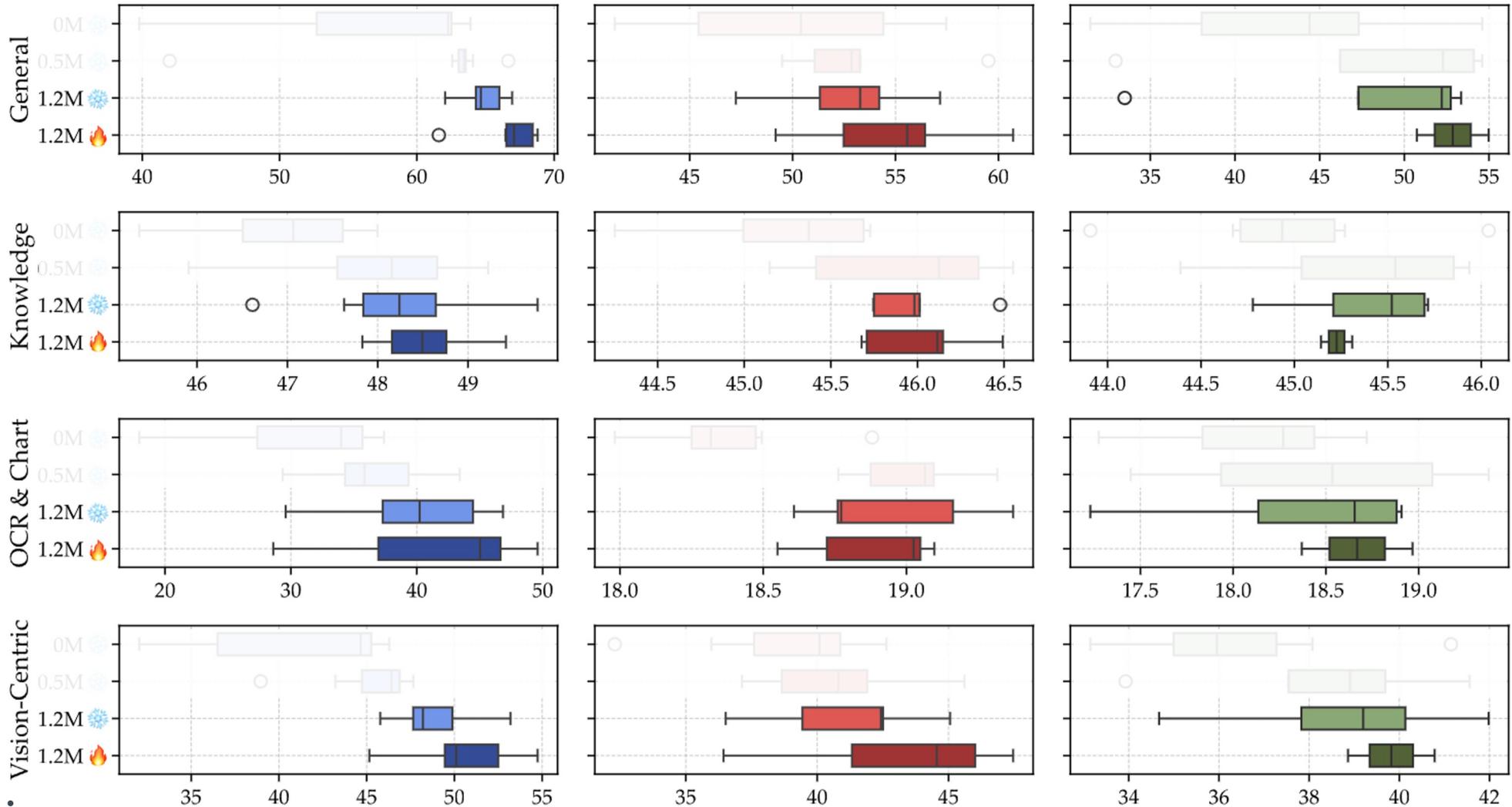
Unfreezing vision encoder helps* 🔥

Instruction Tuning Recipes

Language Supervised

Self-Supervised

Other



Unfreeze

* Can be slow + expensive

Overview



Evaluation Protocol



Instruction Tuning
Recipe



Visual
Representations



Connector Design



Instruction Tuning
Data

Overview



Evaluation Protocol



Instruction Tuning
Recipe



Visual
Representations

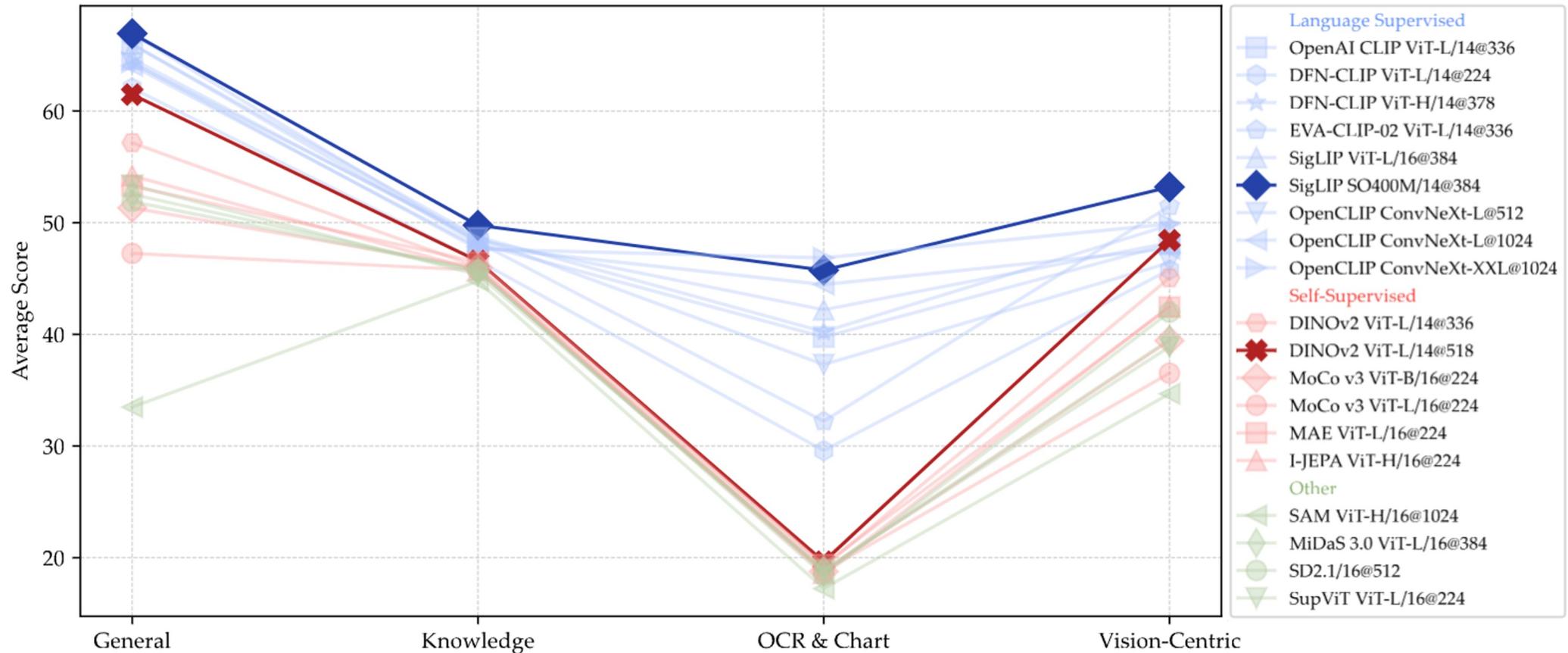


Connector Design



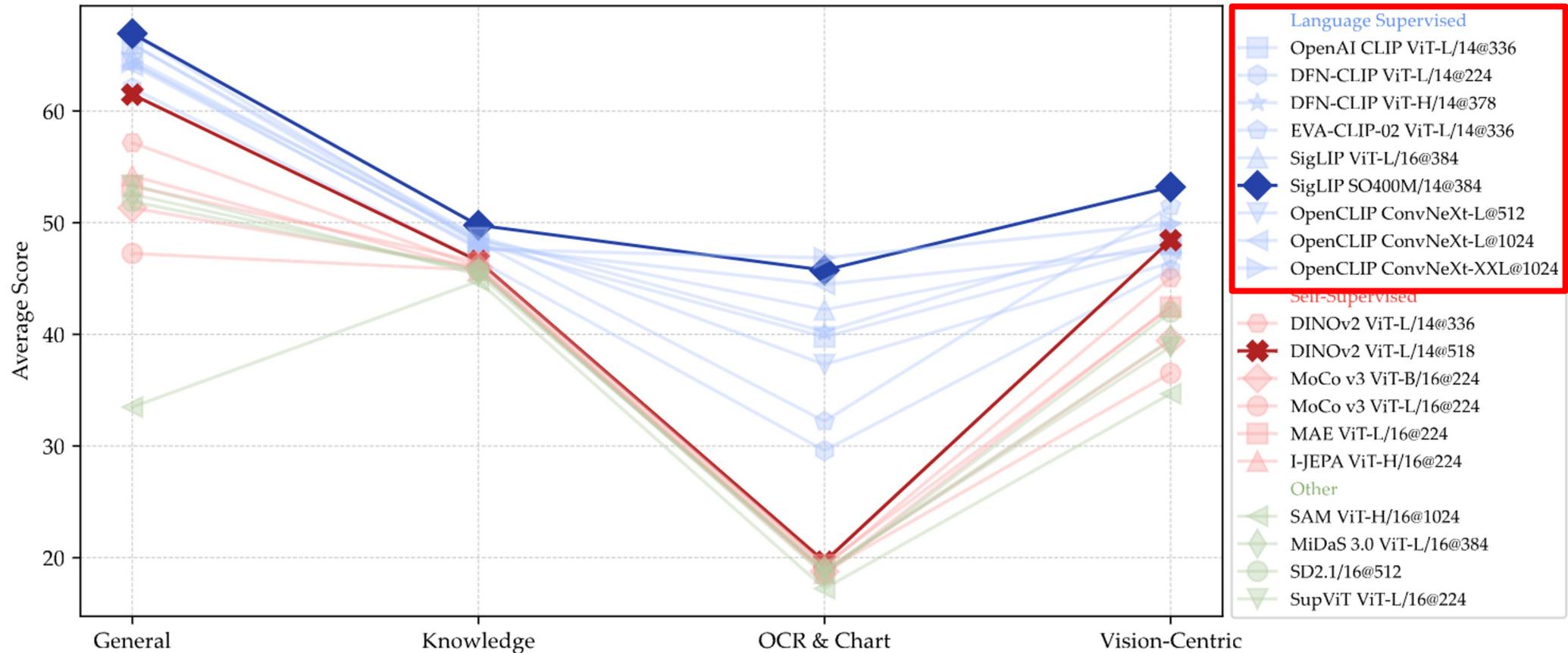
Instruction Tuning
Data

Evaluating Visual Representations

 **1.2M**


Evaluating Visual Representations

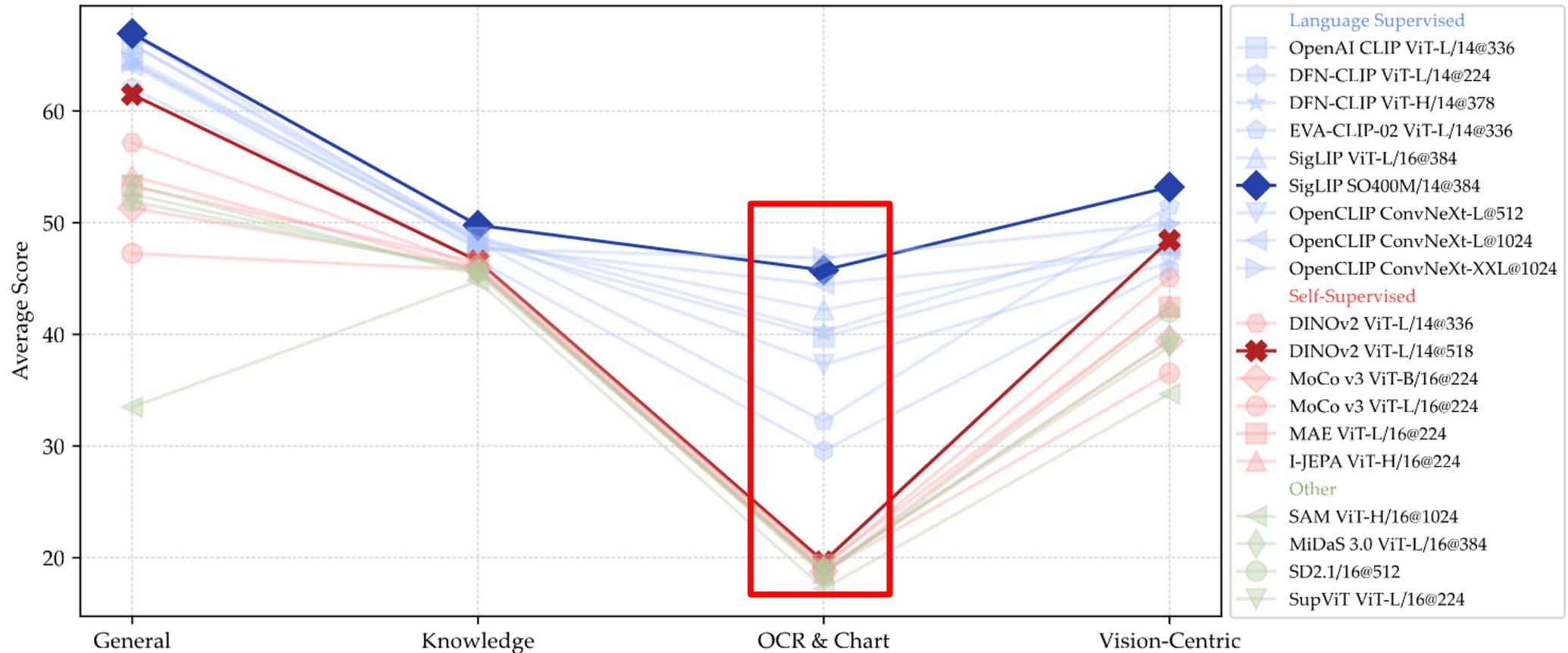
#1 Language Supervised Models are better



Evaluating Visual Representations

#1 Language Supervised Models are better

#2 Gap is largest in OCR & Chart

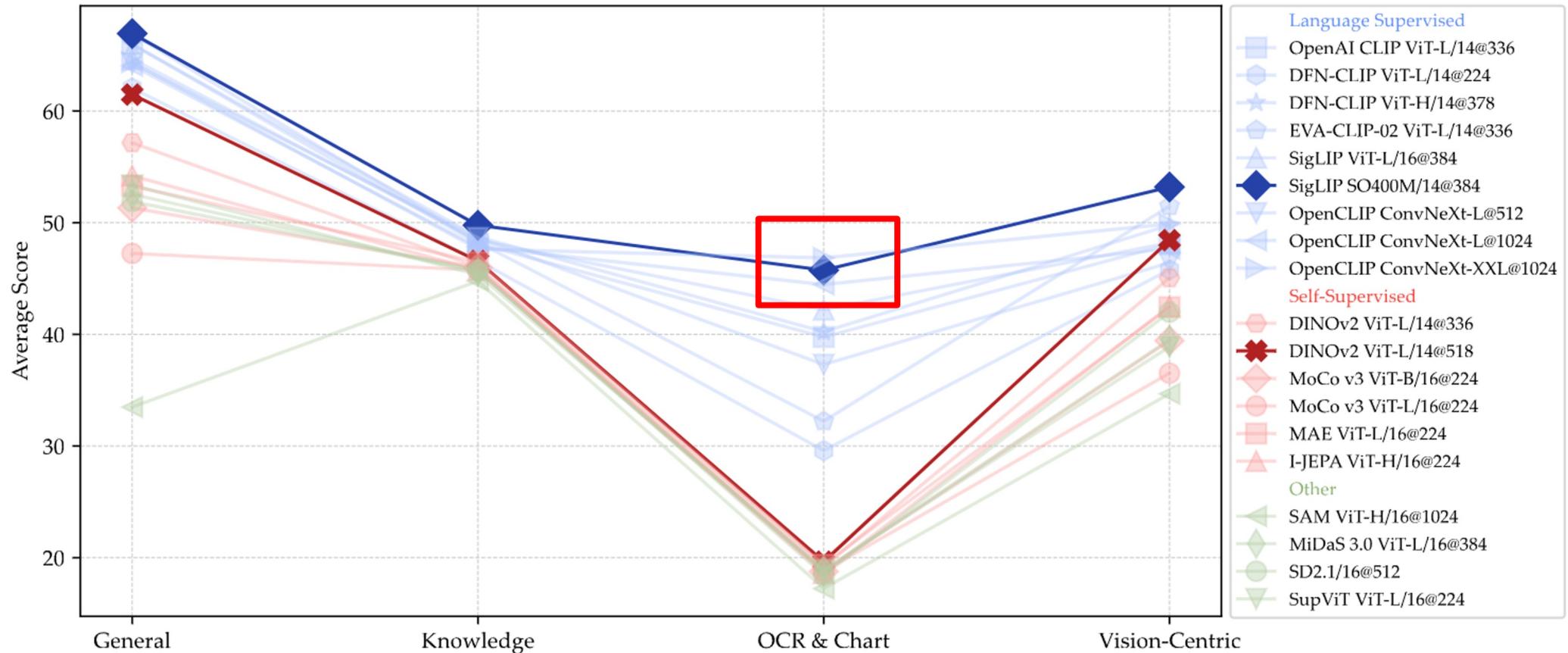


Evaluating Visual Representations

#1 Language Supervised Models are better

#2 Gap is largest in OCR & Chart

#3 ConvNets are good at OCR



Evaluating Visual Representations

#1 Language Supervised Models are better

#2 Gap is largest in OCR & Chart

#3 ConvNets are good at OCR

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

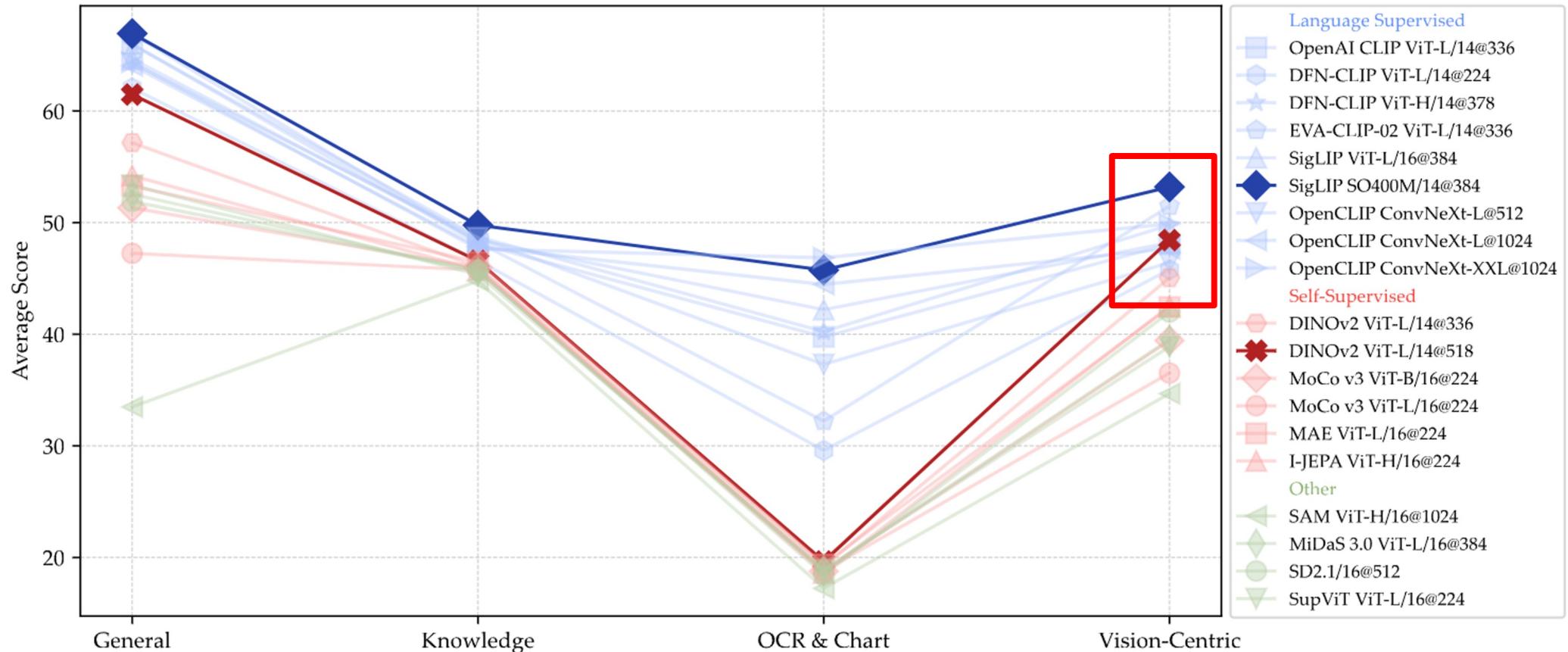
Evaluating Visual Representations

#1 Language Supervised Models are better

#3 ConvNets are good at OCR

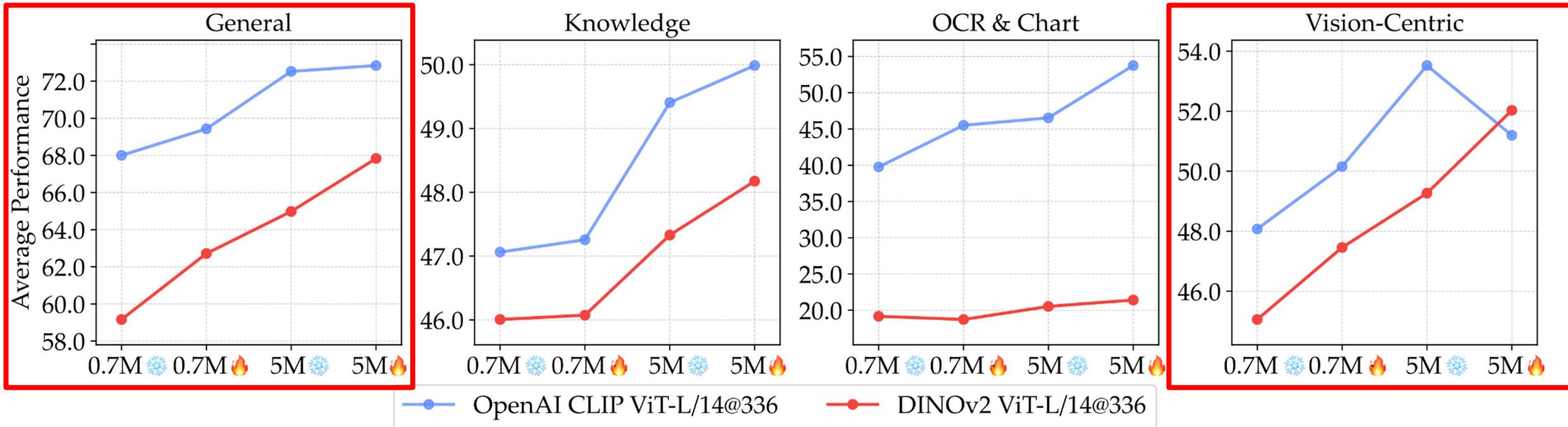
#2 Gap is largest in OCR & Chart

#4 Best SSL Model good at vision-centric



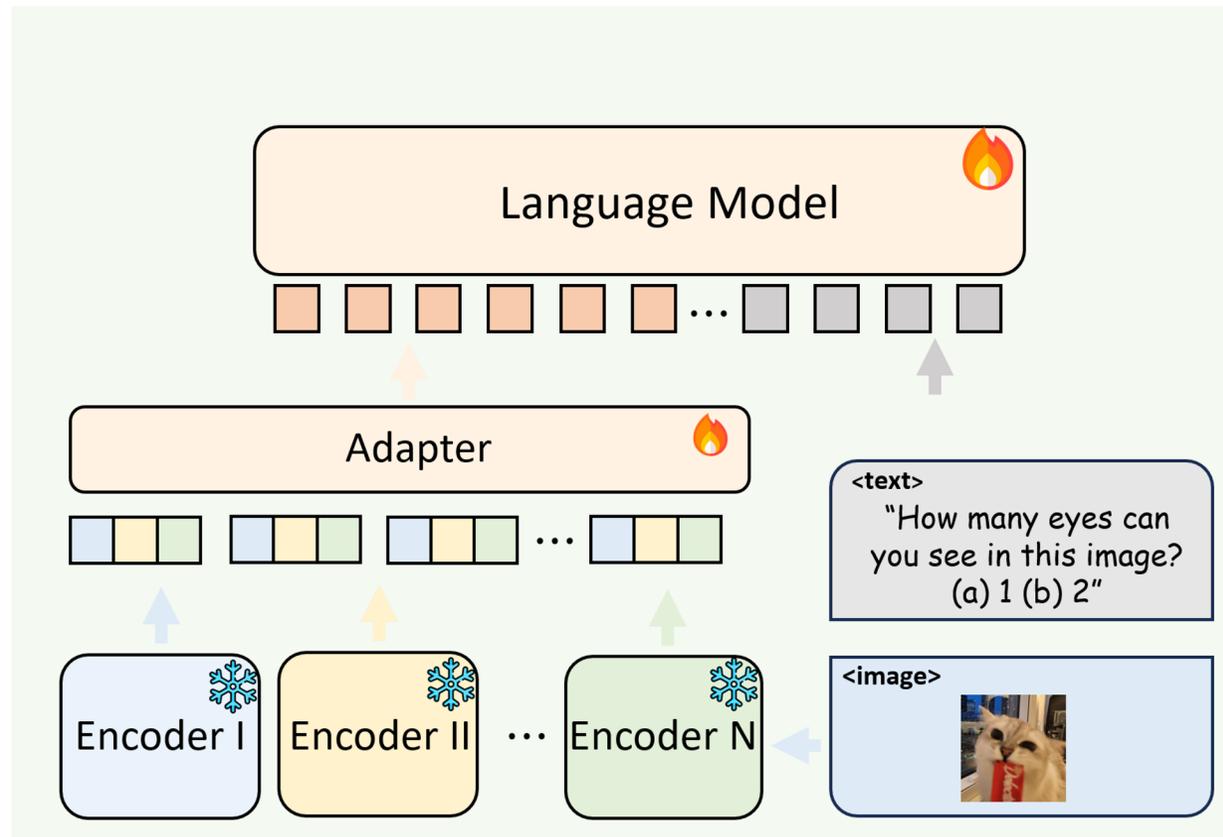
Evaluating Visual Representations

Further training *narrows the gap* between Language-Supervised and Self-Supervised models



Combine Multiple Visual Representations?

→ Simple concatenation of multiple vision embeddings improves performance



Combine Multiple Visual Representations?

→ Simple concatenation of multiple vision embeddings improves performance

Method	Vision Backbone	Average	General				Knowledge				OCR & Chart				Vision-Centric			
			MME ^P	MMB	SEED ^I	GQA	SQA ^I	MMMU ^V	MathVista ^M	AI2D	ChartQA	OCRBench	TextVQA	DocVQA	MMVP	RealWorldQA	CV-Bench ^{2D}	CV-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

But: has bottlenecks...

1. It interpolates all features to the same size
2. It pays equal attention to all vision models

Overview



Evaluation Protocol



Instruction Tuning
Recipe



Visual
Representations



Connector Design



Instruction Tuning
Data

Overview



Evaluation Protocol



Instruction Tuning
Recipe



Visual
Representations

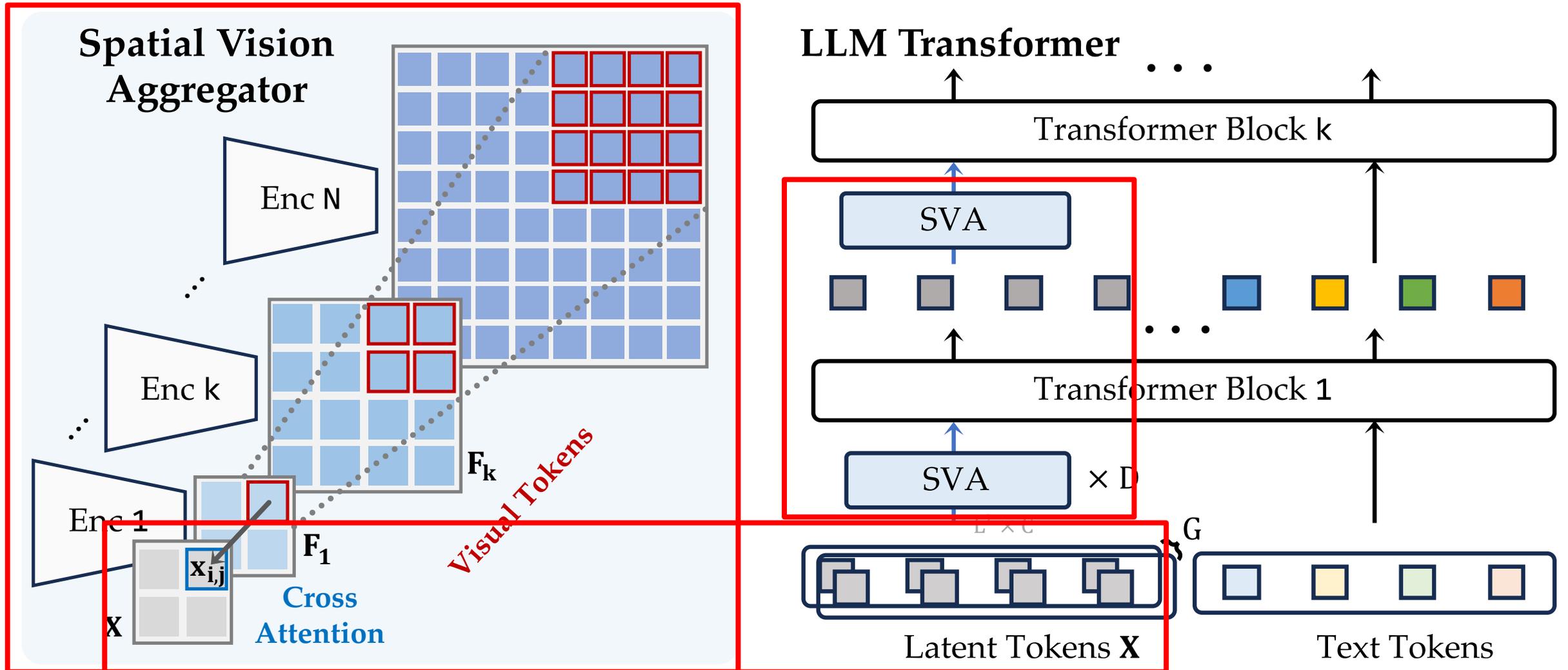


Connector Design



Instruction Tuning
Data

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. [127]	67.2	48.9	50.1	52.6
Resampler [59]	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. [127]	67.2	48.9	50.1	52.6
Resampler [59]	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

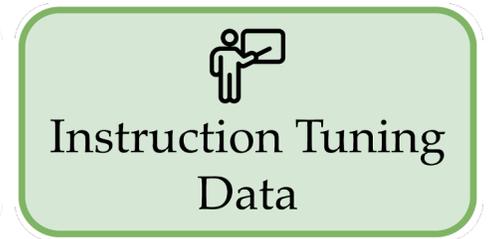
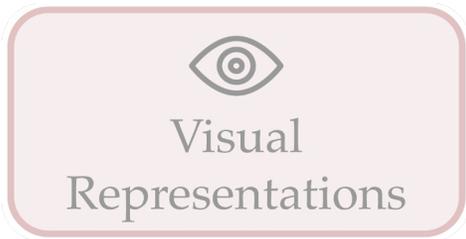


Connector Design



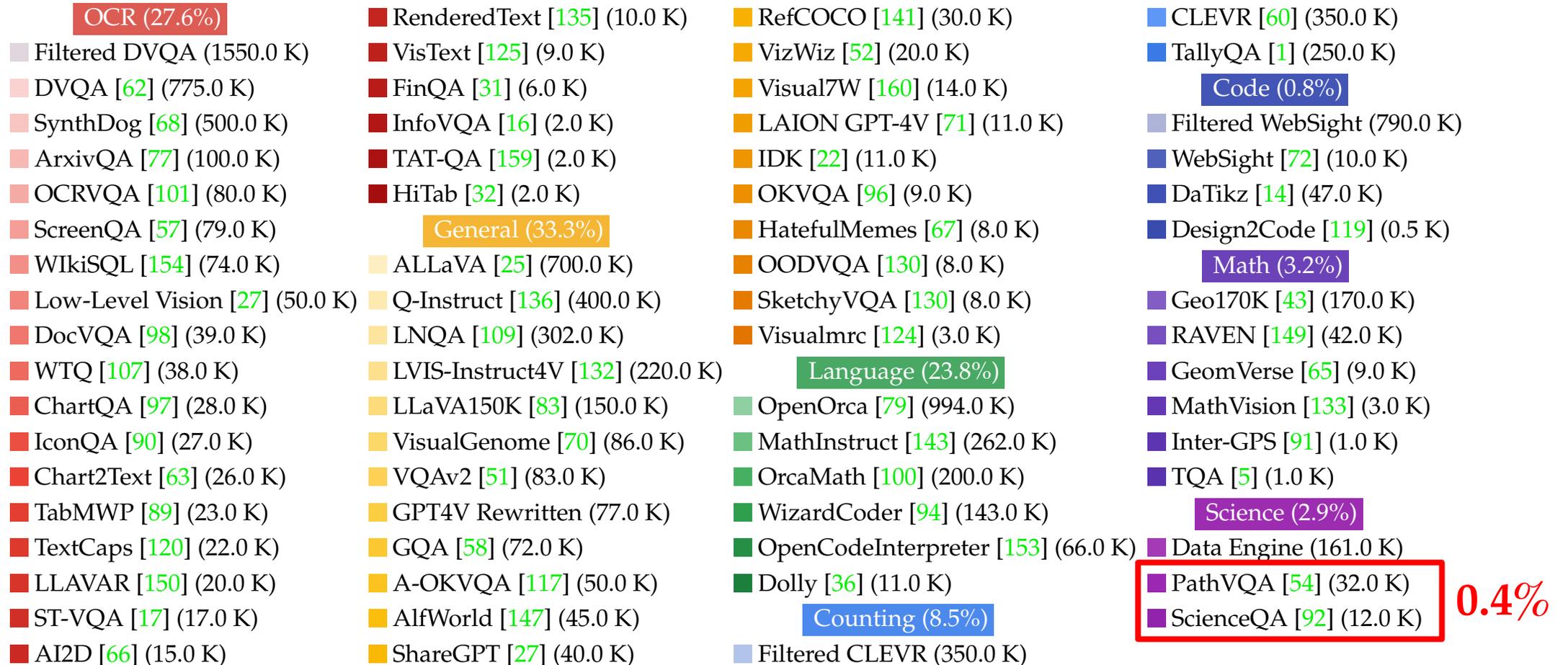
Instruction Tuning
Data

Overview



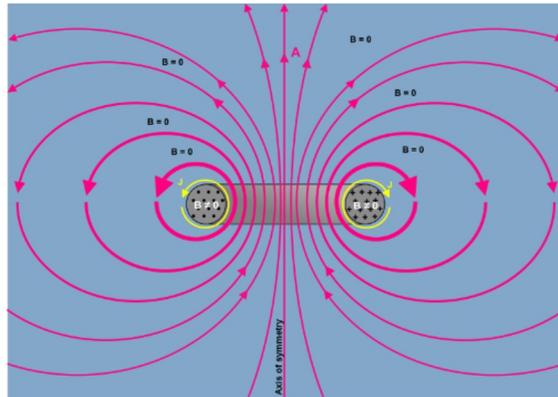
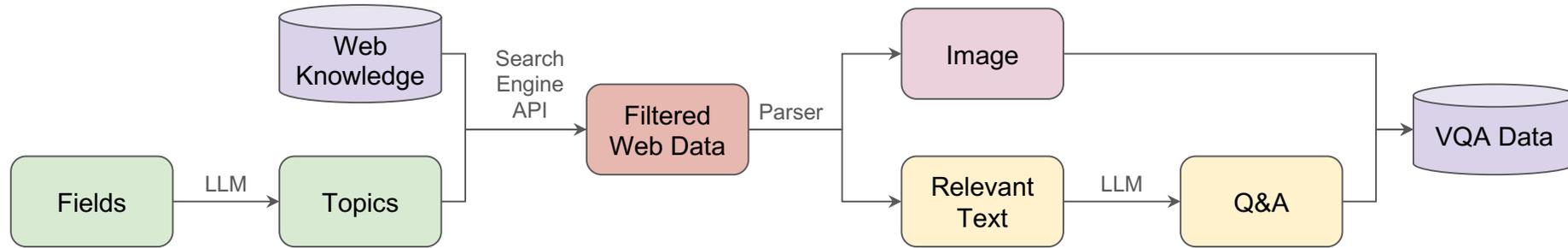
Instruction Tuning Data: Collection

Collect *all available* Instruction Tuning Data



Instruction Tuning Data: Collection

Targeted Data Engine

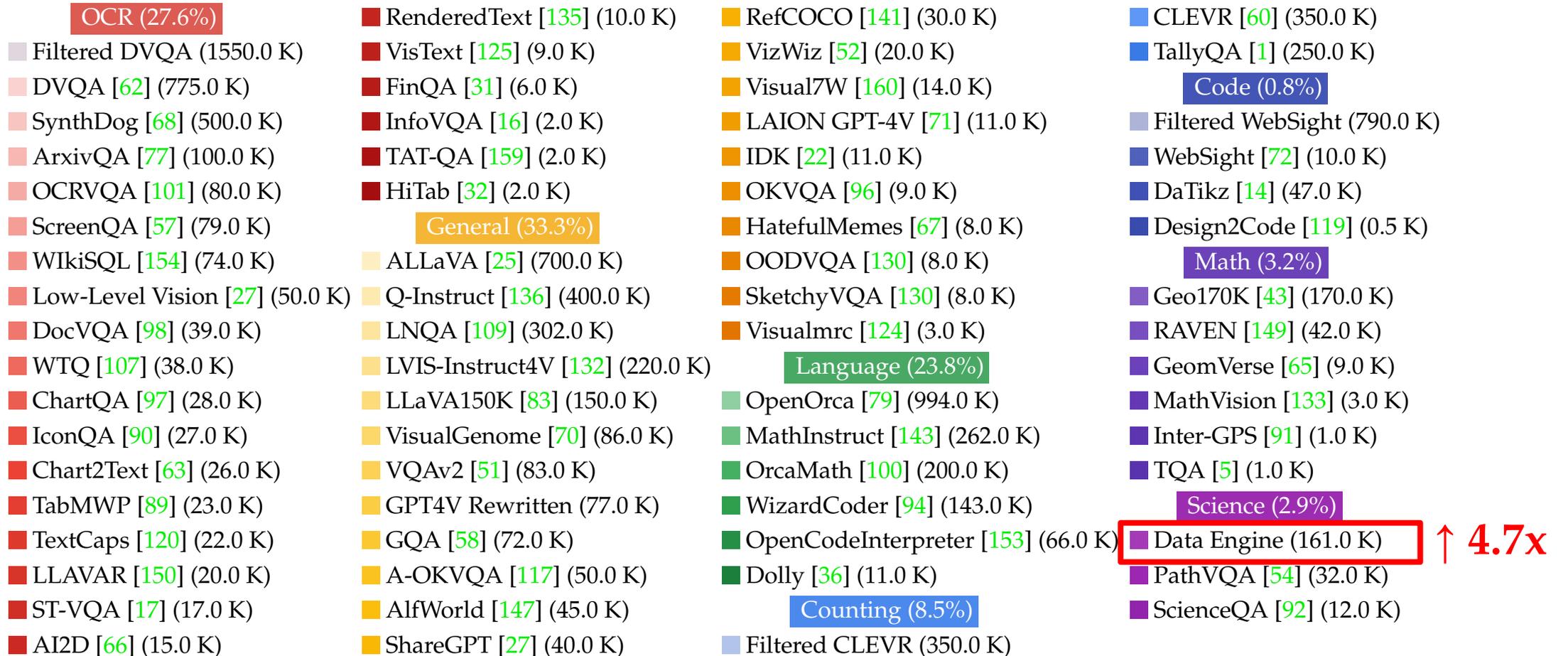


Question: What do the thicker lines signify?

Answer: The thicker lines in the image represent paths of higher average intensity in the magnetic vector potential A field. Shorter paths have higher intensity, ensuring that the path integral remains constant. These lines are drawn to visually convey the general behavior and intensity distribution of the A -field around the toroidal inductor.

Instruction Tuning Data: Collection

Collect *all available* Instruction Tuning Data



Instruction Tuning Data: **Curation**



Data Balancing
(Filtering)

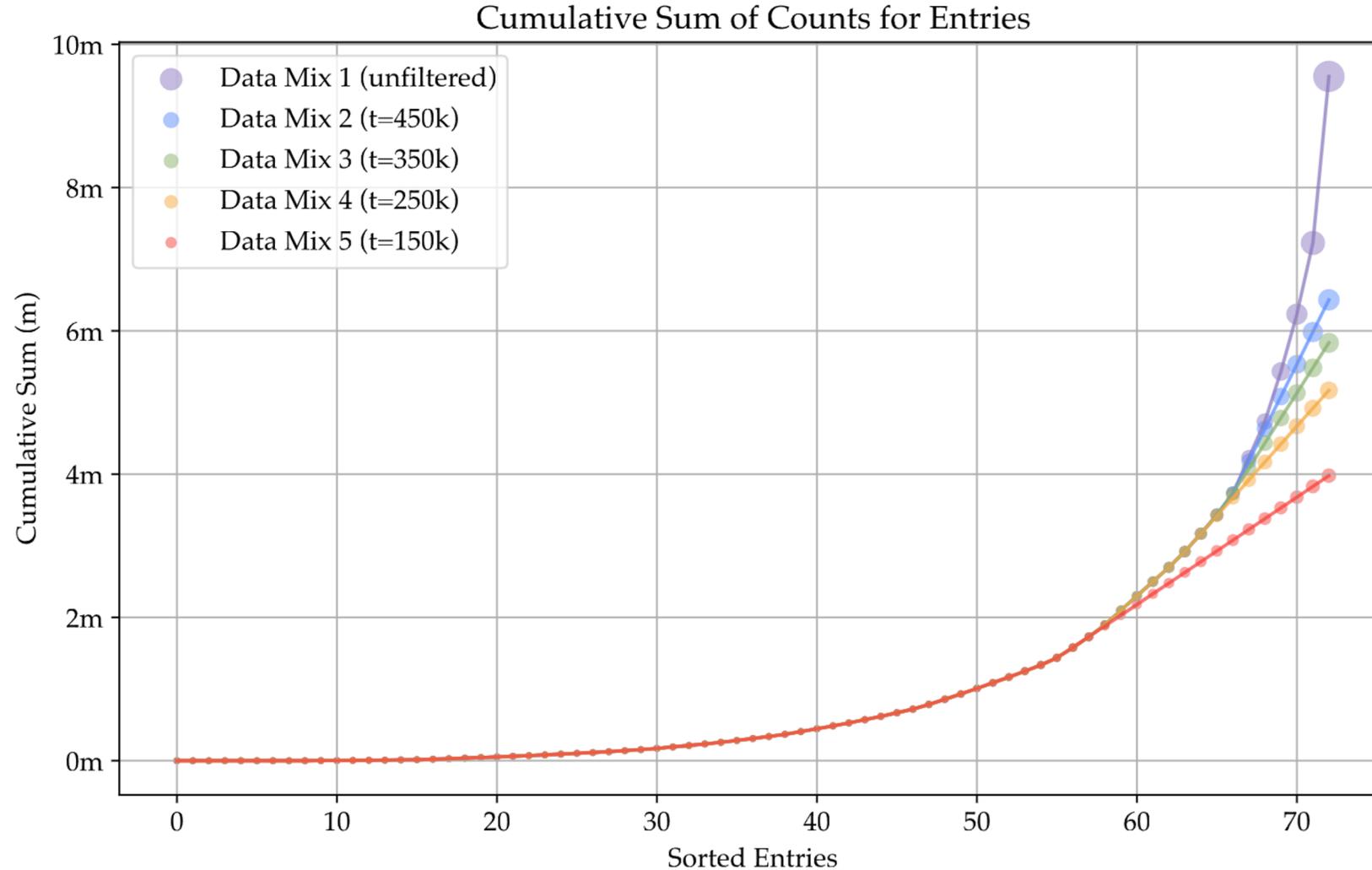


Data Ratio



Data Preprocessing

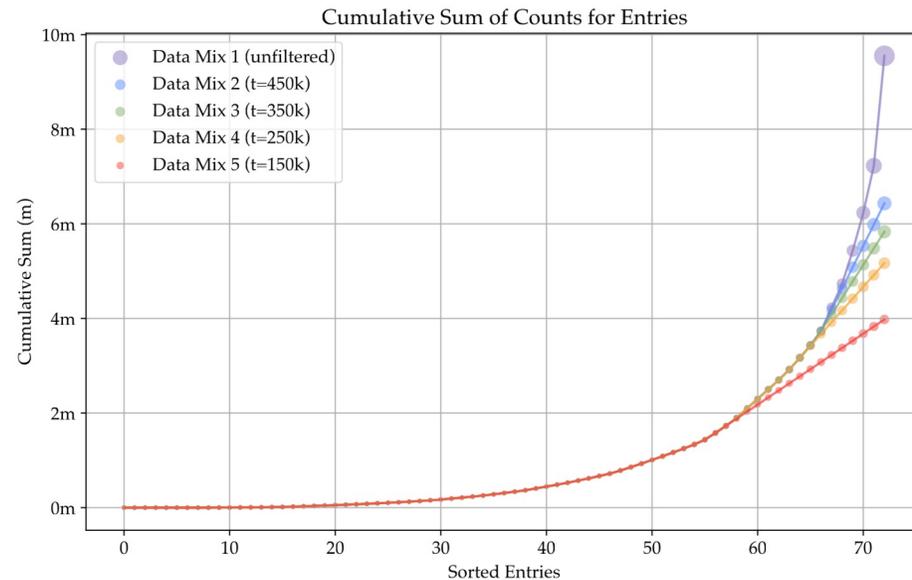
Curation: Data Balancing (Filtering)



Curation: Data Balancing (Filtering)

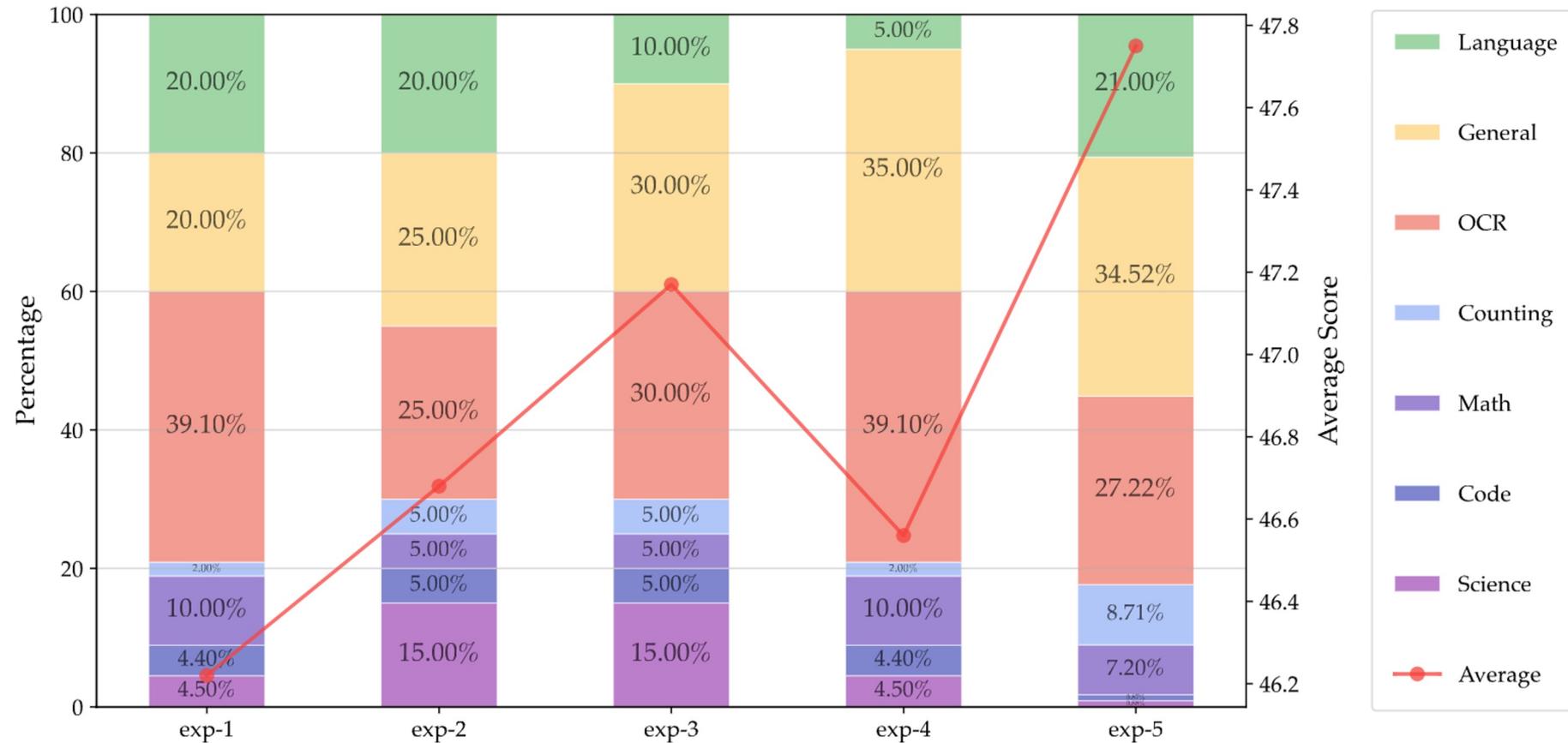
Elbow Effect in Data Balancing

	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



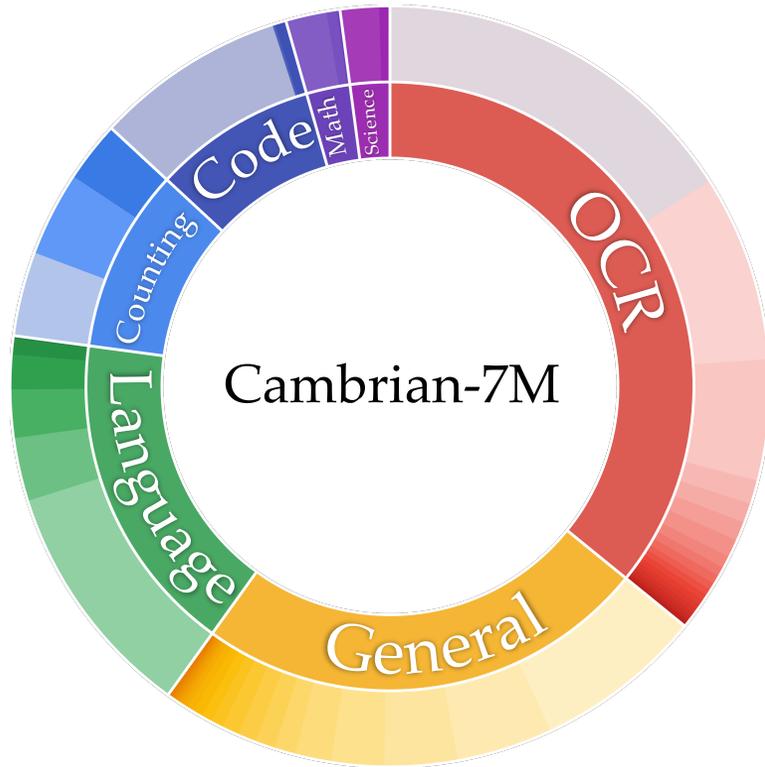
Curation: Data Ratio

Different Categories in Data Matters



Cambrian 7M: High-quality subset

Filtering + Category Balancing



👉 [nyu-visionx/Cambrian-10M](https://nyu-visionx.github.io/Cambrian-10M)

	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

Curation: Text Preprocessing

“Answer Machine Phenomenon”

The model tends to output short answer-like responses...

How many people are in this image?



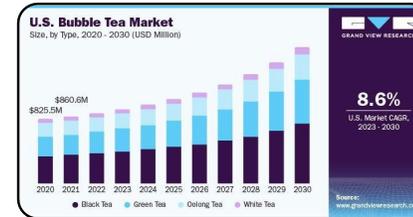
There are six people in the image.

Describe this image please.



Cat in bathtub

Based on this chart, what is the most popular tea?



Black Tea

What city is this? If I want to tour around the landmark in this picture, where should I go?



New York City

Curation: Text Preprocessing

Solution: Add system prompts during instruction finetuning

Index	Response formatting prompts
1	Answer the question using a single word or phrase.
2	Answer the question using a single number or phrase.
3	Answer with the option's letter from the given choices directly.
4	Give the short answer directly.
5	Answer the question using a single word or phrase.
6	When the provided information is insufficient, respond with <no answer>.
7	Directly provide the HTML code.
8	First show your reasoning process and then give the final answer.
9	When the provided information is insufficient, respond with 'Unanswerable'. Answer the question using a single word or phrase.
10	Answer with the letter.

Curation: Text Preprocessing

System Prompt alleviates “Answer Machine Phenomenon”

w/o system prompt

How many people are in this image?



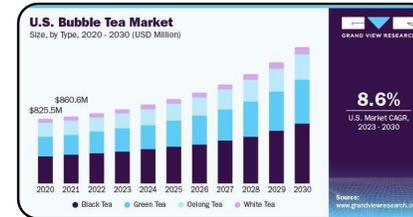
There are six people in the image.

Describe this image please.



Cat in bathtub

Based on this chart, what is the most popular tea?



Black Tea

What city is this? If I want to tour around the landmark in this picture, where should I go?



New York City

Curation: Text Preprocessing

System Prompt alleviates “Answer Machine Phenomenon”

w/o system prompt

How many people are in this image?



There are six people in the image.

There are six people in the image.

Describe this image please.

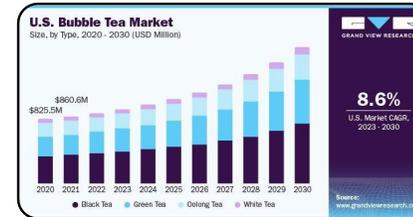


Cat in bathtub

The image captures a serene moment in a bathroom setting. A fluffy, gray cat is comfortably seated on the edge of a white bathtub. The cat's gaze is directed upwards, possibly at something outside the frame, giving it a contemplative expression. The backdrop is a simple white tiled wall, which contrasts with the cat's fur, making it the focal point of the image. The overall scene suggests a quiet, peaceful moment in a domestic setting.

w/ system prompt

Based on this chart, what is the most popular tea?



Black Tea

Based on the chart, black tea appears to be the most popular type of tea in the U.S. Bubble Tea Market, as it consistently has the highest sales volume across all years displayed.

What city is this? If I want to tour around the landmark in this picture, where should I go?



New York City

This is New York City, and the prominent landmark in the image is the Washington Square Arch, which is located in the Greenwich Village neighborhood. To tour around this landmark, you should visit the nearby Washington Square Park, which is a popular spot for locals and tourists alike. The park offers a variety of activities, including walking paths, a fountain, and a statue of George Washington. Additionally, the surrounding area is rich with historical buildings, art galleries, and restaurants, making it a vibrant and culturally diverse place to explore.



Cambrian-1 Models



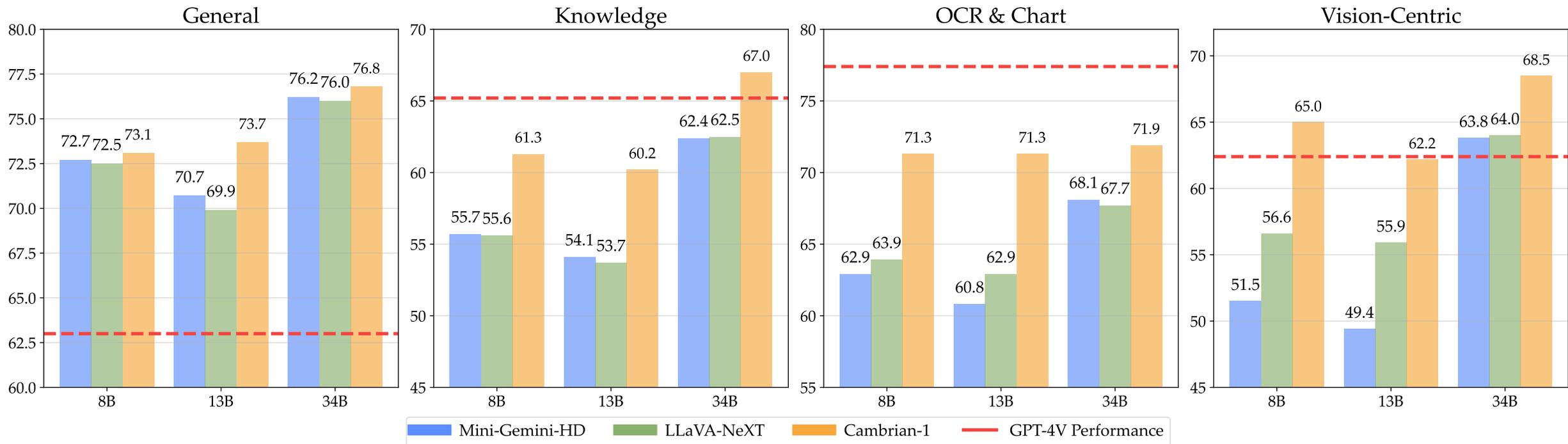
Cambrian-1 Models

cambrian-1-8b [Llama-3]

cambrian-1-13b [Vicuna-1.5]

cambrian-1-34b [Hermes2-Yi]

“SOTA” Performance



Takeaways & What's Next?

- We need to develop better visual representations
 - There is no perfect representation
 - We can use MLLMs as vision model evaluator

Takeaways & What's Next?

- We need to develop better visual representations
 - There is no perfect representation
 - We can use MLLMs as vision model evaluator
- Continuously improve evaluation benchmarks
 - We can breakdown and group benchmarks into categories
 - Model developments go far beyond current benchmarks e.g. chat, creativity, reliability, etc

Takeaways & What's Next?

- We need to develop better visual representations
 - There is no perfect representation
 - We can use MLLMs as vision model evaluator
- Continuously improve evaluation benchmarks
 - We can breakdown and group benchmarks into categories
 - Model developments go far beyond current benchmarks e.g. chat, creativity, reliability, etc
- The tradeoff between # of Tokens and High-Resolution
 - Using Hi-Res encoders (eg, ConvNets) helps to process high-res efficiently and effectively
 - Spatial Inductive Bias in SVA is also important

Takeaways & What's Next?

- We need to develop better visual representations
 - There is no perfect representation
 - We can use MLLMs as vision model evaluator
- Continuously improve evaluation benchmarks
 - We can breakdown and group benchmarks into categories
 - Model developments go far beyond current benchmarks e.g. chat, creativity, reliability, etc
- The tradeoff between # of Tokens and High-Resolution
 - Using Hi-Res encoders (eg, ConvNets) helps to process high-res efficiently and effectively
 - Spatial Inductive Bias in SVA is also important
- **Data! Data! Data!**
 - Data Collection, Balancing, Ratio and Preprocessing

Takeaways & What's Next?

- We need to develop better visual representations
 - There is no perfect representation
 - We can use MLLMs as vision model evaluator
- Continuously improve evaluation benchmarks
 - We can breakdown and group benchmarks into categories
 - Model developments go far beyond current benchmarks e.g. chat, creativity, reliability, etc
- The tradeoff between # of Tokens and High-Resolution
 - Using Hi-Res encoders (eg, ConvNets) helps to process high-res efficiently and effectively
 - Spatial Inductive Bias in SVA is also important
- Data! Data! Data!
 - Data Collection, Balancing, Ratio and Preprocessing
- Training & Post-Training
 - End-to-end training instead of fine-tuning LLM
 - Post-Training: Using RL, potentially bypasses limitations of SFT, e.g. Decision-Making Abilities

The background is a dense, light-colored collage of various Cambrian fossils, including trilobites, graptolites, and other marine organisms, rendered in a semi-transparent, artistic style.

Thank you!

<https://cambrian-mlm.github.io/>

Poster 11am–2pm!
East Exhibit Hall A-C #3700

II: Spatial Intelligence

1. [VSI-Bench*](#): Benchmarking Spatial Intelligence
2. [SIMS-V](#): Spatial Training via Embodied Simulator
3. [Test-Set Stress Test](#): Debiasing Benchmarks

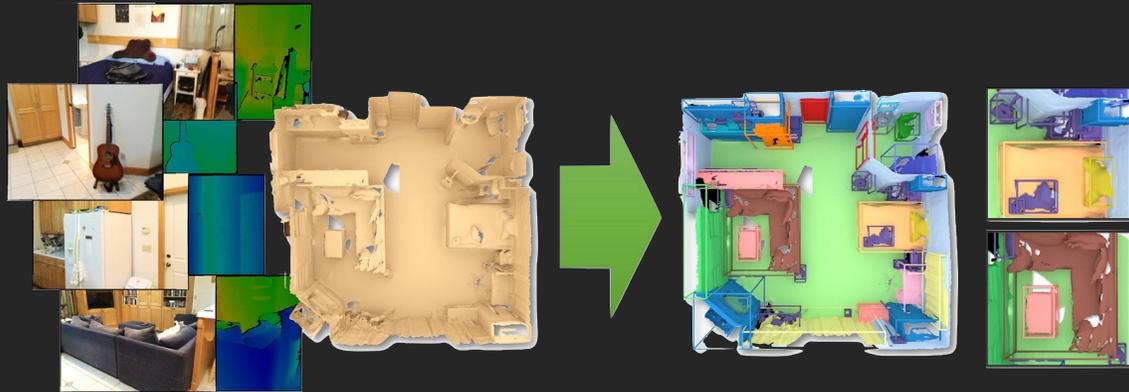
Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces

Jihan Yang*, Shusheng Yang*, Anjali W. Gupta*, Rilyn Han*,
Li Fei-Fei, Saining Xie



In computer vision...

We study *space*, but not *thinking*...



[ScanNet, Dai et al. 2017]

We study *thinking*, but not in *space*...



On what date did the individual in the video leave a place that Simon thought was very important to him?

- A. May 31, 2022. B. June 9, 2021. C. May 9, 2021. D. June 31, 2021.



[Video-MME, Fu et al. 2024]

Watch the video and answer the question



How many chairs are there in this room?

Your Answer: ?

Ground Truth: 9

Gemini-1.5 Pro Answer: 4

Watch the video and answer the question



If I am standing by the nightstand and facing the chair, is the closet to the left or the right of the chair?

A. Left B. Right

Your Answer: ?

Ground Truth: Left

Gemini-1.5 Pro Answer: Right



How do humans do this?
Can models do this? How?



VSI-Bench

Benchmark Formulation

Video



Question

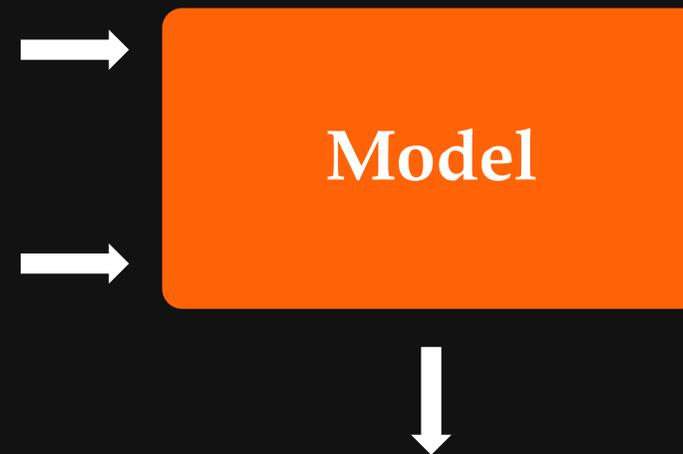
If I am standing by the nightstand and facing the chair, is the closet to the left or the right of the chair?

GT

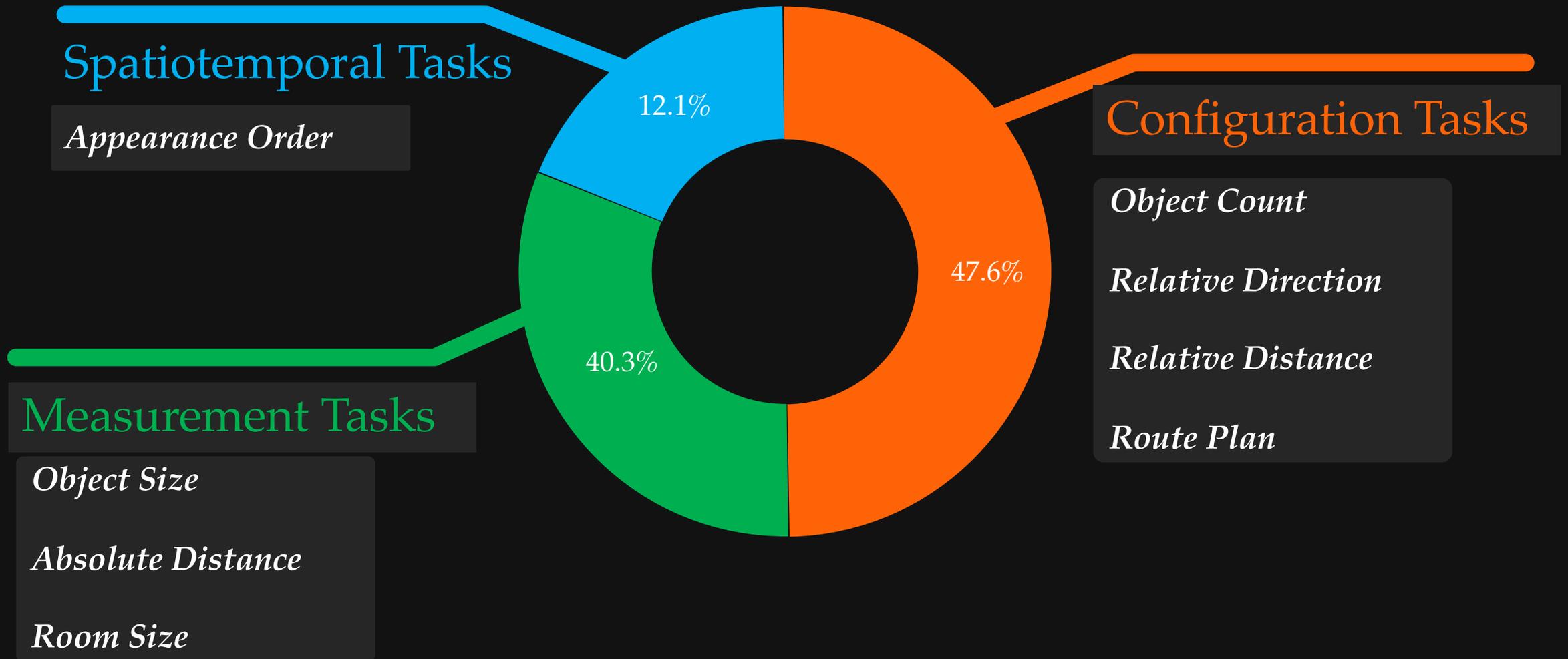
Left



Prediction



Task Definition



Spatiotemporal Tasks

Appearance Order

Configuration Tasks

Object Count

Relative Direction

Relative Distance

Route Plan

Measurement Tasks

Object Size

Absolute Distance

Room Size

How can we construct the benchmark?

Real-world Video



Ground Truth

Object Counts
Room Size
Direction
Distance
...

Build From Scratch



Repurposing Existing 3D Dataset!

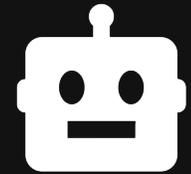
ScanNet
ScanNet++
ARKitScenes



Object Category
3D Boxes
Segmentation Map
...



Object Counts
Object Size
Room Size
Distance
Direction
Appearance
...
Meta Information



Automatic QA
Generation

Human In the Loop Verifying and Filtering

5K+ High Quality QA Pairs
Affordable Human Efforts



Benchmarking MLLMs on VSI-Bench

Chance Level



Gemini 1.5 Pro

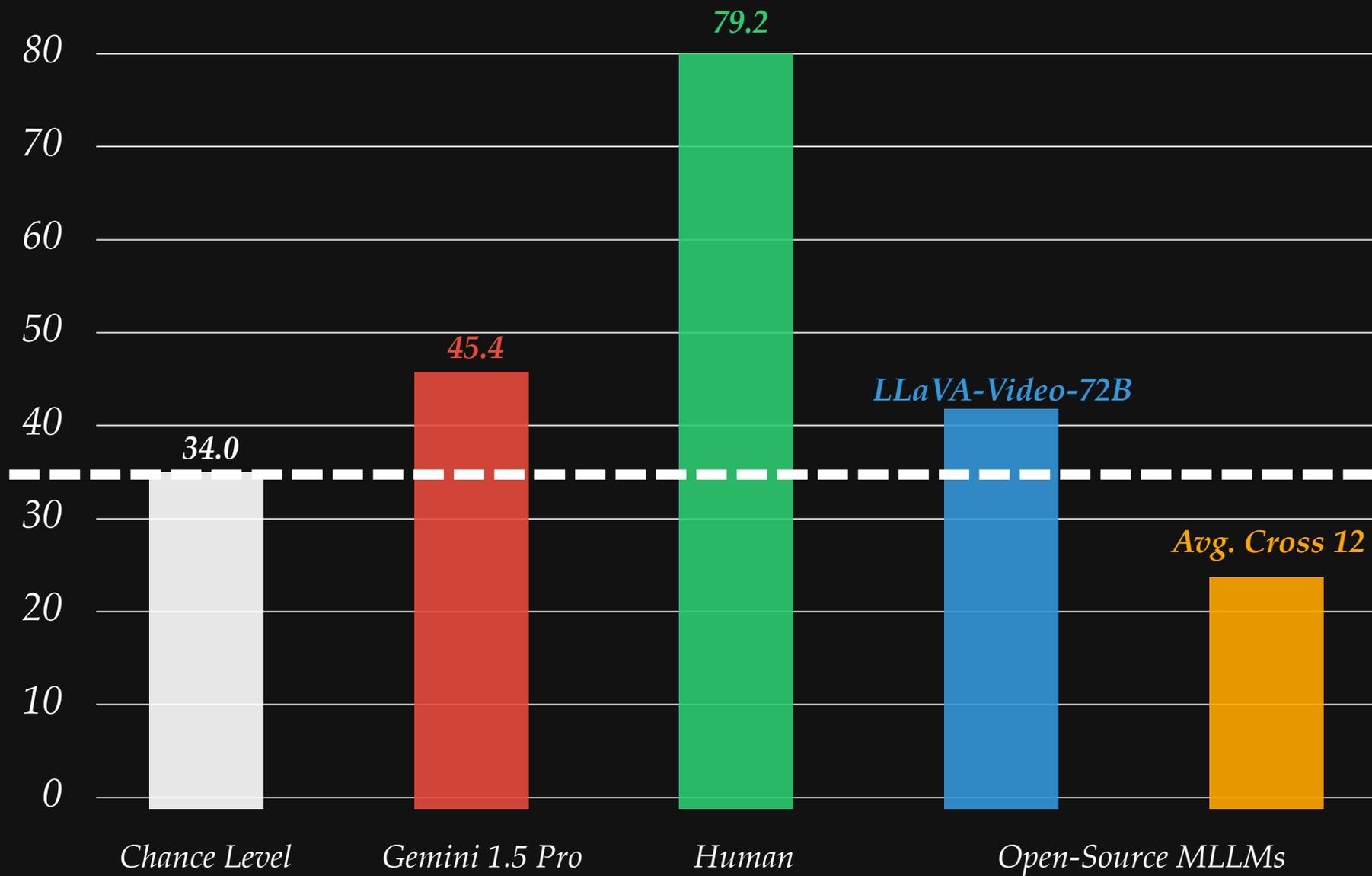


Human

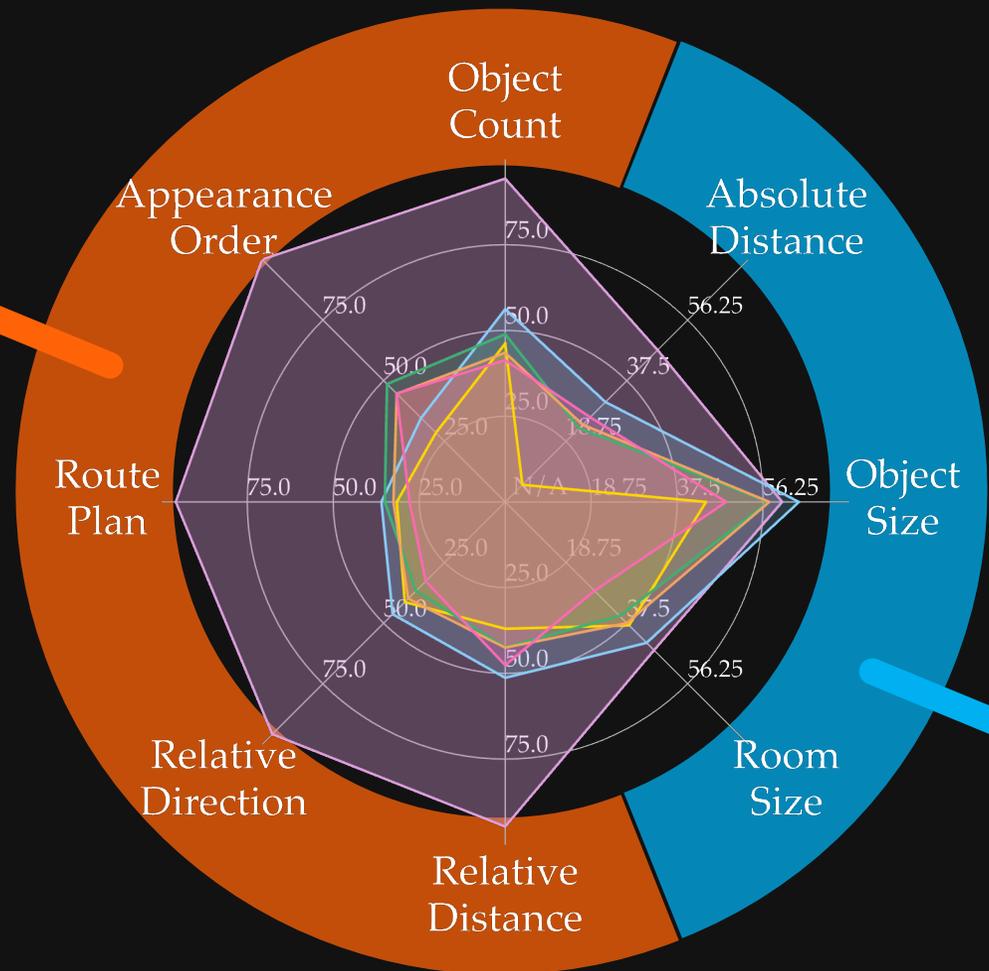


Open-Source MLLMs





Configuration and Spatial Temporal Tasks

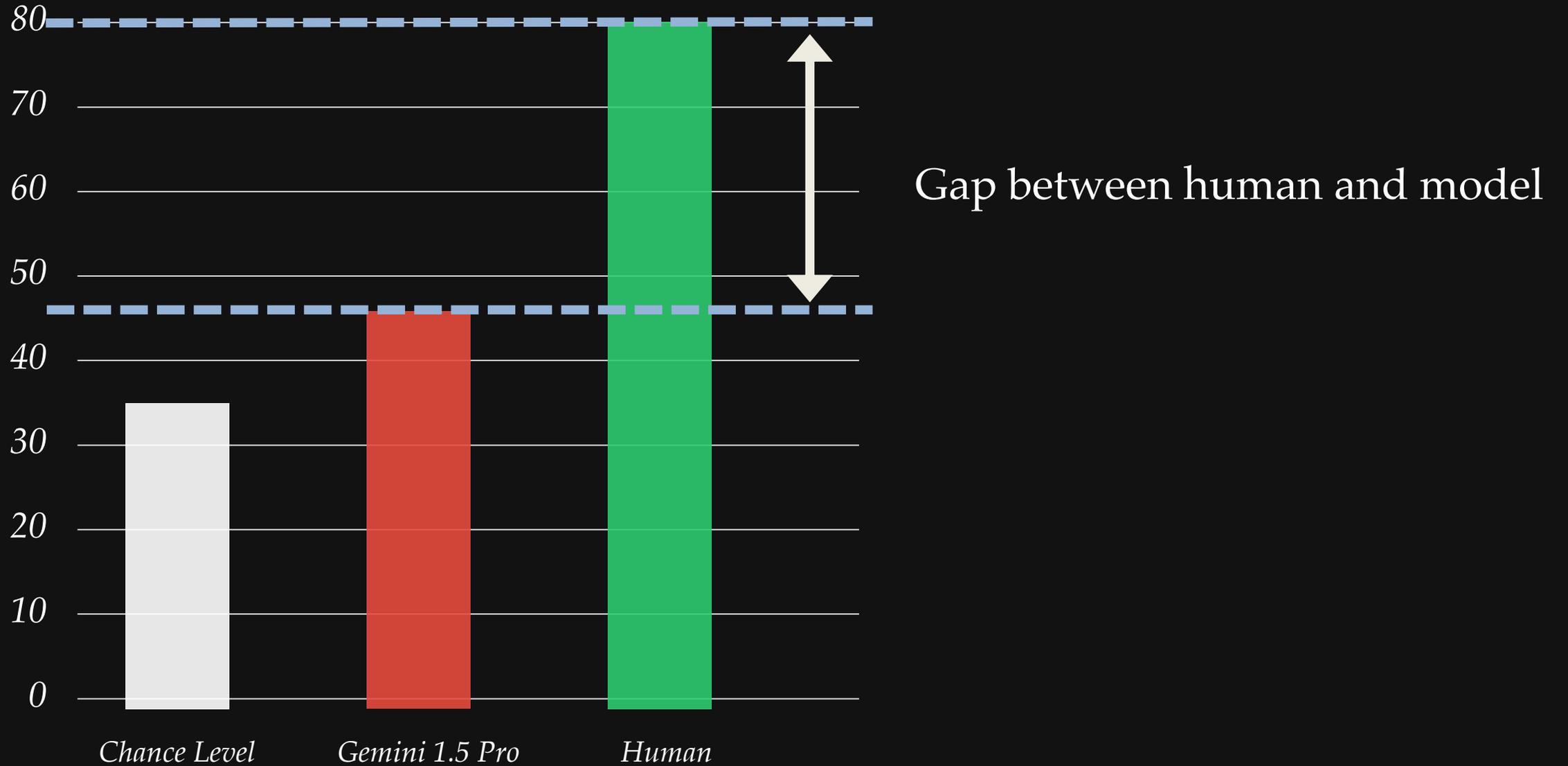


Measurement Tasks

- Human-Level
- Gemini-1.5 Pro
- GPT-4o
- LLaVA-Video-72B
- LLaVA-OneVision-72B
- InternVL2-40B

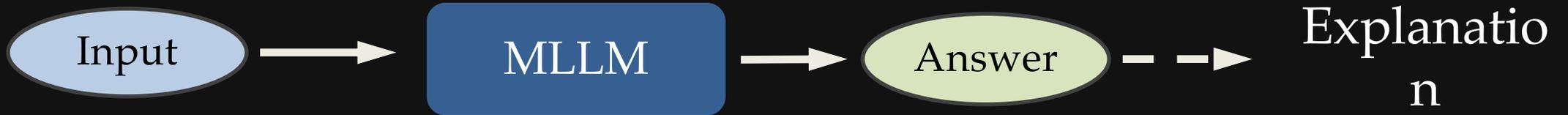
How Do MLLMs Think in
Space?

How Do MLLMs Think in Space?



How Do MLLMs Think in Space?

Prompt Model to Explain Itself

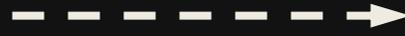


Analysis by Self-Explanation



Error Breakdown

Visual Perception



Recognition error

Linguistic Intelligence



Logic/Math reasoning error

Relational Reasoning



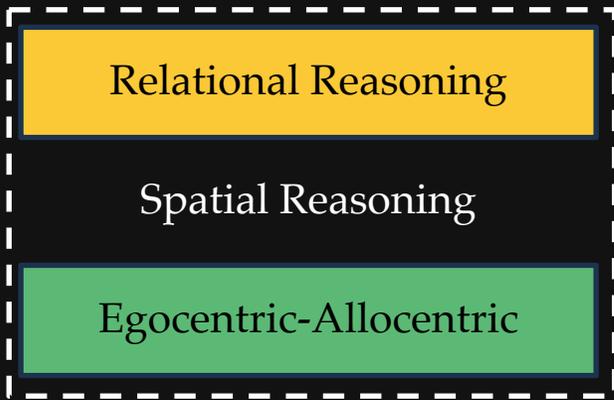
Distance/Size/Direction reasoning error

Spatial Reasoning

Egocentric-Allocentric



Perspective-taking and spatial layouts error

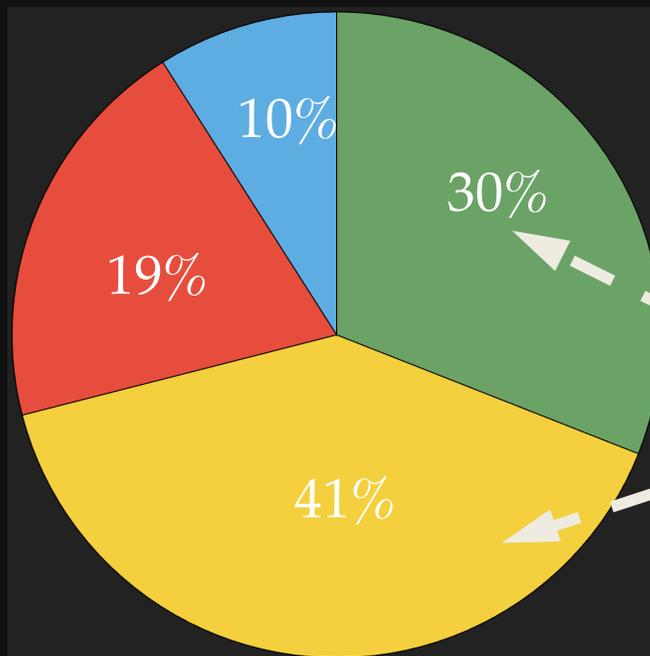


Analysis by Self-Explanation

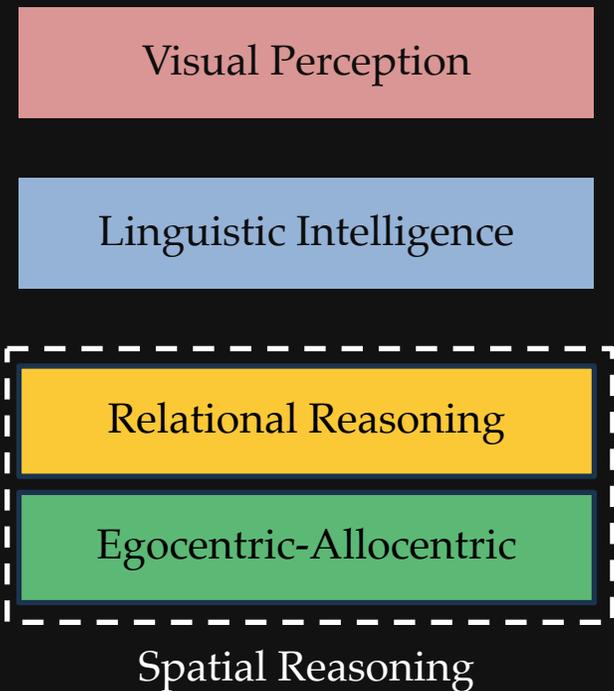


Error Breakdown

From 163 incorrect samples



71% spatial reasoning errors



Spatial reasoning is the main bottleneck for MLLMs on VSI-Bench

Takeaways

MLLMs exhibit sub-human visual-spatial intelligence.

Spatial reasoning, as opposed to visual perception or linguistic intelligence, is the main bottleneck to performance on VSI-Bench.

Linguistic prompting techniques, although effective in language reasoning and general visual tasks, are harmful for spatial reasoning.

When remembering spaces, a MLLM forms a series of local world models rather than a single global model.

So models fail at spatial reasoning...

Limiting factor = the training data.

How do we fix this?

⇒ 3D Simulators!

- **Idea:** Use simulators to generate spatial training data with *perfect* GT annos. No human labeling needed
- **SAT:** 175K spatial QA pairs from ProcTHOR scenes
- **Key result:** *Sim* training transfers to *real* evals
- **Limitation:** single-image only
 - *Video* is much harder
 - No help on VSI-Bench!

Simulated SAT Training



SAT: Dynamic Spatial Aptitude Training for Multimodal Language Models

Arijit Ray,¹ Jiafei Duan,^{2†} Ellis Brown,^{5†} Reuben Tan,^{1,4} Dina Bashkirova,¹ Rose Hendrix,³ Kiana Ehsani,³ Aniruddha Kembhavi,^{2,3} Bryan A. Plummer,¹ Ranjay Krishna,^{2,3*} Kuo-Hao Zeng,^{3*} Kate Saenko^{1*}

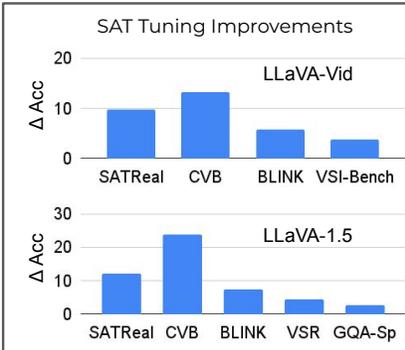
¹Boston University, ²University of Washington, ³Allen Institute for AI, ⁴Microsoft Research, ⁵New York University

<https://arijitray.com/SAT/>

Real-world improvements on spatial reasoning

Static and *dynamic* (reasoning about effect of ego and object motion)

Static (Existing Bench)	Perspective	Action Consequence	Goal Aim
			
Is the dog on the motorcycle? SAT (us): No, it is laying beside...	If I sit by the dog and face the same direction as the dog, will the lamp be to my left or right? SAT (us): Right	Is it safe to turn left and move forward in this picture? SAT (us): No	If I need to go to the back of the train, do I turn left or right? SAT (us): left
Object Movement	Egocentric Movement		
			
Were any of the objects moved from the 1st scene to 2nd? SAT (us): Chair was moved left	How did the camera rotate from the 1st scene to 2nd? SAT (us): Rotated right		



SIMS-V: Simulated Instruction-Tuning for Spatial Video Understanding

Ellis Brown¹ Arijit Ray² Ranjay Krishna³ Ross Girshick⁴ Rob Fergus¹ Saining Xie¹

¹New York University ²Boston University ³AllenAI ⁴Vercept

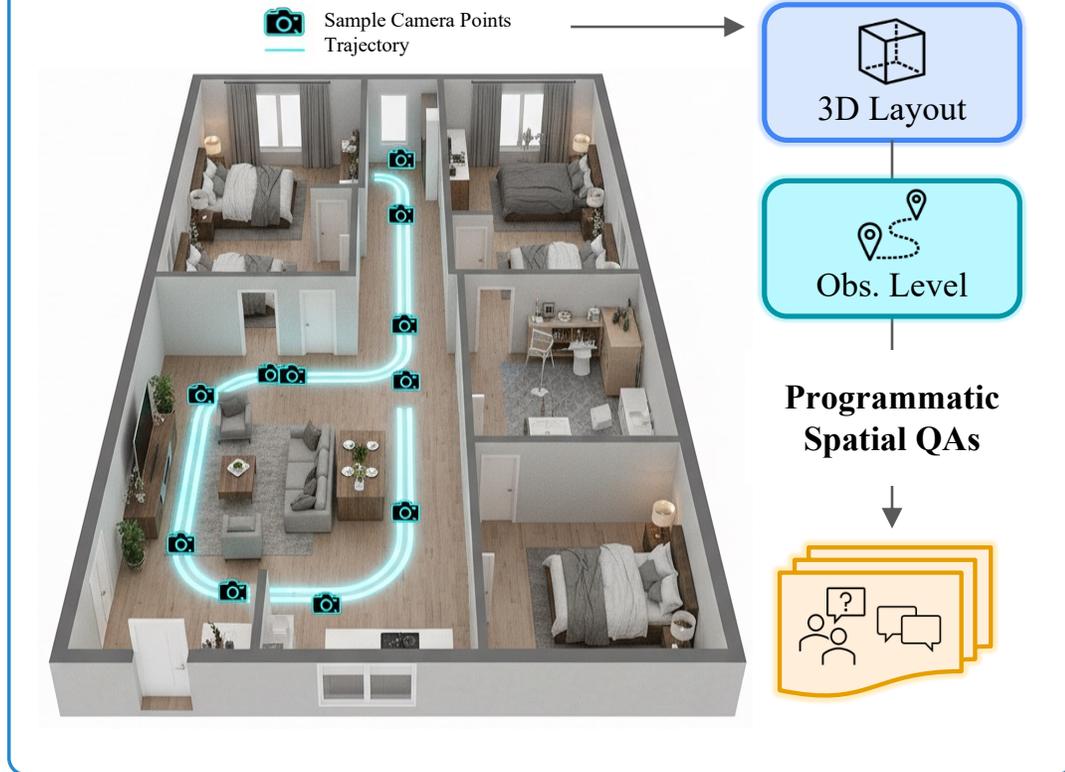
<https://ellisbrown.github.io/sims-v>

Motivation

- Video LLMs have strong general performance but lack spatial understanding
 - → Thinking in Space
- Real-world training data is hard to get
 - Unscalable

Overview

Simulating 3D-consistent spatial video training data ...



... that improves spatial understanding in *real* videos
on targeted “*in-domain*” tasks

VSI-Bench

+ 8.4%
LLaVA-Vid
+ 5.4%
LLaVA-OV

Q: What is the distance between the TV and the keyboard, in meters?

and on *out-of-domain* tasks

OpenEQA

Q: Can another cookie jar fit on the cookie jar shelf?

+ 8.6%
LLaVA-Vid

MME-RealWorld

Q: What is the future state of the white SUV in the middle?

+ 4.5%
LLaVA-Vid

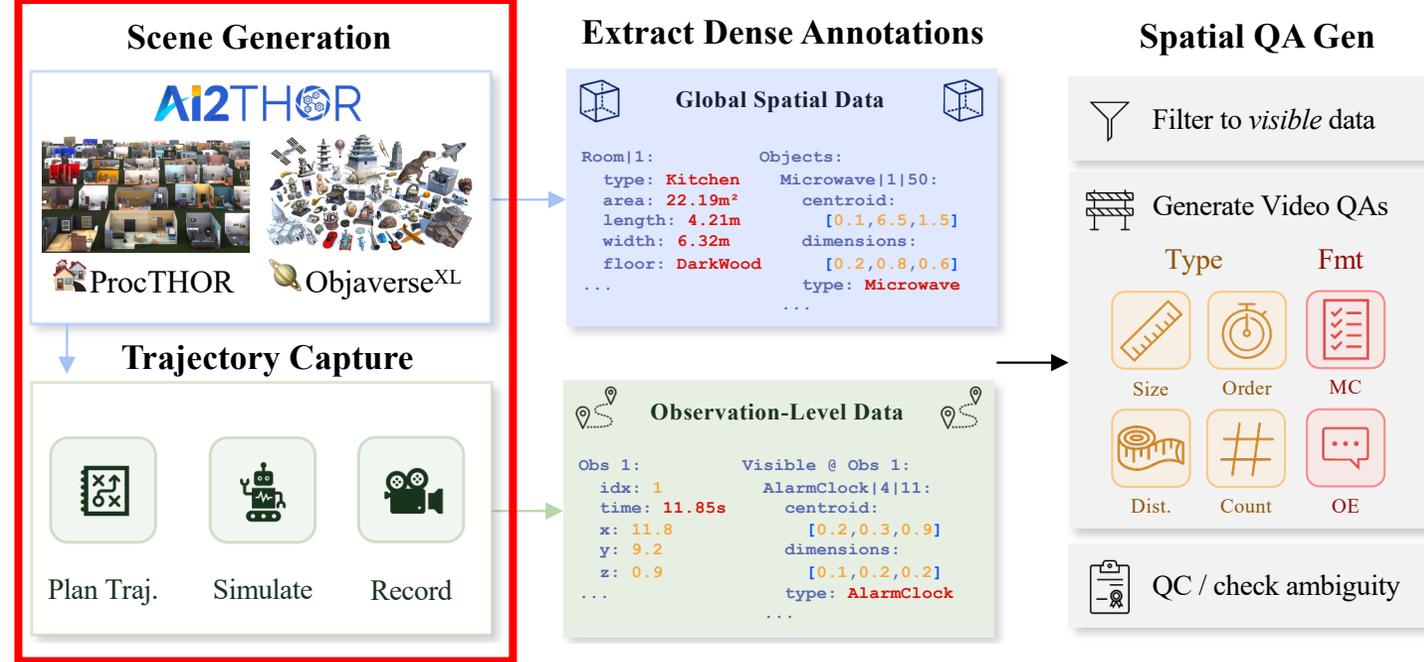
- **SIMS-V** leverages the spatial consistency of simulated 3D environments to create diverse, accurately-annotated training examples without the prohibitive costs of manual real-world data collection.
- ⇒ Demonstrate effective *Sim2Real* transfer

Simulated Spatial Instruction-Tuning

Scene Generation



Simulated Spatial Instruction-Tuning



1. Video Generation

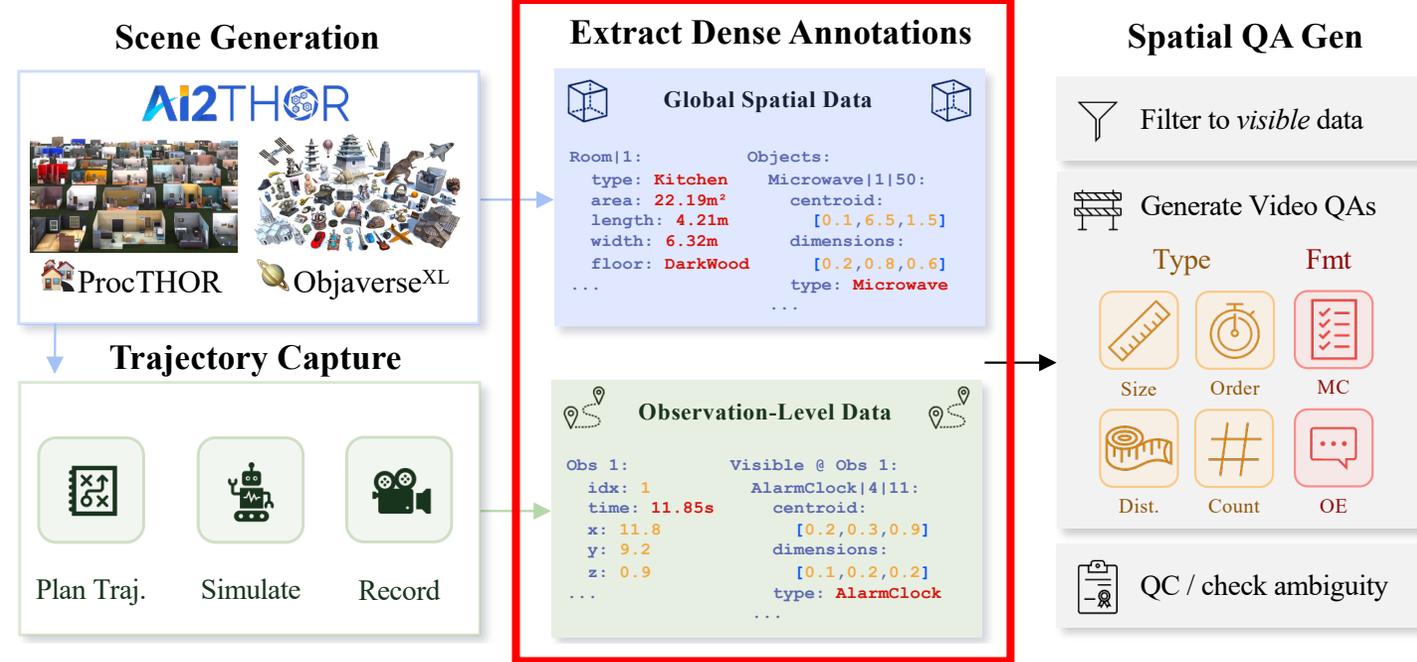
i. Scene Generation

- i. Simulator: Ai2THOR
- ii. Environments: ProcTHOR + Objaverse objects

ii. Trajectory capture

- i. Planner: use GT to plan S.P. traversal through every room in scene
- ii. Simulate/Record: execute plan in Ai2THOR (ultra quality), record raw observations

Simulated Spatial Instruction-Tuning



2. Extract Dense Annos

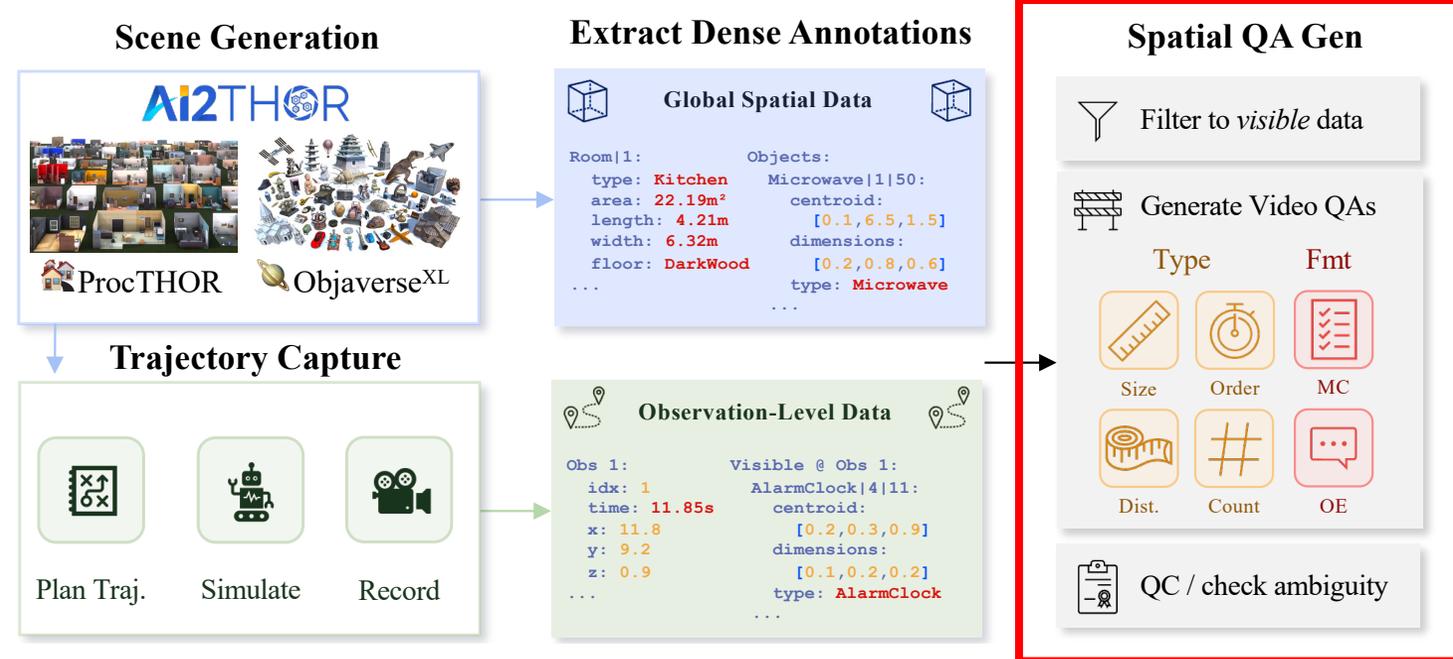
i. Observation-level Data

- i. Detailed info about what was seen @ each observation
 - i. Agent position
 - ii. Time
 - iii. Visible objects

ii. Global Spatial Data

- i. Comprehensive scene-level info directly from environment spec
 - i. Room layouts
 - ii. 3D obj positions, sizes etc + metadata

Simulated Spatial Instruction-Tuning



3. Spatial QA Gen

i. Filter / preprocess metadata

- i. Get set of objects that were visible at any point during trajectory
- ii. Use a minimum % of pixels as a “visibility” threshold

ii. Generate Qas

- i. Leverage the filtered global + observation level data with *various* preset templates to generate Qas

iii. Quality Control

- i. Added checks to see if MC answers are ambiguous, semantic category objects too similar, etc.

What matters for real-world transfer?

What properties of simulated data are most important for Sim2Real?

1 – *What Questions to Ask?*

- Endless types of questions we could generate using this flexible framework...
- Rigorous baseline for evaluating sim-to-real transfer: **Generate a *simulated* “trainset” for the VSI-Bench “testset”**
 - → generate questions that mirror the format and structure of those in the VSI-Bench evaluation set

VSI-Bench Question Baseline



Measurement



- Abs Dist Q: What is the distance between the fridge and the counter in meters?
A: 0.5
- Obj Size Q: What is the length of the longest side of the dining table?
A: 178
- Rm Size Q: What is the size of this entire house in sq. meters?
A: 72.3

Perspective



- Rel Dir Q: If you're by the fridge facing the counter, is the painting to the left/right?
A: Left
- Rel Dist Q: Which is closest to the counter: fridge, dining table, chair?
A: Fridge
- Obj Ct Q: How many tables are in this room?
A: 2

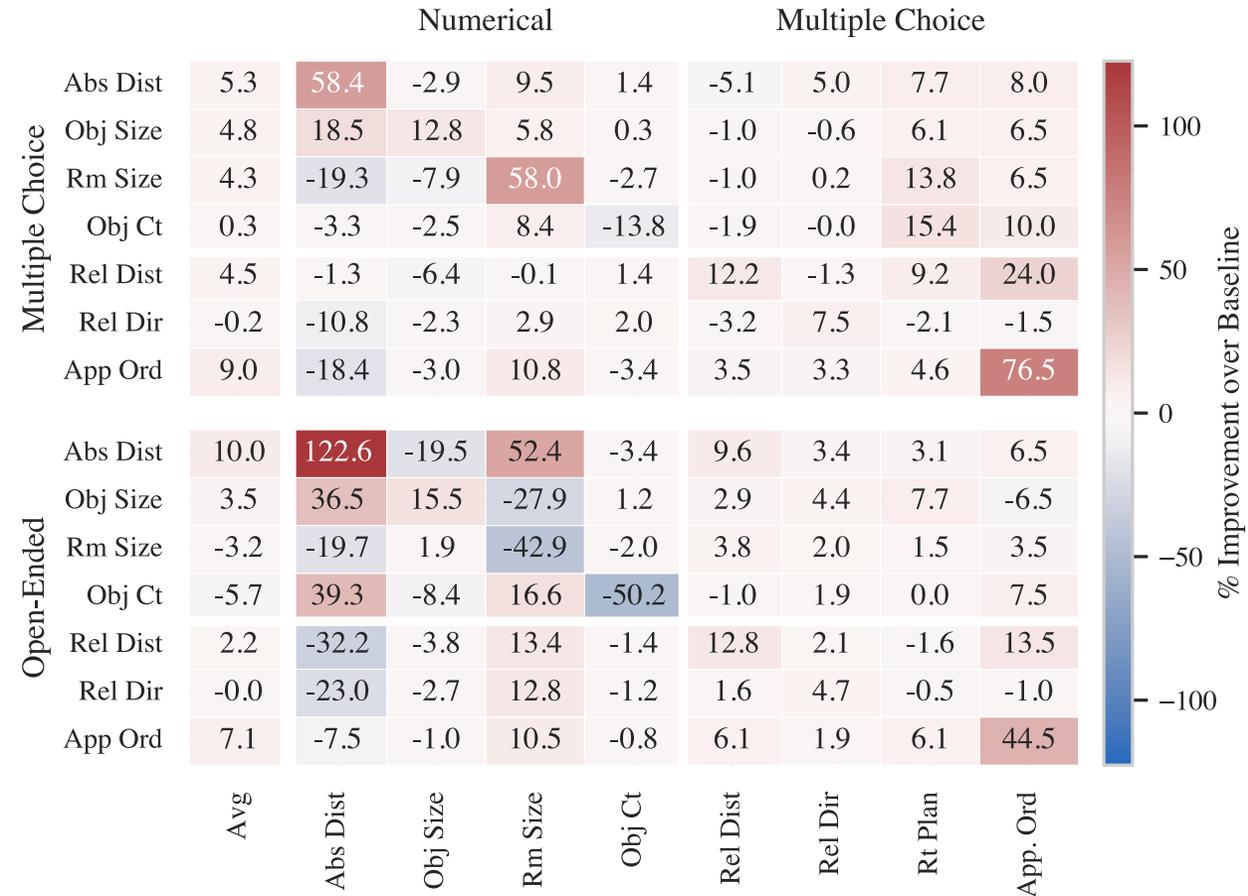
Spatiotemporal



- App Ord Q: What's the first appearance order of the: painting, fridge, window?
A: fridge, window, painting

VSI-Bench Question Baseline

- **Q:** impact of each *train-Q* on its corresponding *test-Q*
 - *Transfer effects?*
- FT one LLaVA-Video-7B for each train Q-type
 - 5K examples of *just* the Q-type
 - Evaluate *all* VSI-Bench test-set Q-types



What mix of questions to use?

VSI-Baseline Mix

- Clone the proportion of questions from the VSI-bench evaluation

3Q Minimal Mix

- *Minimal complementary Q-types* sufficient?
 - *Without* fully mirroring VSI-Bench?

Open-Ended (OE)		Multiple-Choice (MC)	
Q Type	Pct	Q Type	Pct
Abs Dist	16.9	Rel Dist	14.4
Obj Size	19.3	Rel Dir	19.6
Rm Size	5.8	Rt Plan	0.0
Obj Ct	11.4	App Ord	12.5

Three Q-types covering the core dimensions of spatial reasoning:

1. **Measurement:** Absolute distance estimation (open-ended) tests the ability to perceive and estimate metric properties of space.
2. **Perspective:** Relative direction determination (multiple-choice) evaluates understanding of perspective-dependent spatial configurations.
3. **Spatiotemporal:** Appearance order tracking (multiple-choice) assesses the ability to track objects across time and remember their temporal relationships.

Transfer to Real-World Benchmarks (VSI Bench)

		Measurement			Layout			Time	
Methods	Avg.	Abs Dist	Object Size	Room Size	Object Count	Relative Dist	Relative Dir	Route Plan	Appearance Order
		Numerical Answer			Multiple-Choice Answer				
<i>Proprietary Models</i>									
GPT-4o	34.0	5.3	43.8	38.2	46.2	37.0	41.3	31.5	28.5
Gemini-1.5 Flash	42.1	30.8	53.5	54.4	49.8	37.7	41.0	31.5	37.8
Gemini-1.5 Pro	45.4	30.9	64.1	43.6	56.2	51.3	46.3	36.0	34.6
<i>Open-source Models</i>									
LLaVA-Video 7B	35.6	14.0	47.8	24.2	48.5	43.5	42.4	34.0	30.6
LLaVA-Video 72B	40.9	22.8	57.4	35.3	48.9	42.4	36.7	35.0	48.6
<i>LLaVA-Video 7B Finetuned on SIMS</i>									
VSI-Baseline Mix (25k)	41.8	37.8	51.3	25.6	23.4	58.3	48.2	34.5	55.5
Δ Increase	+6.2	+23.8	+3.5	+1.4	-25.1	+14.8	+5.8	+0.5	+24.9
% Increase	+17.4	+170.0	+7.3	+5.8	-51.8	+34.0	+13.7	+1.5	+81.4
3Q Minimal Mix (25k)	44.6	35.5	41.2	38.0	46.7	53.8	47.9	35.1	58.6
Δ Increase	+9.0	+21.5	-6.6	+13.8	-1.8	+10.3	+5.5	+1.1	+28.0
% Increase	+25.3	+153.6	-13.8	+56.9	-3.7	+23.7	+13.0	+3.2	+91.5

Table 1. Performance comparison on VSI-Bench showing that our SIMS data can be used to train models that are competitive with state-of-the-art proprietary models which use significantly more data, parameters, and longer context-lengths. Our 3Q Minimal Mix achieves competitive results against larger proprietary models while showing substantial gains in spatial reasoning tasks, particularly in appearance order (+28.0%) and absolute distance estimation (+25.3%).

Transfer to Real-World Benchmarks (Other)

	EgoSchema
LLaVA-Video 7B	56.9
VSI-Baseline Mix (25k)	57.8
Δ Increase	+0.9
% Increase	+1.6
3Q Minimal Mix (25k)	57.7
Δ Increase	+0.8
% Increase	+1.4

Table 2. Performance on the EgoSchema benchmark demonstrates that models trained on our synthetic SIMS data transfer effectively to real-world egocentric videos. Importantly, our approach not only achieves substantial gains on VSI-Bench (Tab. 1) but also improves performance on this independent benchmark, indicating that our method teaches generalizable spatial reasoning rather than overfitting to a specific evaluation.

Methods	Static Image Spatial Benchmarks		
	BLINK	CV-Bench 2D	CV-Bench 3D
LLaVA-Video 7B	56.8	67.3	63.0
VSI-Baseline Mix (25k)	58.7	65.8	68.2
Δ Increase	1.9	-1.5	5.2
% Increase	3.4	-2.3	8.3
3Q Minimal Mix (25k)	58.4	65.4	67.2
Δ Increase	+1.7	-1.9	+4.2
% Increase	+2.9	-2.8	+6.7

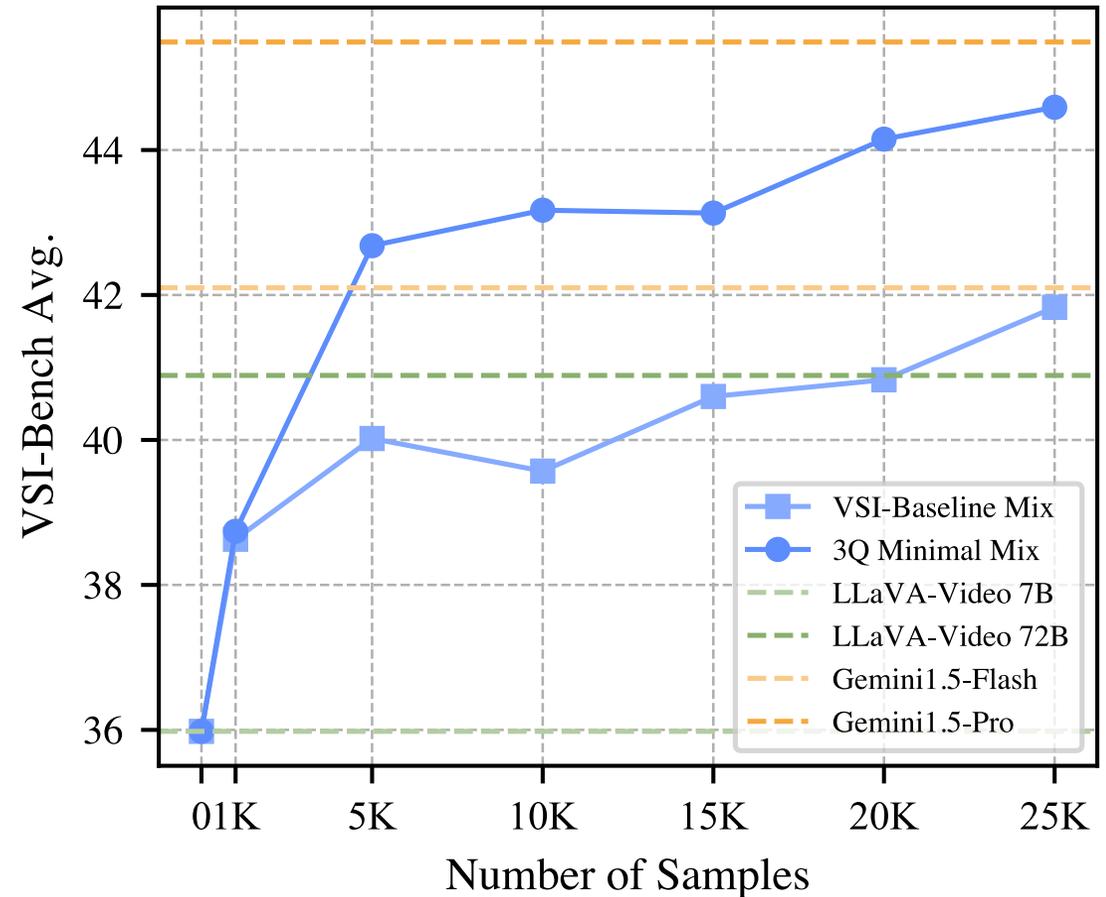
Table 3. Performance comparison on static image spatial benchmarks. Our 3Q Minimal Mix model, despite being trained on synthetic video data, maintains or improves performance on independent image-based spatial reasoning tasks. This demonstrates that our approach teaches general spatial understanding rather than overfitting to specific video scenarios or benchmarks.

Data Scaling

Goal: understand *how many simulated examples are required* to achieve strong spatial reasoning capabilities

3Q Minimal Mix:

- Surpasses 72B model + Gemini1.5-Flash with <5k samples
- Competitive with *SOTA* Gemini1.5-Pro after ~20k samples



Implications for Simulator-Based Training

- **Computational Efficiency:** Lower-quality rendering requires less computational resources, enabling faster and more efficient data generation at scale.
- **Domain-Invariant Features:** Success with abstract visual representations suggests models are learning domain-invariant spatial reasoning that transfers well despite visual domain gaps.
- **Structure Over Appearance:** The strong performance of training on simplified representations indicates that consistent spatial structure and accurate annotations are more important than visual realism.

Nice results!

But during these experiments,
something strange happened that
led to a whole new paper...

Benchmark Designers Should “Train on the Test Set” to Expose Exploitable Non-Visual Shortcuts

Ellis Brown Jihan Yang Shusheng Yang Rob Fergus Saining Xie

New York University

Origin Story

Rewind: March 2025

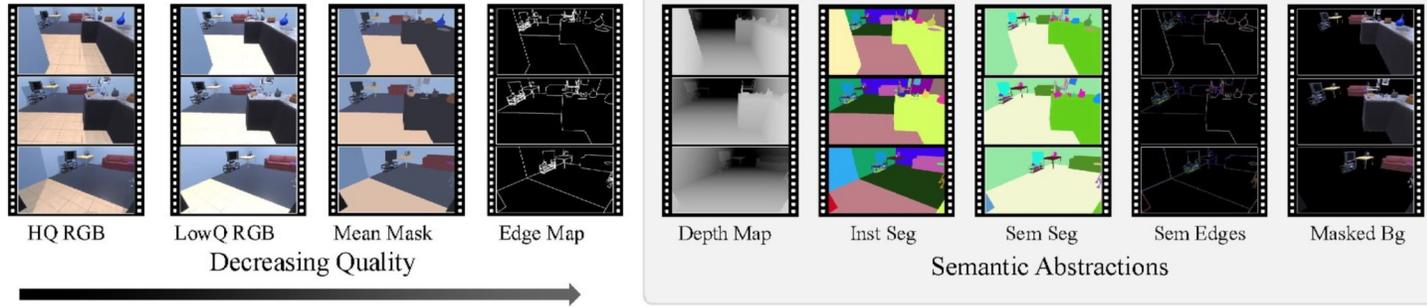
Can *Simple* Simulators Elicit Spatial Understanding in Video Language Models?

3/18/2025

ICCV 2025 submission

Ellis Brown, Arijit Ray, Ranjay Krishna, Ross Girshick, Rob Fergus, Saining Xie

Rewind: March 2025



Decreasing Quality

- HQ RGB (default)
- LowQ RGB: simplified textures, lighting
- Mean Mask: avg RGB value in each inst. mask
- Edge Map: black w/ white edges around instances

Semantic Abstractions

- Depth Map
- Instance Seg
- Semantic Seg → colored by obj semantic category
- Semantic Edges → edges ^^
- Masked BG → floors+walls deleted

- Low-quality visual inputs were not much worse than the best RGB!

- hmmm

Rewind: March 2025

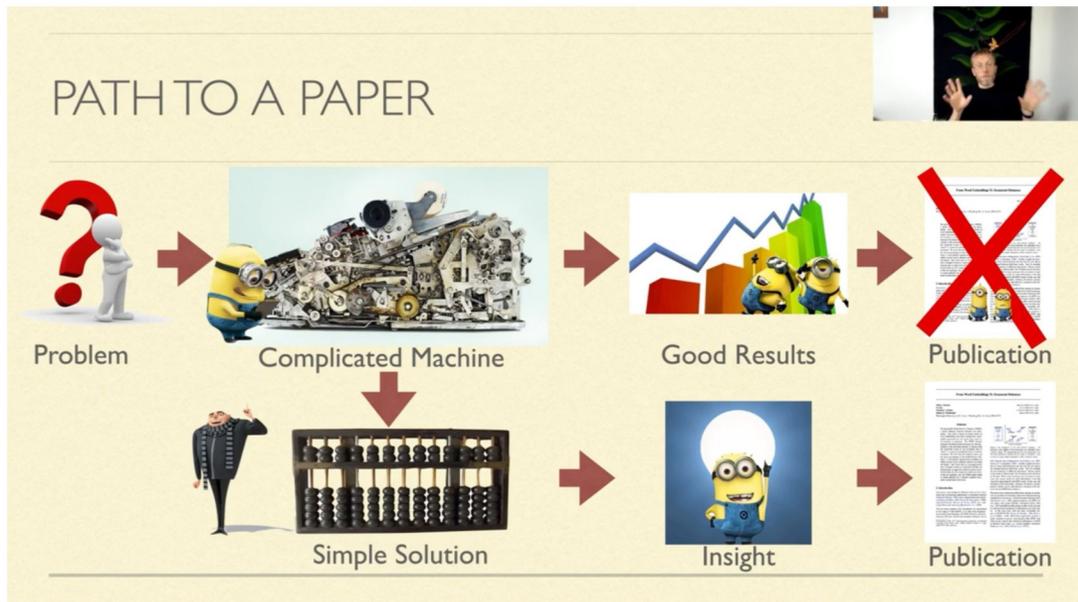
Open Questions

- WHY does the visual quality not matter?
- What actually matters?
- What do we have to do to break performance gains?

Rewind: March 2025

On the Importance of Deconstruction in Machine Learning Research – Killian Weinberger

- Did exactly this!



- **Strive for simplicity** instead of complexity
- Complex solutions can be an important means to an end
- Don't be afraid to (brutally) **deconstruct** your findings!
- Publish your new understanding of the problem

Rewind: March 2025

How simple can we ***actually*** go?

- Question: is it possible that the model is just learning the question formats and the visual input does not matter at all?
- HARD Visual inputs:
 - All-black input
 - Subsample / downscale → 8x, 32x, 64x
 - Shuffle temporal order
 - Single frame input (first or random frame)
 - No input (just train on text)

+6%

with NO visual input

**What are these
models *actually*
learning?**

The Provocative Claim

"Train on the test set"

Not to cheat; to *diagnose shortcuts*

Core argument: **If your benchmark CAN be gamed, it WILL be**

Proactive auditing > reactive discovery

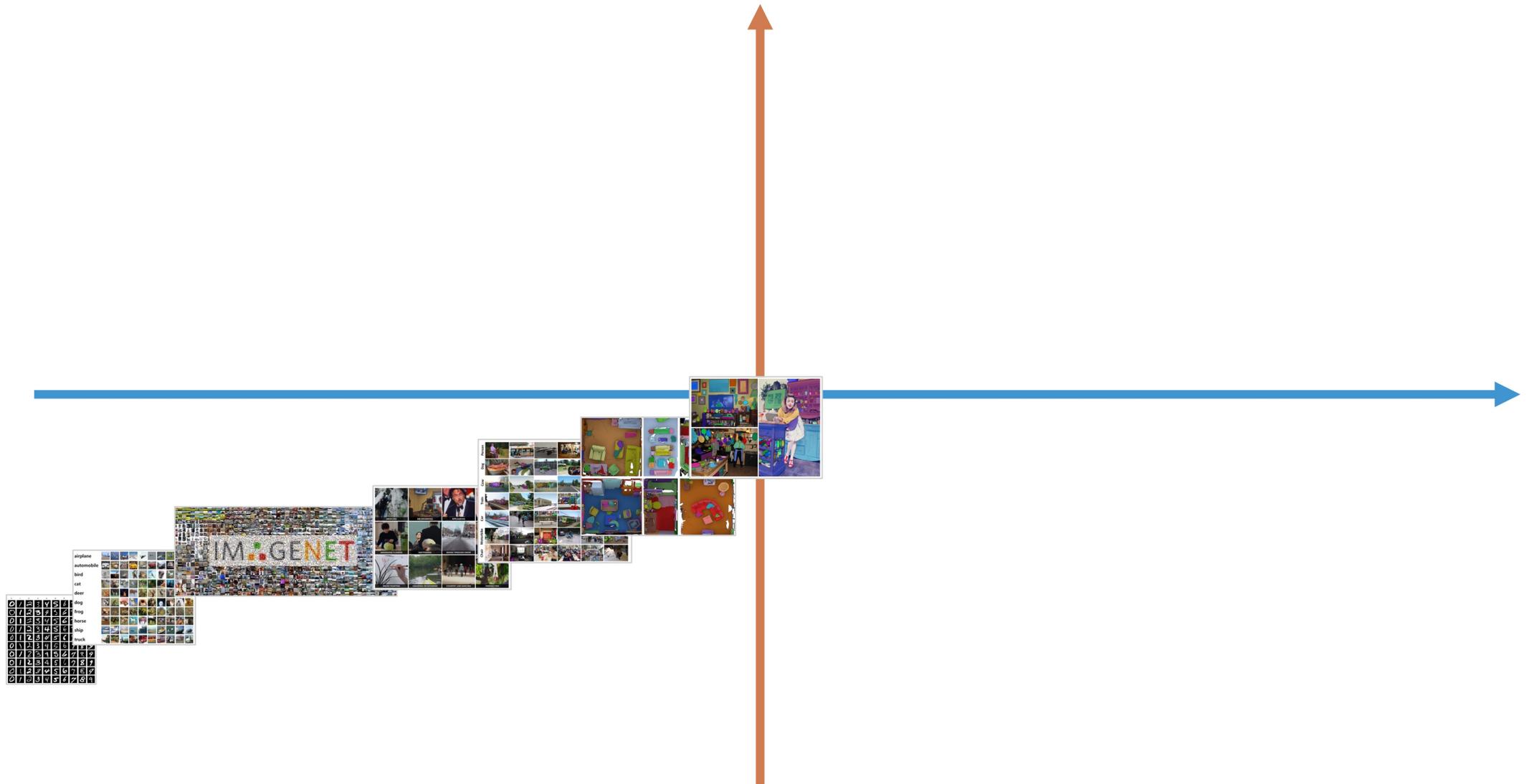
This talk:

- I. The problem (*non-visual shortcuts*)
- II. TsT framework (*diagnosis*)
- III. IBP (*debiasing*)

I. The Problem

Non-visual shortcuts

Benchmarking Visual Understanding



Cambrian-1 “Blind” Test

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, Ziteng Wang,
Rob Fergus, Yann LeCun, Saining Xie[†]

New York University

Abstract

We introduce Cambrian-1, a family of multimodal LLMs (MLLMs) designed with a vision-centric approach. While stronger language models can enhance multimodal capabilities, the design choices for vision components are often insufficiently explored and disconnected from visual representation learning research. This gap hinders accurate sensory grounding in real-world scenarios. Our study uses LLMs and visual instruction tuning as an interface to evaluate

Cambrian-1 “Blind” Test

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, Ziteng Wang,
Rob Fergus, Yann LeCun, Saining Xie[†]

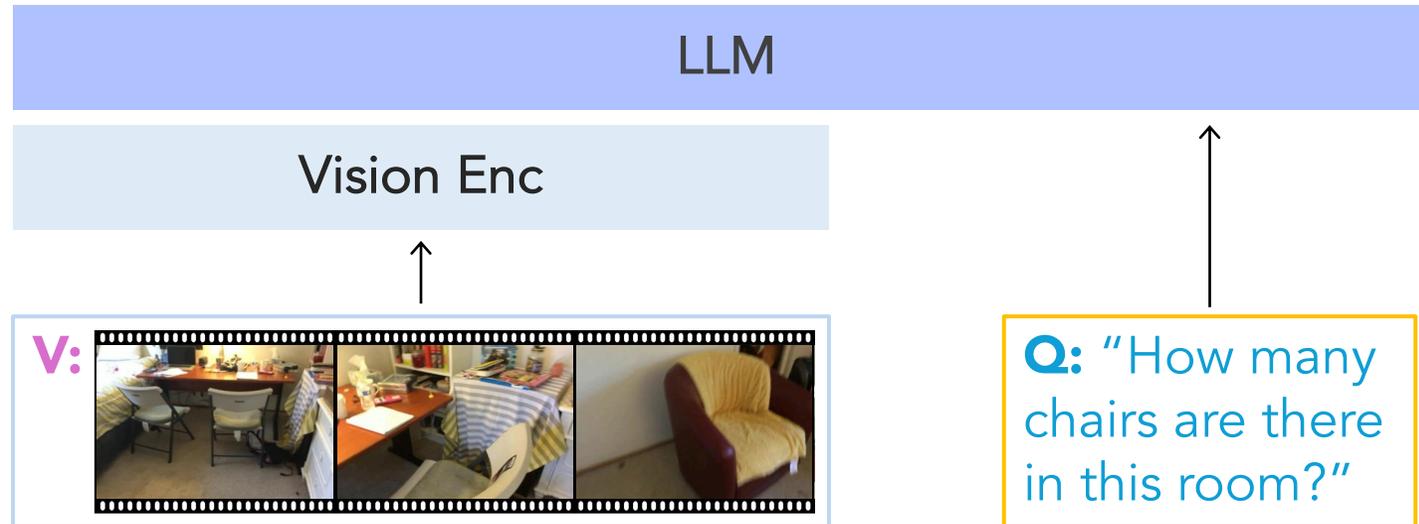
New York University

Abstract

We introduce Cambrian-1, a family of multimodal LLMs (MLLMs) designed with a vision-centric approach. While stronger language models can enhance multimodal capabilities, the design choices for vision components are often insufficiently explored and disconnected from visual representation learning research. This gap hinders accurate sensory grounding in real-world scenarios. Our study uses LLMs and visual instruction tuning as an interface to evaluate various visual representations, offering new insights into different models and architectures—self-supervised, strongly supervised, or combinations thereof—based on experiments with over 20 vision encoders. We critically examine existing MLLM benchmarks, address the difficulties involved in consolidating and interpreting results from various tasks, and introduce a new vision-centric benchmark, CV-Bench. To further improve visual grounding, we propose the Spatial Vision Aggregator (SVA), a dynamic and spatially-aware connector that integrates high-resolution vision features with LLMs while reducing the number of tokens. Additionally, we discuss the curation of high-quality visual instruction-tuning data from publicly available sources, emphasizing the importance of data source balancing and distribution ratio. Collectively, Cambrian-1 not only achieves state-of-the-art performance but also serves as a comprehensive, open cookbook for instruction-tuned MLLMs. We provide model weights, code, supporting tools, datasets, and detailed instruction-tuning and evaluation recipes. We hope our release will inspire and accelerate advancements in multimodal systems and visual representation learning.

Website <https://cambrian-llm.github.io>
Code <https://github.com/cambrian-llm/cambrian>
Models <https://huggingface.co/nyu-visionx/>
Data <https://huggingface.co/datasets/nyu-visionx/Cambrian-10M>
CV-Bench <https://huggingface.co/datasets/nyu-visionx/CV-Bench>
Evaluation <https://github.com/cambrian-llm/cambrian#evaluation>

*Project Lead
[†]Corresponding Author



Cambrian-1 "Blind" Test

arXiv:2406.16860v2 [cs.CV] 4 Dec 2024

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, Ziteng Wang, Rob Fergus, Yann LeCun, Saining Xie[†]

New York University

Abstract

We introduce Cambrian-1, a family of multimodal LLMs (MLLMs) designed with a vision-centric approach. While stronger language models can enhance multimodal capabilities, the design choices for vision components are often insufficiently explored and disconnected from visual representation learning research. This gap hinders accurate sensory grounding in real-world scenarios. Our study uses LLMs and visual instruction tuning as an interface to evaluate various visual representations, offering new insights into different models and architectures—self-supervised, strongly supervised, or combinations thereof—based on experiments with over 20 vision encoders. We critically examine existing MLLM benchmarks, address the difficulties involved in consolidating and interpreting results from various tasks, and introduce a new vision-centric benchmark, CV-Bench. To further improve visual grounding, we propose the Spatial Vision Aggregator (SVA), a dynamic and spatially-aware connector that integrates high-resolution vision features with LLMs while reducing the number of tokens. Additionally, we discuss the curation of high-quality visual instruction-tuning data from publicly available sources, emphasizing the importance of data source balancing and distribution ratio. Collectively, Cambrian-1 not only achieves state-of-the-art performance but also serves as a comprehensive, open cookbook for instruction-tuned MLLMs. We provide model weights, code, supporting tools, datasets, and detailed instruction-tuning and evaluation recipes. We hope our release will inspire and accelerate advancements in multimodal systems and visual representation learning.

Website <https://cambrian-llm.github.io>
Code <https://github.com/cambrian-llm/cambrian>
Models <https://huggingface.co/nyu-visionx/>
Data <https://huggingface.co/datasets/nyu-visionx/Cambrian-10M>
CV-Bench <https://huggingface.co/datasets/nyu-visionx/CV-Bench>
Evaluation <https://github.com/cambrian-llm/cambrian#evaluation>

*Project Lead
[†]Corresponding Author



Cambrian-1 "Blind" Test

arXiv:2406.16860v2 [cs.CV] 4 Dec 2024

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, Ziteng Wang, Rob Fergus, Yann LeCun, Saining Xie*

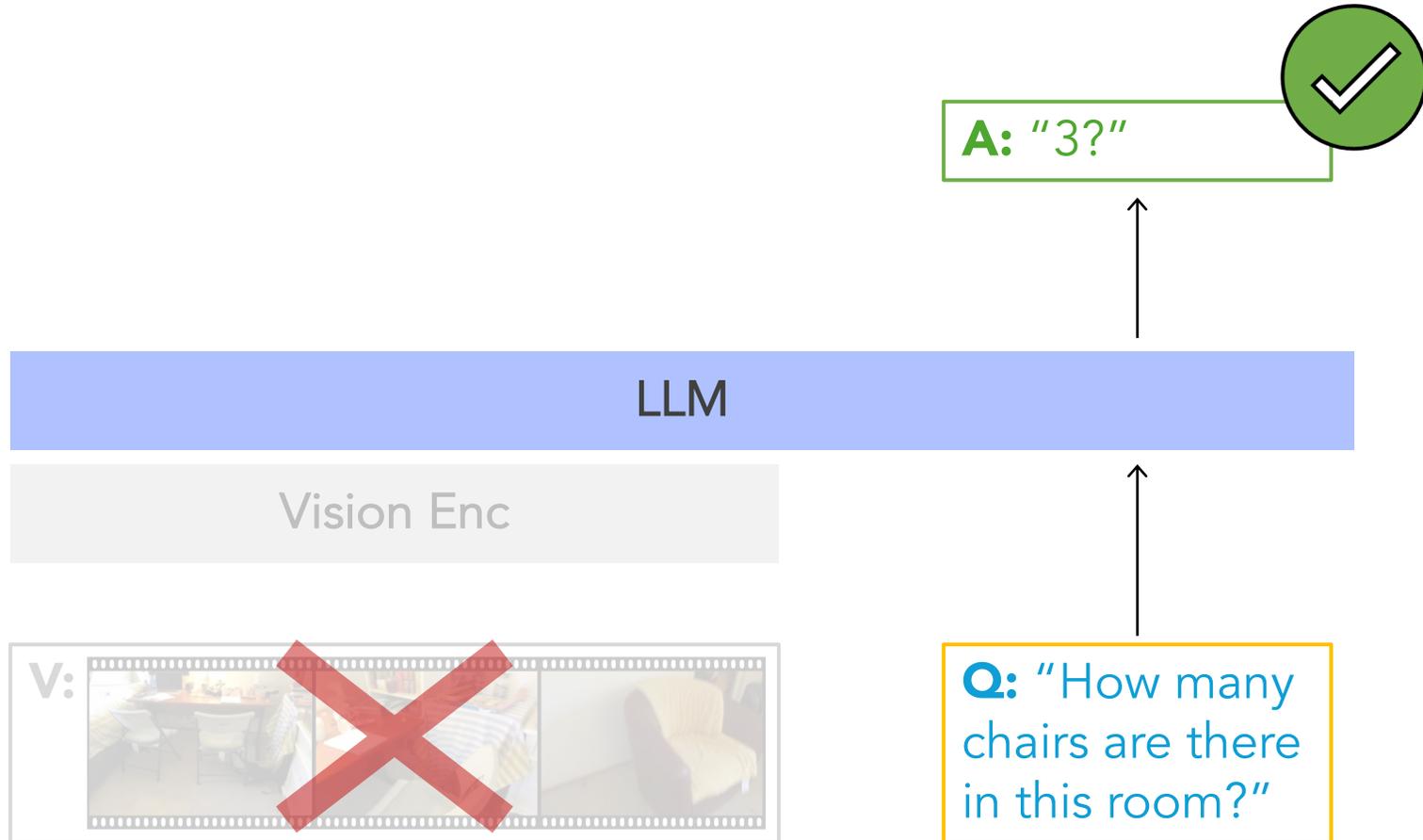
New York University

Abstract

We introduce Cambrian-1, a family of multimodal LLMs (MLLMs) designed with a vision-centric approach. While stronger language models can enhance multimodal capabilities, the design choices for vision components are often insufficiently explored and disconnected from visual representation learning research. This gap hinders accurate sensory grounding in real-world scenarios. Our study uses LLMs and visual instruction tuning as an interface to evaluate various visual representations, offering new insights into different models and architectures—self-supervised, strongly supervised, or combinations thereof—based on experiments with over 20 vision encoders. We critically examine existing MLLM benchmarks, address the difficulties involved in consolidating and interpreting results from various tasks, and introduce a new vision-centric benchmark, CV-Bench. To further improve visual grounding, we propose the Spatial Vision Aggregator (SVA), a dynamic and spatially-aware connector that integrates high-resolution vision features with LLMs while reducing the number of tokens. Additionally, we discuss the curation of high-quality visual instruction-tuning data from publicly available sources, emphasizing the importance of data source balancing and distribution ratio. Collectively, Cambrian-1 not only achieves state-of-the-art performance but also serves as a comprehensive, open cookbook for instruction-tuned MLLMs. We provide model weights, code, supporting tools, datasets, and detailed instruction-tuning and evaluation recipes. We hope our release will inspire and accelerate advancements in multimodal systems and visual representation learning.

Website <https://cambrian-llm.github.io>
Code <https://github.com/cambrian-llm/cambrian>
Models <https://huggingface.co/nyu-visionx/>
Data <https://huggingface.co/datasets/nyu-visionx/Cambrian-10M>
CV-Bench <https://huggingface.co/datasets/nyu-visionx/CV-Bench>
Evaluation <https://github.com/cambrian-llm/cambrian#evaluation>

*Project Lead
*Corresponding Author



Cambrian-1 "Blind" Test



Non-Visual Biases!
Solvable without vision



Q: "How many chairs are there in this room?"

A: "3?"



Cambrian-1 *"Blind" Test*



Non-Visual Biases!
Solvable without vision



BUT: offers *no insight* into

- *why* specific samples are exploitable
- *how* to fix the shortcuts

Models Can Ace Multimodal Benchmarks Without Vision

High scores \neq visual understanding.

\Rightarrow Particularly problematic for “*vision-centric*” benchmarks that explicitly claim to require vision.

Knowledge Shortcuts

Exploiting prior world knowledge

Example:

Q: "How tall is the refrigerator?"

\rightarrow Model knows fridges are \sim 1.7m without looking

Primary source: LLM Pretraining.

\rightarrow *can* be useful, but still be a shortcut!

Statistical Shortcuts

Exploiting **idiosyncrasies** of the dataset / QA pairs themselves

Example:

Q: "How many chairs?"

\rightarrow $>$ 50% of answers are "2", so just guess "2"

Sources: sampling logic, procedural gen., templates, human filtering, etc...

Orthogonal vulnerabilities \rightarrow a benchmark can suffer from one, both, or neither.

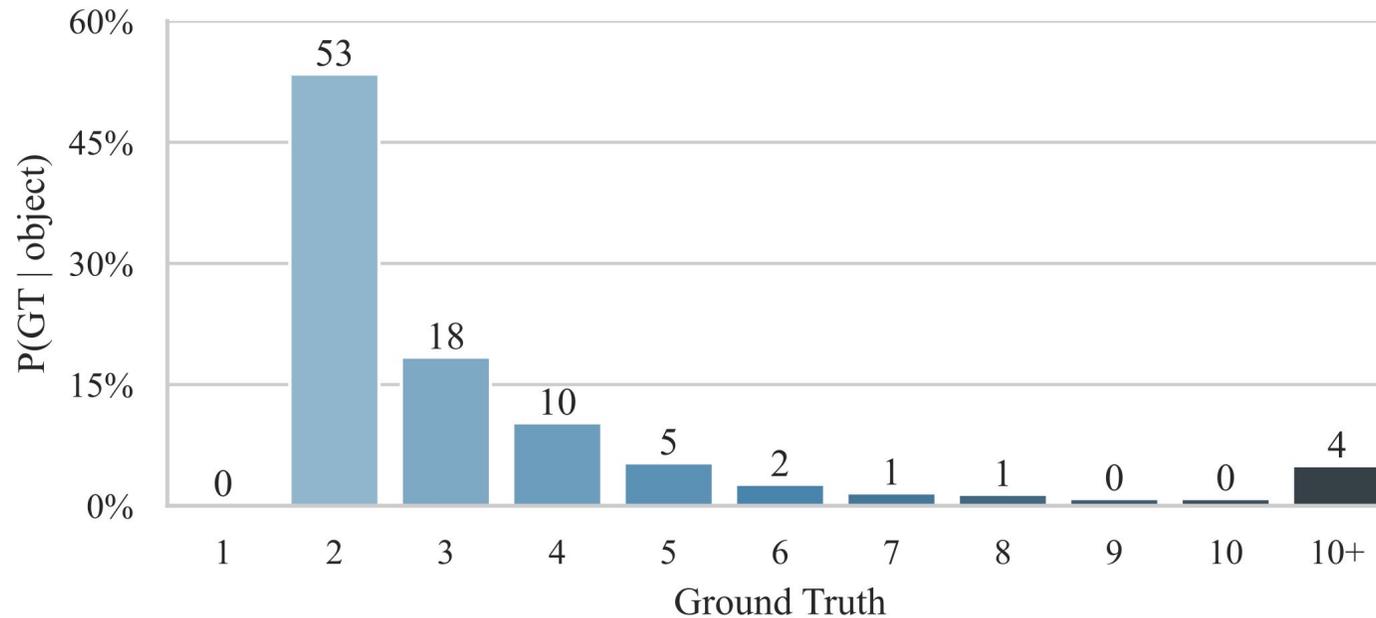
What Constitutes a Non-Visual Shortcut?

A pattern (whether reflecting natural world knowledge, real-world statistical regularities, or procedural generation artifacts) becomes *an exploitable non-visual shortcut* if it renders the visual input redundant for a task designed to measure visual understanding

Statistical Biases are Pervasive



Q: How many objects are in the image?



(a) Counting

>50%

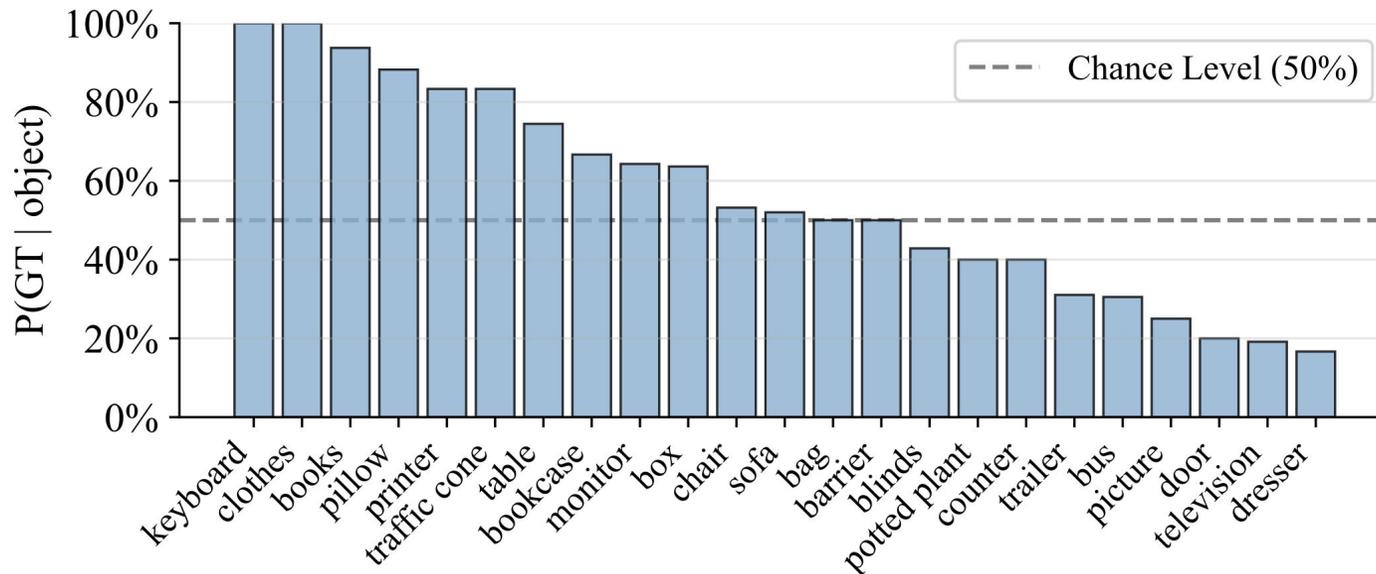
of ground truth
answers are ≤ 3

→ Just guess "2" and
score well

Statistical Biases are Pervasive



Q: Which is closer to object 1: object 2 or object 3?



100%

of "keyboard" and "clothes" choices are correct *whenever* they appear as options

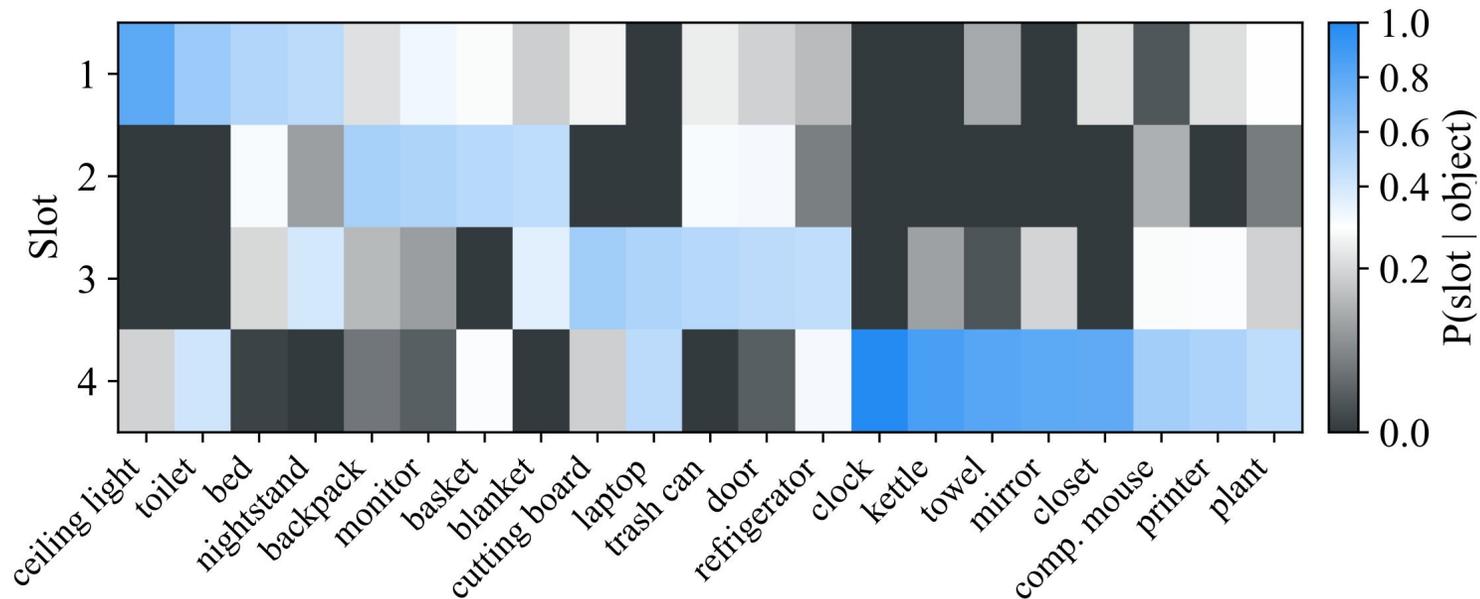
(b) Spatial relation

All patterns exploitable without looking at a single frame

Statistical Biases are Pervasive



Q: What is the appearance order for obj 1, obj 2, obj 3, obj 4?



100%

of the time "clock" appears in a GT sequence, it is in slot 4 (n=50)

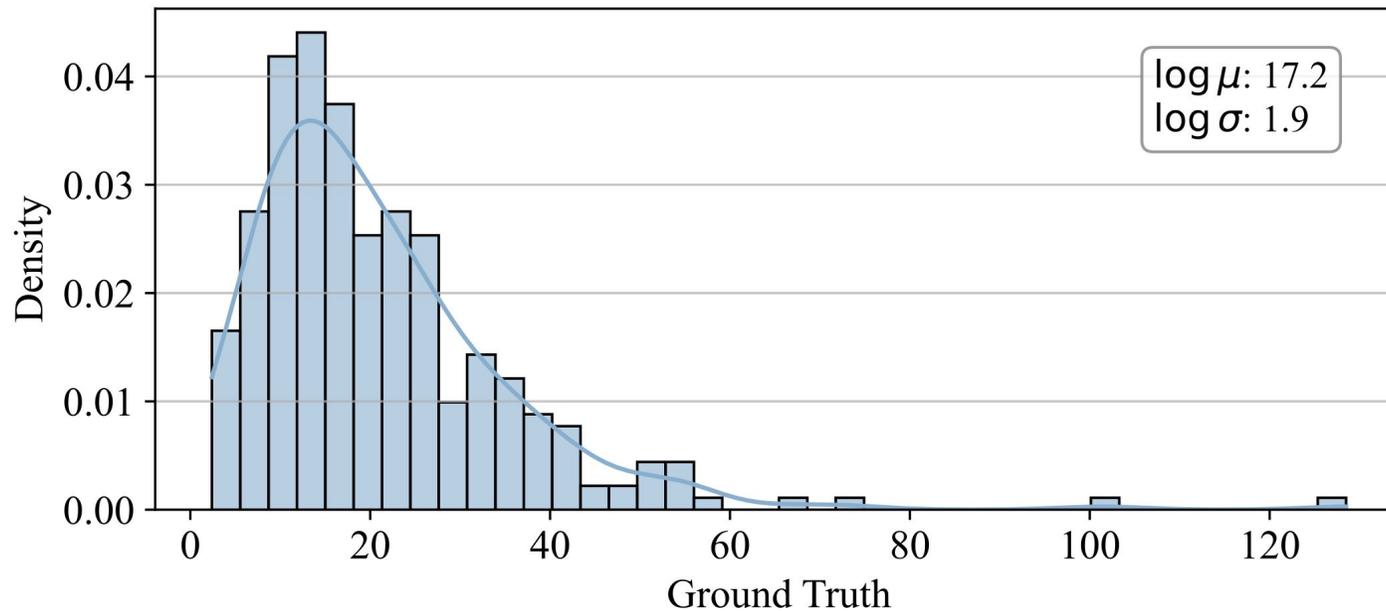
(c) Appearance order

All patterns exploitable without looking at a single frame

Statistical Biases are Pervasive



Q: What is the length/width/size for the object/room?



(d) Size estimation

Log-normal

Sizes and measurements follow predictable distributions

Models Readily Exploit These Shortcuts

VSI-Train-10k → a 10k *in-dist* train-set for VSI-Bench following the same generation pr

Configuration	Vision	Blind	Δ_{V-B}
LLaVA-Video 7B (Base)	36.7	25.9	10.8
+ VSI-Train-10k FT	57.1	44.7	12.4
Δ due to FT	+20.4	+18.8	+1.6
Chance (frequency)		34.0	

LLaVA-Video-7B on VSI-Bench before & after FT on VSI-Train-10k

The problem: Fine-tuning improves blind and vision *almost equally*

Models learn statistical shortcuts that benefit both configurations. Does not *require* visual understanding!

CORE PRINCIPLE

If a benchmark *can* be gamed,
it *will* be

Designers should **proactively** try to game their own benchmarks first

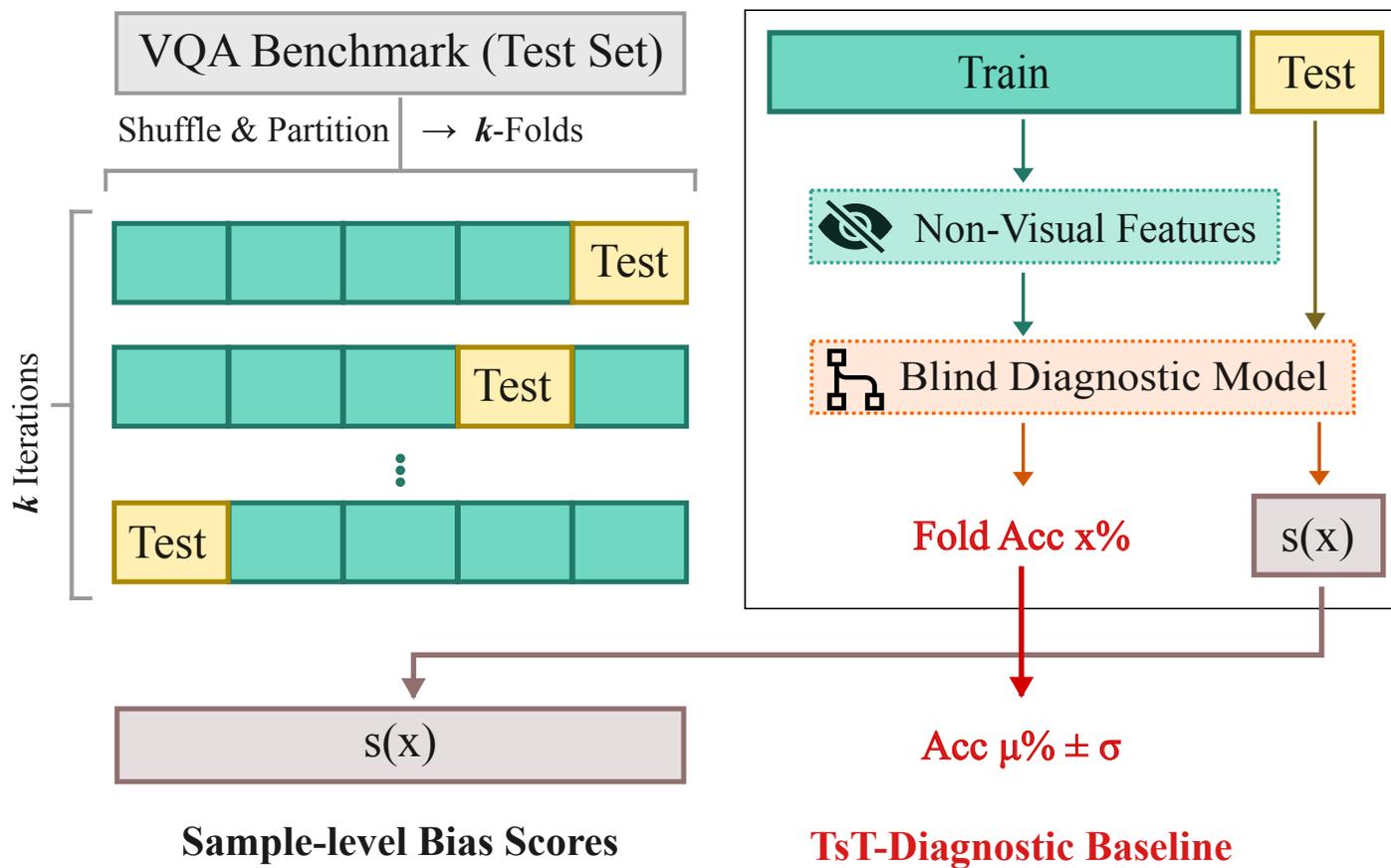
The most rigorous stress test = training directly on the test set

Not to overfit; to **diagnose**

II. Diagnosing Shortcuts

TsT: Test-set Stress-Testing

TsT: Test-set Stress-Test



INPUT

Test Set Questions

Non-visual features only

K-Fold Cross-Validation ($k=5$)

Train diagnostic model on text/metadata only

OUTPUT 1

TsT Accuracy

Global exploitability measure

OUTPUT 2

Bias Scores $s(x)$

Per-sample vulnerability

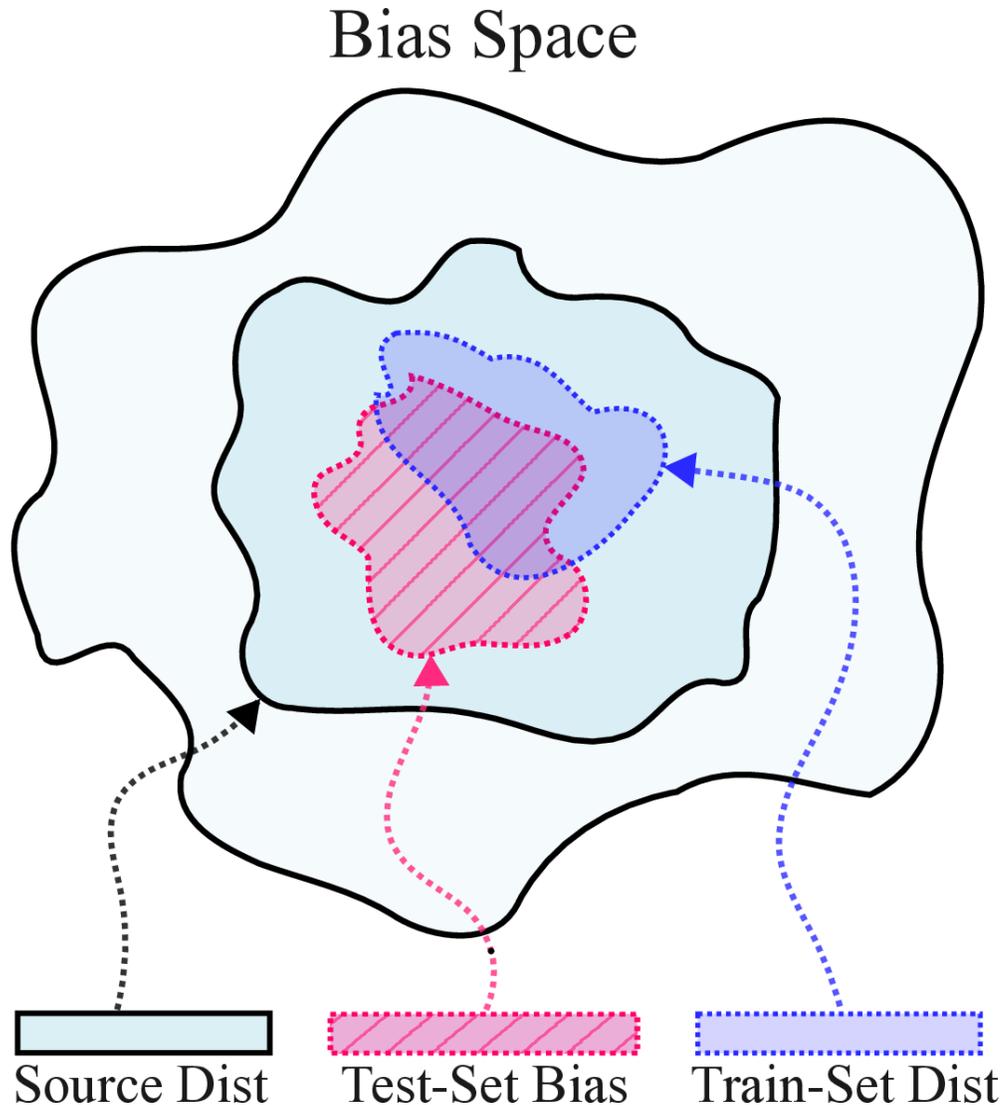
TsT: Test-set Stress-Test



Why the “*Test Set*” Specifically?

*“Is it actually crucial to train on the test set specifically?
Why not just use a held-out training set?”*

Why the “Test Set” Specifically?



“Is it actually crucial to train on the test set specifically? Why not just use a held-out training set?”

YES!

Training on held-out data (like VSI-Train) reveals *generalizable* biases...

...but **MISSES** idiosyncratic artifacts unique to the test set:

- Sampling quirks
- Template artifacts
- Human filtering decisions

TsT *directly* probes the *specific artifact* that will *actually* be evaluated on.

TsT: Test-set Stress-Test

Two complementary variants:

TST-LLM

Model: Qwen2.5-7B + LoRA
Features: Raw question text
Compute: ~20 min on 4×A100
Interpretability: Limited

✓ **Complex patterns**
✓ **Any benchmark**
✓ **Trivial to impl**

Best for: catching complex shortcuts

TST-RF

Model: Random Forest
Features: Hand-crafted
Compute: Minutes on CPU
Interpretability: High (Gini)

✓ **Interpretability**
✓ **Fast iteration**
✗ **Feature Eng**

Best for: understanding why

TsT Results: Exploitability Across Benchmarks

TsT-RF Accuracy

Benchmark	Chance Acc.	Majority Acc.	TsT-RF CV-Acc.
CV-Bench	33.3	43.1	75.5
VSI-Bench	-	34.0	43.5

TsT-LLM Accuracy

Benchmark	Blind ZS Acc.	TsT-LLM CV-Acc.	Δ_{TsT}
CV-Bench	40.1	73.4	+33.3
VSI-Bench	25.0	56.4	+31.4
MMMU (val)	34.9	43.5	+8.6
VideoMME	35.3	41.7	+6.4

Applicable to **template-based** BMs

→ dramatically exploitable
Even with a simple blind RF model!

Works on **freeform/natural language** BMs!
Strong learnable biases
Simple first-pass diagnostic

TsT-RF gives *interpretable* insights

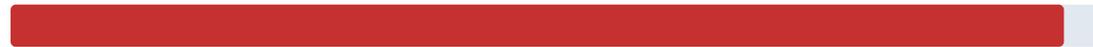


Task: Size Estimation

"What is the length of the longest dimension of {object} in cm?"

Gini Feature Importance

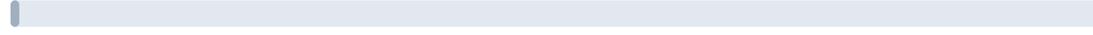
obj_val_log_mean



object (category)



obj_val_log_std



obj_count



One feature dominates:
`obj_val_log_mean` = 0.968

Translation: The RF just memorizes typical object sizes

- "dishwashers are ~90cm"
- "beds are ~216cm" ...

✓ Actionable Insight:

Questions about low-variance size objects (dishwasher, bed, washer) contribute more to bias on this question type

III. Debiasing Shortcuts

Iterative Bias Pruning (IBP)

IBP: Diagnosis → Debiasing

Iterative Bias Pruning

→ Compute bias scores $s(x)$ for all samples

→ Remove batch of highest-scoring samples

Re-compute $s(x)$ ← Critical!

→ Repeat until threshold met

```
while budget > 0 and TST_acc > t:  
    scores = compute_TST_scores(  
        remaining_data)  
    remove top_b samples by score  
    budget -= b
```

IBP: Diagnosis → Debiasing

Iterative Bias Pruning

→ Compute bias scores $s(x)$ for all samples

→ Remove batch of highest-scoring samples

Re-compute $s(x)$ ← Critical!

→ Repeat until threshold met

Why re-compute?

Removing samples **changes the statistical landscape**. Without re-scoring, you just shift which biases dominate.

Example: Removing all “count=2” questions might make “count=3” the new dominant pattern. Iterative re-scoring helps catch this.

Result: Prevents “*shifting bias under the rug*” — keeps the benchmark honest across multiple potential exploits.

Does Debiasing Actually Help?

Model Configuration	VSI-Bench (Original)			VSI-Bench-Debiased		
	Vis.	Blind	Δ_{V-B}	Vis.	Blind	Δ_{V-B}
LLaVA-Video 7B (Base)	36.7	25.9	10.8	31.3	20.3	11.0
+ VSI-Train-10k FT	57.1	44.7	12.4	48.7	32.0	16.6
<i>Increase in Δ due to FT</i>	20.4	18.8	1.6	17.4	11.7	5.6
<i>Chance (frequency)</i>		34.0			34.0	

Key comparison: FT improvement

Original: Vision +20.4, Blind +18.8 (almost equal!)

Debiased: Vision +17.4, Blind +11.7 (vision benefits more)

937 samples removed
→ 30.7% of 3,056 questions

Gap after FT:

12.4 → 16.6

+4.2 point improvement

✓ Validation

The debiased benchmark better isolates actual visual reasoning improvements.

Takeaways

- 1. Non-visual shortcuts are pervasive** — even in "careful" benchmarks
- 2. TST provides systematic diagnosis**
K-fold CV on text-only features → sample-level bias scores
- 3. IBP provides principled mitigation**
Iterative removal with score recomputation → validated by widening vision-blind gap
- 4. The meta-point:** If it CAN be gamed, it WILL be — audit proactively

 Hot Take

Benchmark Designers Should "Train on the Test Set"

...to expose and mitigate exploitable non-visual shortcuts.

Ellis Brown Jihan Yang Shusheng Yang Rob Fergus Saining Xie

New York University

Preprint, Nov. 2025

 arXiv

 PDF

 GitHub

 Thread

 VSI-Train-10k

 VSI-Bench (Debiased)

<https://vision-x-nyu.github.io/test-set-training>

III: Towards World Models

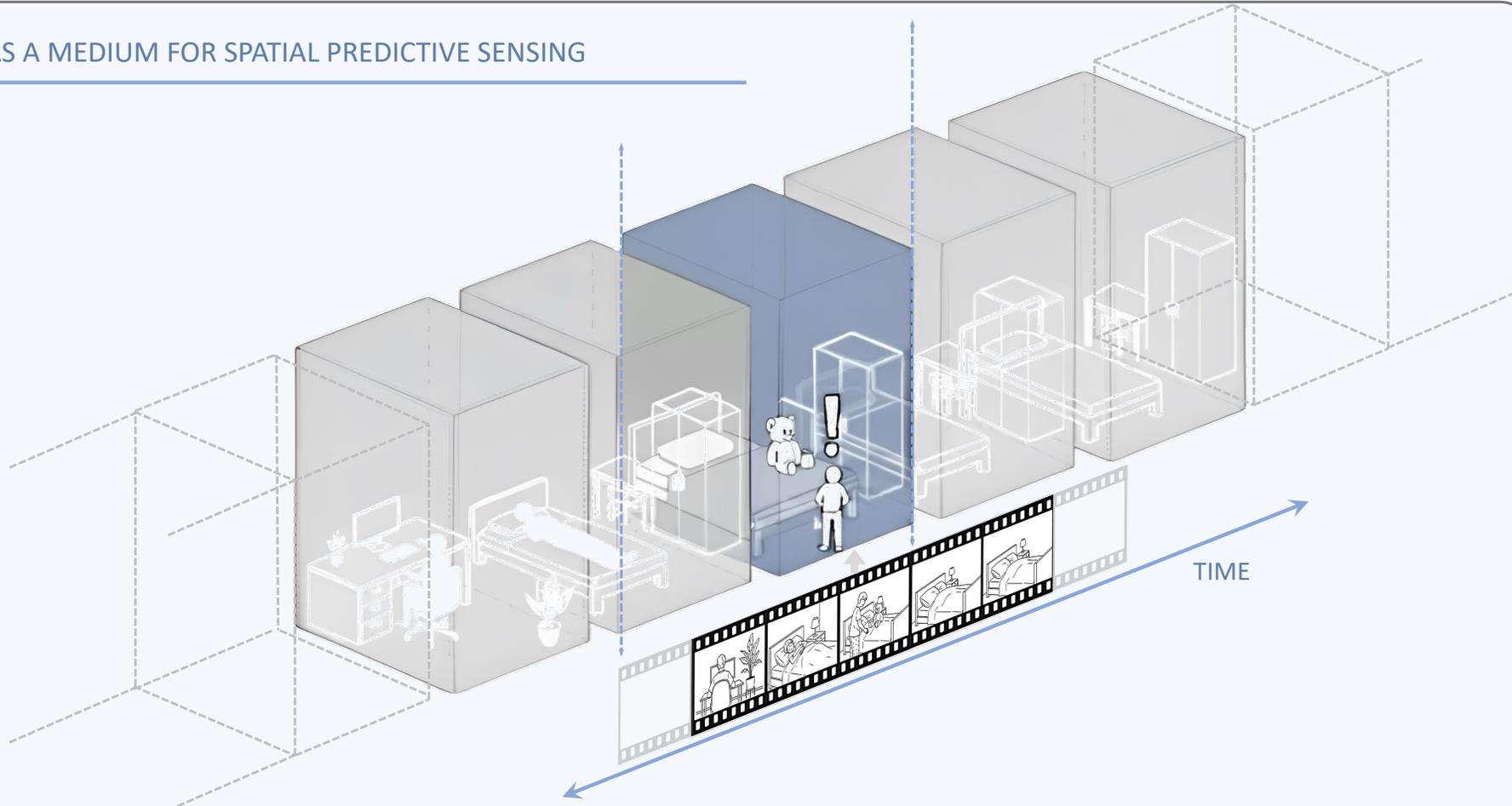
1. [Cambrian-S](#): Spatial “Super-Sensing”
2. [Unified Multimodal Models](#): Unified Multimodal Pretraining

Cambrian-S: Towards Spatial Supersensing in Video

Shusheng Yang*, Jihan Yang*, Pinzhi Huang[†], Ellis Brown[†],
Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan
Wang, Daohan Lu, Rob Fergus, Yann LeCun, Li Fei-Fei, Saining Xie

New York University & Stanford University

VIDEO AS A MEDIUM FOR SPATIAL PREDICTIVE SENSING



Supersensing in Video

Towards Spatial Supersensing in Video

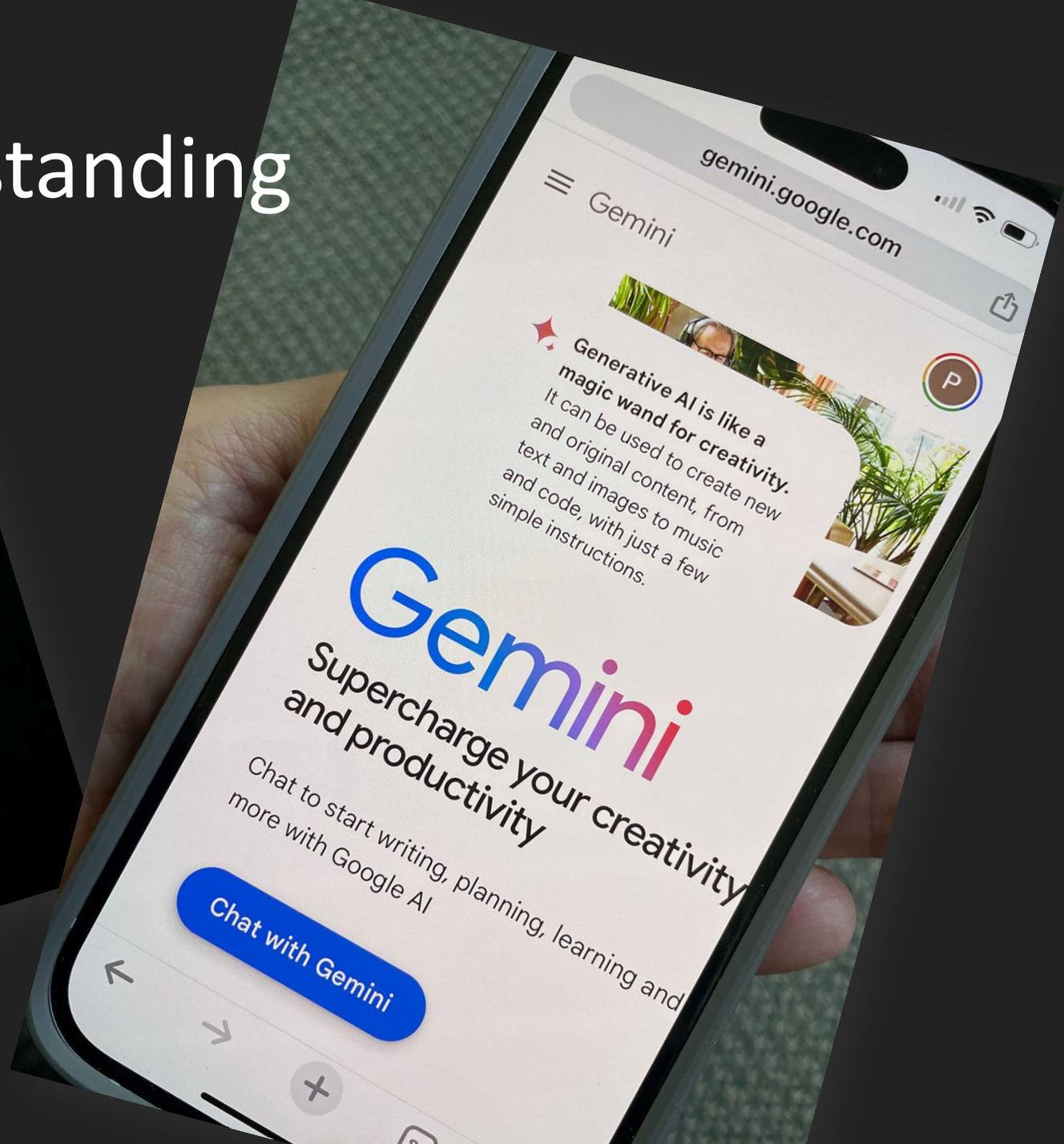
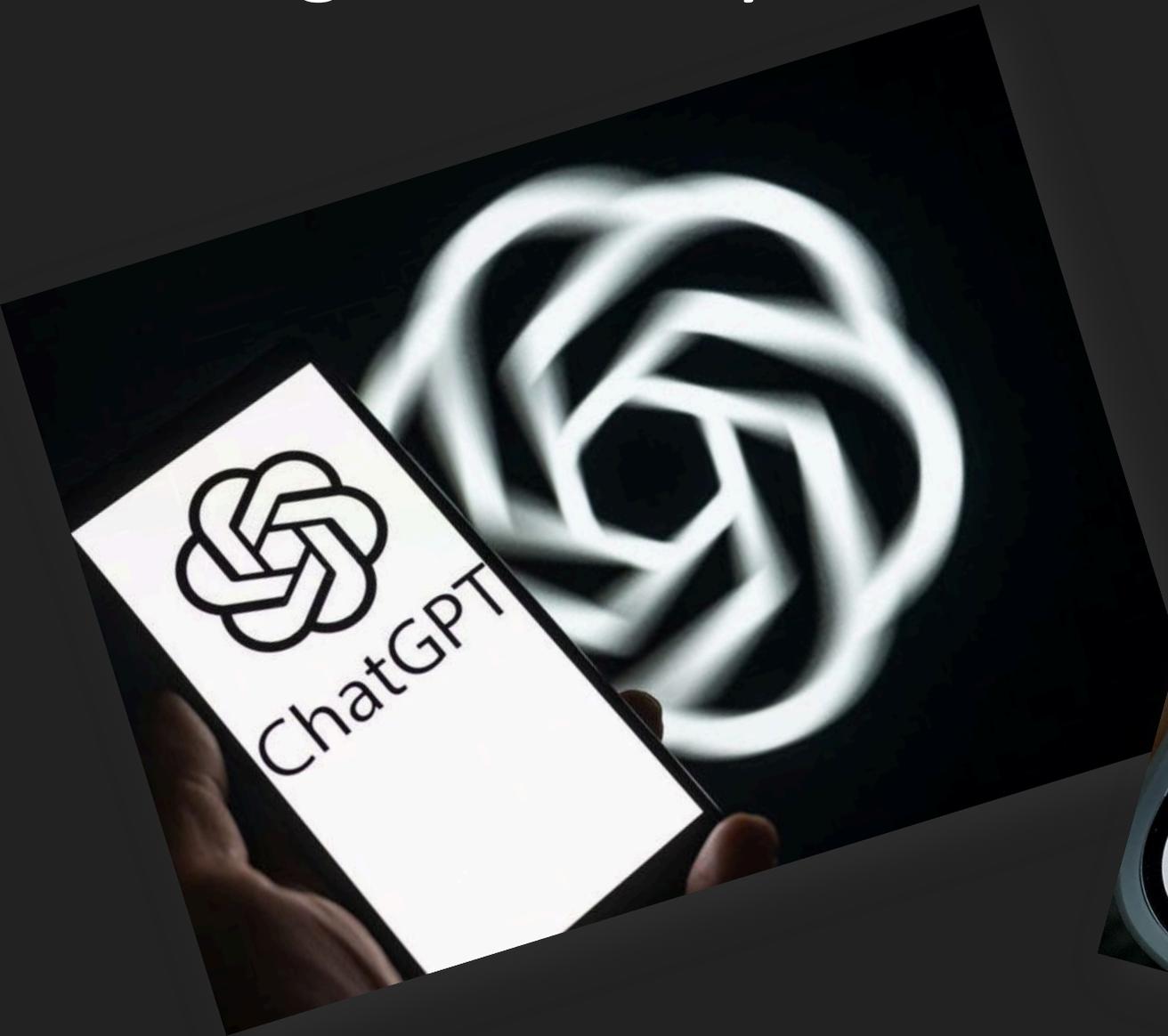


Linguistic-Only Understanding

*Knowledge recall;
no sensory modeling*

Linguistic-only understanding: no multimodal intelligence; reasoning is confined to text and symbols without sensory grounding. Current MLLMs have progressed beyond this stage, yet they still retain traces of its bias.

Linguistic-only understanding



Towards Spatial Supersensing in Video



Linguistic-Only Understanding

*Knowledge recall;
no sensory modeling*

Linguistic-only understanding: no multimodal intelligence; reasoning is confined to text and symbols without sensory grounding. Current MLLMs have progressed beyond this stage, yet they still retain traces of its bias.

Towards Spatial Supersensing in Video



Linguistic-Only Understanding

*Knowledge recall;
no sensory modeling*



Semantic Perception

*Naming and describing things
for user prompts*

Semantic perception: parsing pixels into objects, attributes, and relations. This corresponds to the strong “show and tell” capabilities present in MLLMs.

Semantic Perception

Autumnal Country Store Scene

System Instructions
Optional tone and style instructions for the model

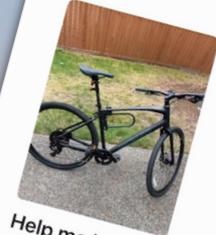


User

describe this image

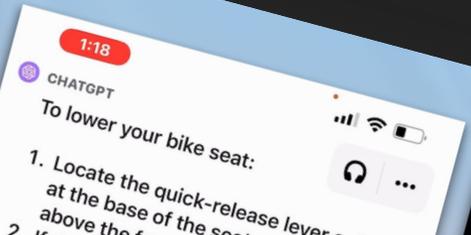
Model

This image captures a charming autumn scene at a country store. The store, named "1836 Country Store" as indicated by a display window, is situated amidst a courtyard paved with flagstones, and several stuffed animals are arranged to depict a variety of animals, including bears, tigers, giraffes, and dogs. Several people are relaxing in chairs around the stuffed animals, suggesting a leisurely atmosphere. A small table with baskets and a sale. A flock of Canada geese strolls through the courtyard, adding to the rustic ambiance. The trees surrounding the store are in various stages of autumn foliage, displaying vibrant hues of orange, red, and yellow. Fallen leaves litter the ground, and the lighting is soft and diffused, possibly suggesting an overcast day or late afternoon light. The image evokes a sense of peace and tranquility, characteristic of a country store scene.



RYAN
Help me lower my bike seat

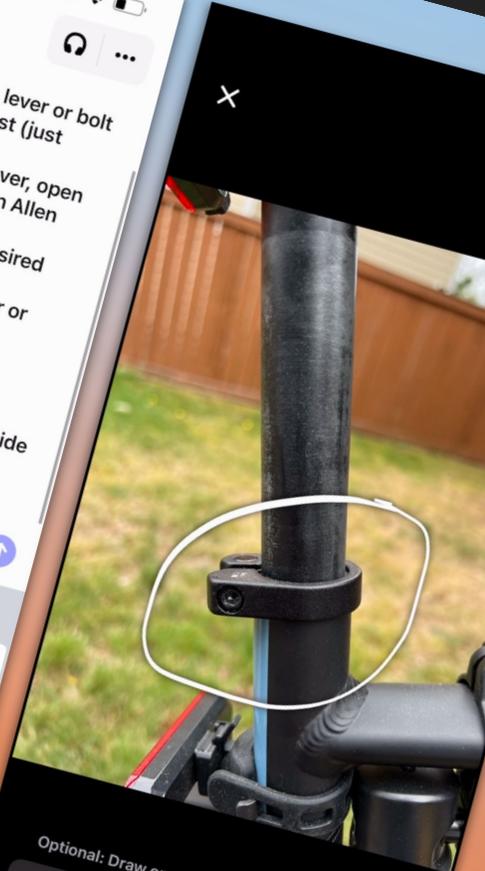
CHATGPT



CHATGPT
To lower your bike seat:

1. Locate the quick-release lever or bolt at the base of the seat post (just above the frame).
2. If there's a quick-release lever, open it. If it's a bolt, you'll need an Allen wrench.
3. Slide the seat down to the desired height.
4. Tighten the quick-release lever or bolt to secure the seat in place.
5. Ensure the seat is straight and doesn't wobble before riding.

If you have tools, show me, and I'll guide you further.



Optional: Draw on your image to focus ChatGPT

Retake

Towards Spatial Supersensing in Video



Linguistic-Only Understanding

*Knowledge recall;
no sensory modeling*



Semantic Perception

*Naming and describing things
for user prompts*

Semantic perception: parsing pixels into objects, attributes, and relations. This corresponds to the strong “show and tell” capabilities present in MLLMs.

Towards Spatial Supersensing in Video



Linguistic-Only Understanding

*Knowledge recall;
no sensory modeling*



Semantic Perception

*Naming and describing things
for user prompts*



Streaming Event Cognition

*Always-on sensing for open-
ended streams; memory
across time; proactive
answering*

Streaming event cognition: processing live, unbounded streams while proactively interpreting and responding to ongoing events. This aligns with efforts to make MLLMs real-time assistants.

Streaming event cognition



Towards Spatial Supersensing in Video



Linguistic-Only Understanding

*Knowledge recall;
no sensory modeling*



Semantic Perception

*Naming and describing things
for user prompts*



Streaming Event Cognition

*Always-on sensing for open-
ended streams; memory
across time; proactive
answering*

Streaming event cognition: processing live, unbounded streams while proactively interpreting and responding to ongoing events. This aligns with efforts to make MLLMs real-time assistants.

Towards Spatial Supersensing in Video



Linguistic-Only Understanding

*Knowledge recall;
no sensory modeling*



Semantic Perception

*Naming and describing things
for user prompts*



Streaming Event Cognition

*Always-on sensing for open-
ended streams; memory
across time; proactive
answering*

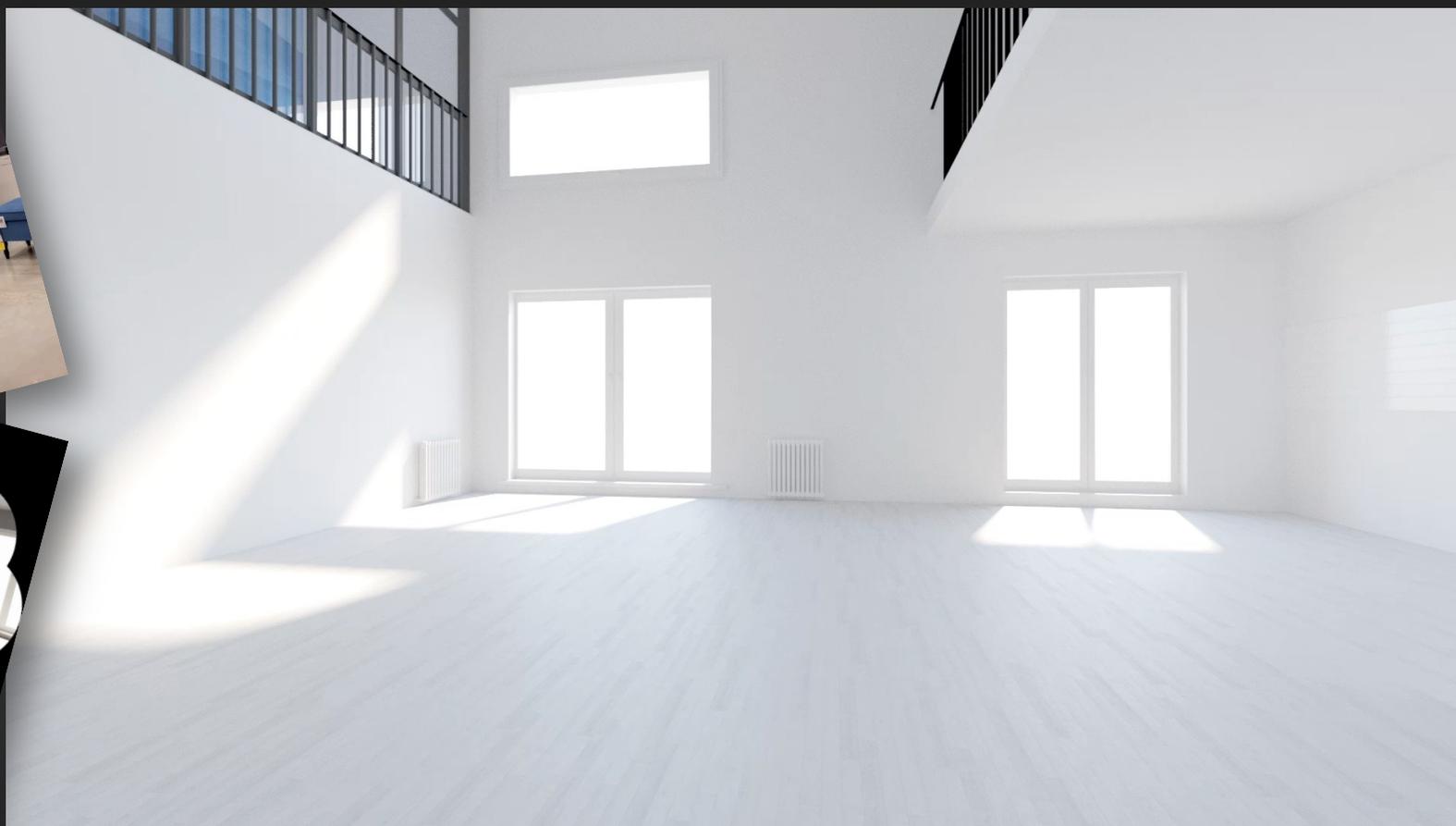


Spatial Cognition

*Seeing the world behind the
video; implicit 3D*

Implicit 3D spatial cognition: understanding video as projections of a 3D world. Agents must know what is present, where, how things relate, and how configurations change over time. Today's video models remain limited here.

Implicit 3D spatial cognition



Towards Spatial Supersensing in Video



Linguistic-Only Understanding

*Knowledge recall;
no sensory modeling*



Semantic Perception

*Naming and describing things
for user prompts*



Streaming Event Cognition

*Always-on sensing for open-
ended streams; memory
across time; proactive
answering*



Spatial Cognition

*Seeing the world behind the
video; implicit 3D*

Implicit 3D spatial cognition: understanding video as projections of a 3D world. Agents must know what is present, where, how things relate, and how configurations change over time. Today's video models remain limited here.

Towards Spatial Supersensing in Video



Linguistic-Only Understanding

*Knowledge recall;
no sensory modeling*



Semantic Perception

*Naming and describing things
for user prompts*



Streaming Event Cognition

*Always-on sensing for open-
ended streams; memory
across time; proactive
answering*



Spatial Cognition

*Seeing the world behind the
video; implicit 3D*



Predictive World Modeling

*Unconscious inference;
Predictive, selective, and self-
updating world model*

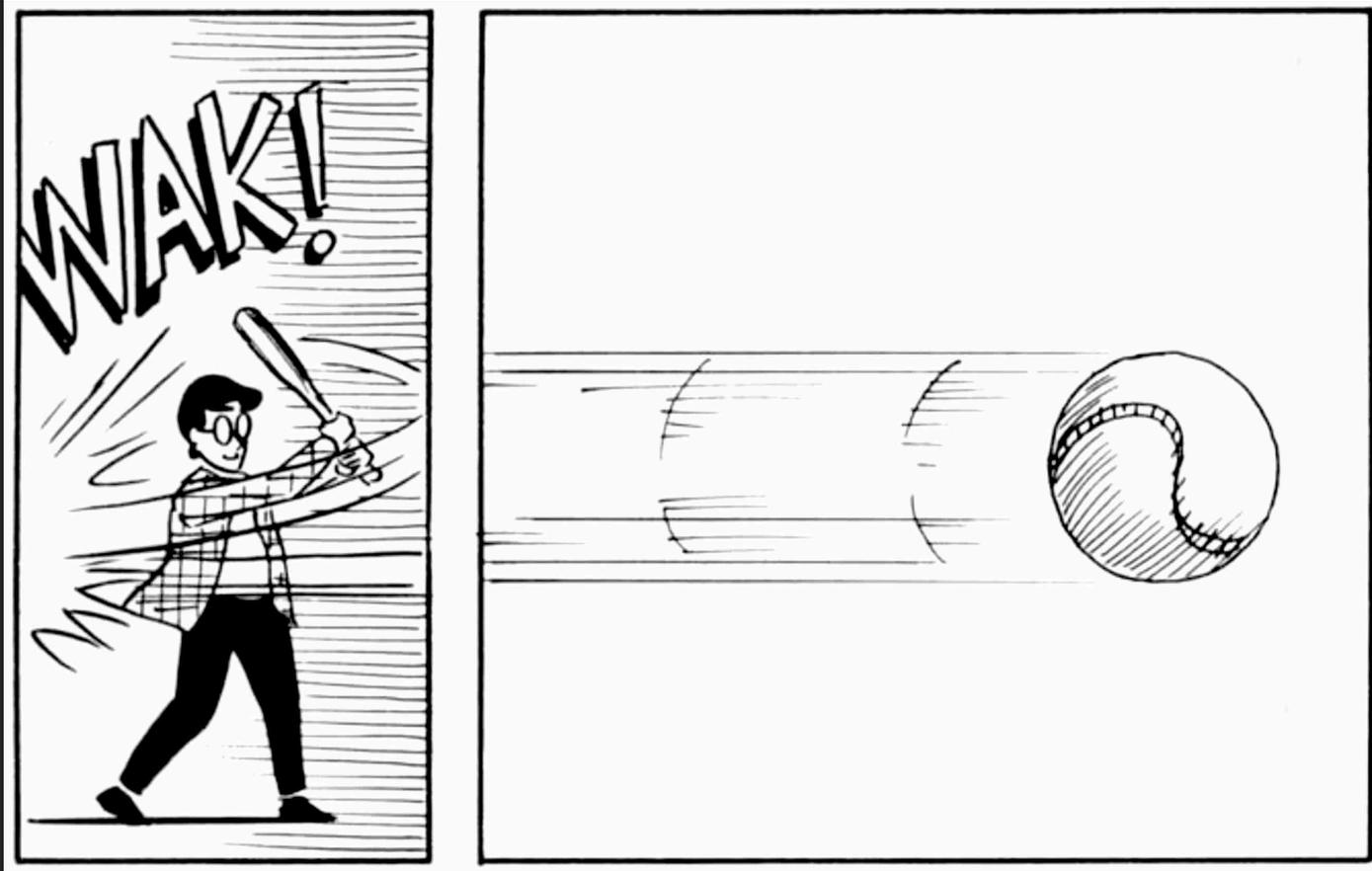
Predictive world modeling: anticipating future states with an internal model that uses expectation and surprise to organize perception for memory and decision making. This process mirrors human "unconscious inference" and is largely absent in current systems.

Predictive World Modeling

- How human's brain works so efficiently but so powerful on processing life-long visual streams?

Predictive world modeling (Predictive mind or unconscious inference) of human brain helps us allocate our attention (computation) and memory (storage)

Predictive World Modeling



“The reason we are able to hit a 100mph fastball is due to our ability to instinctively predict when and where the ball will go. For professional players, this all happens subconsciously.

...

They can quickly act on their predictions of the future without the need to consciously roll out possible future scenarios to form a plan.”

Towards Spatial Supersensing in Video



Linguistic-Only Understanding

*Knowledge recall;
no sensory modeling*



Semantic Perception

*Naming and describing things
for user prompts*



Streaming Event Cognition

*Always-on sensing for open-
ended streams; memory
across time; proactive
answering*



Spatial Cognition

*Seeing the world behind the
video; implicit 3D*



Predictive World Modeling

*Unconscious inference;
Predictive, selective, and self-
updating world model*

TASK-DRIVEN

WORLD MODELING

Our hierarchy of supersensing in video

Evaluation is NOT READY!

Data is not READY!

Architecture is not READY!

Evaluation is NOT READY!

Data is not READY!

Architecture is not READY!

*How can we benchmark spatial
supersensing?*

Are existing benchmarks enough?

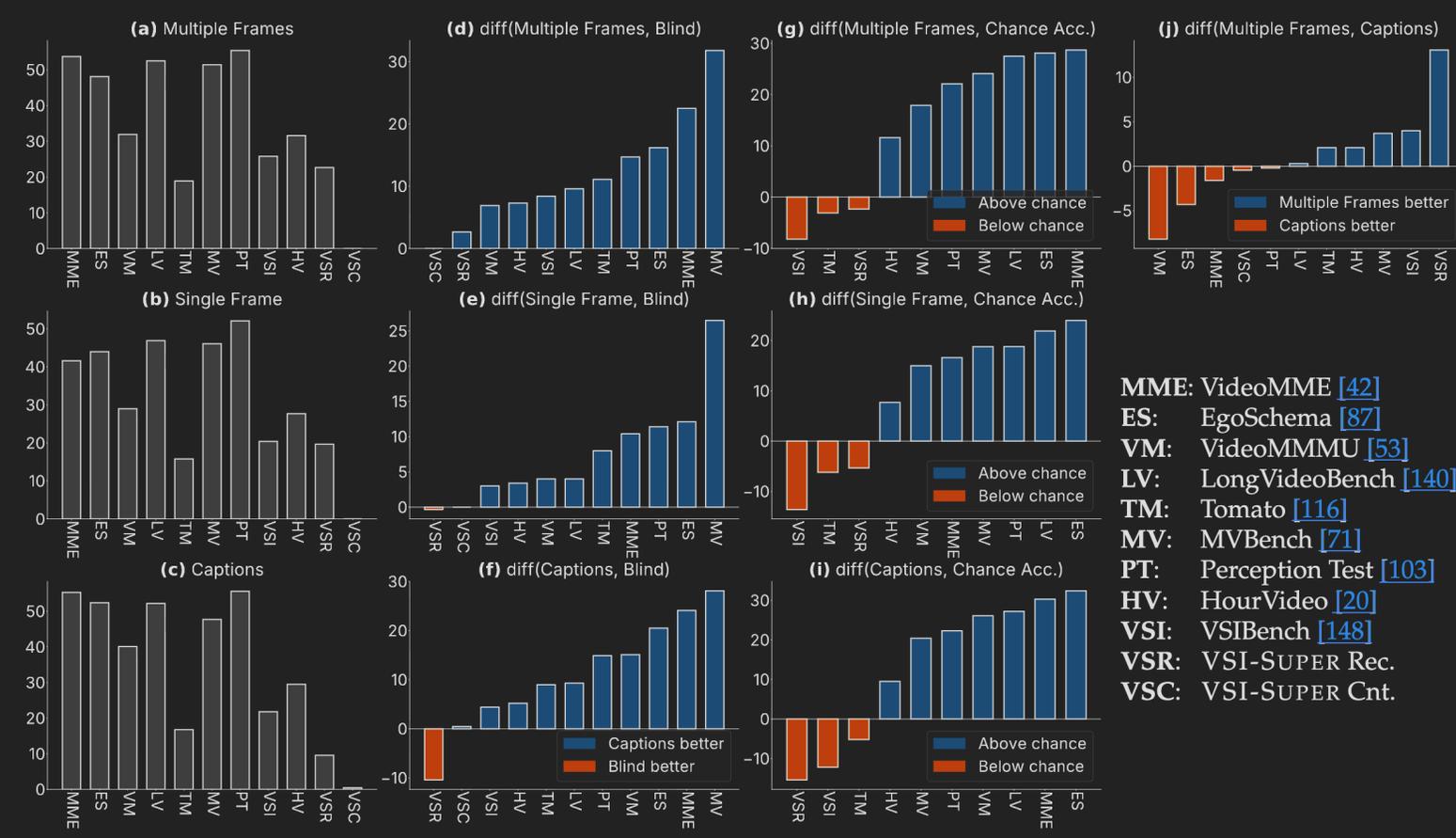
Deconstructing Existing Video Benchmarks

- **Multiple Frames:** Model processes 32 uniformly sampled frames from each video clip — standard video representation method.
- **Single Frame:** Model uses only the middle frame of the clip to test performance with minimal visual context.
- **Frame Captions:** Model receives captions for the same 32 sampled frames (no visual input) to test task solvability without perceptual grounding. Captions generated using the Gemini-2.0-Flash API.

Deconstructing Existing Video Benchmarks

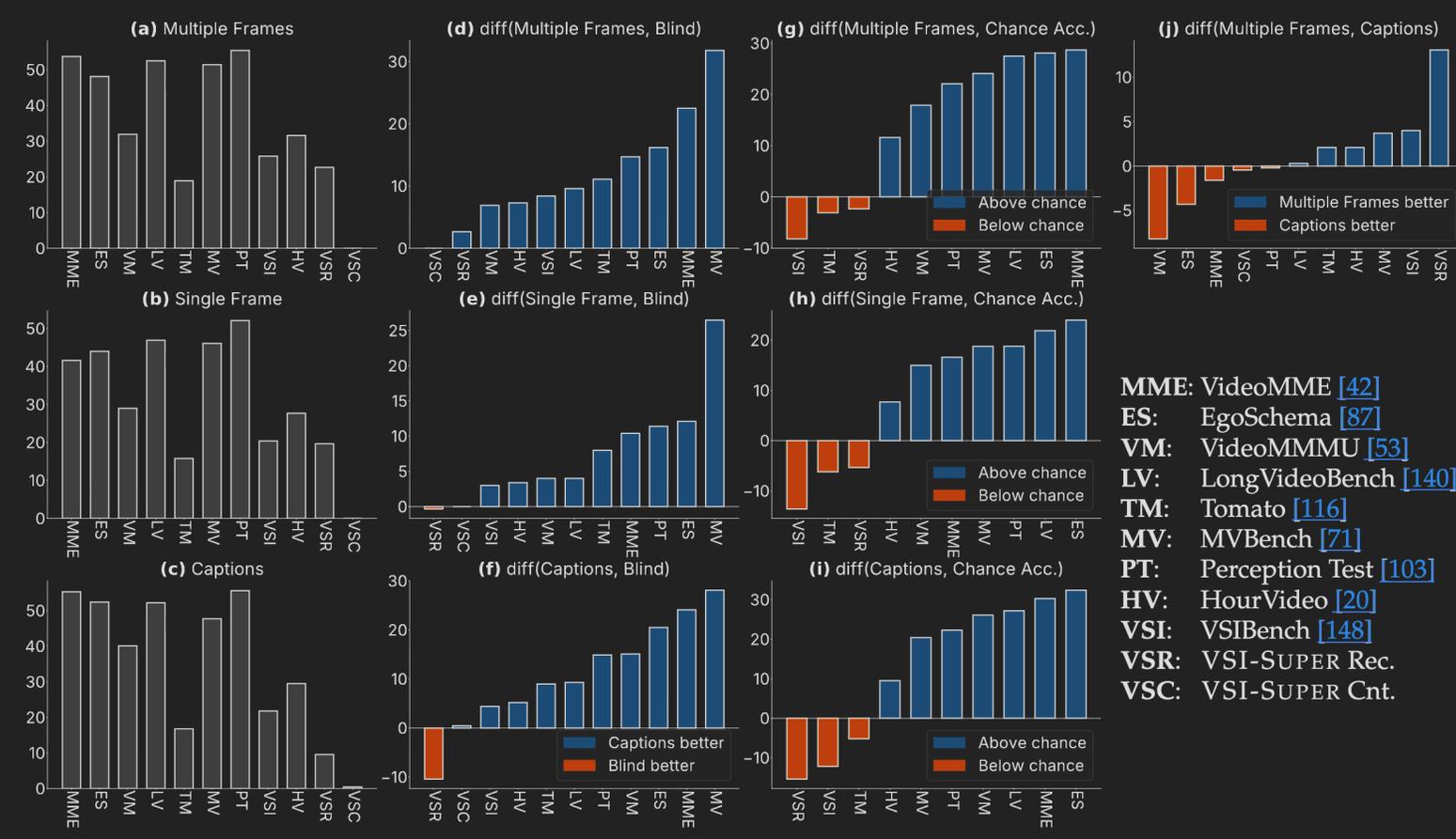
- *Blind: No visual inputs.*
- *Chance: Random choice.*

Current benchmarks are NOT READY



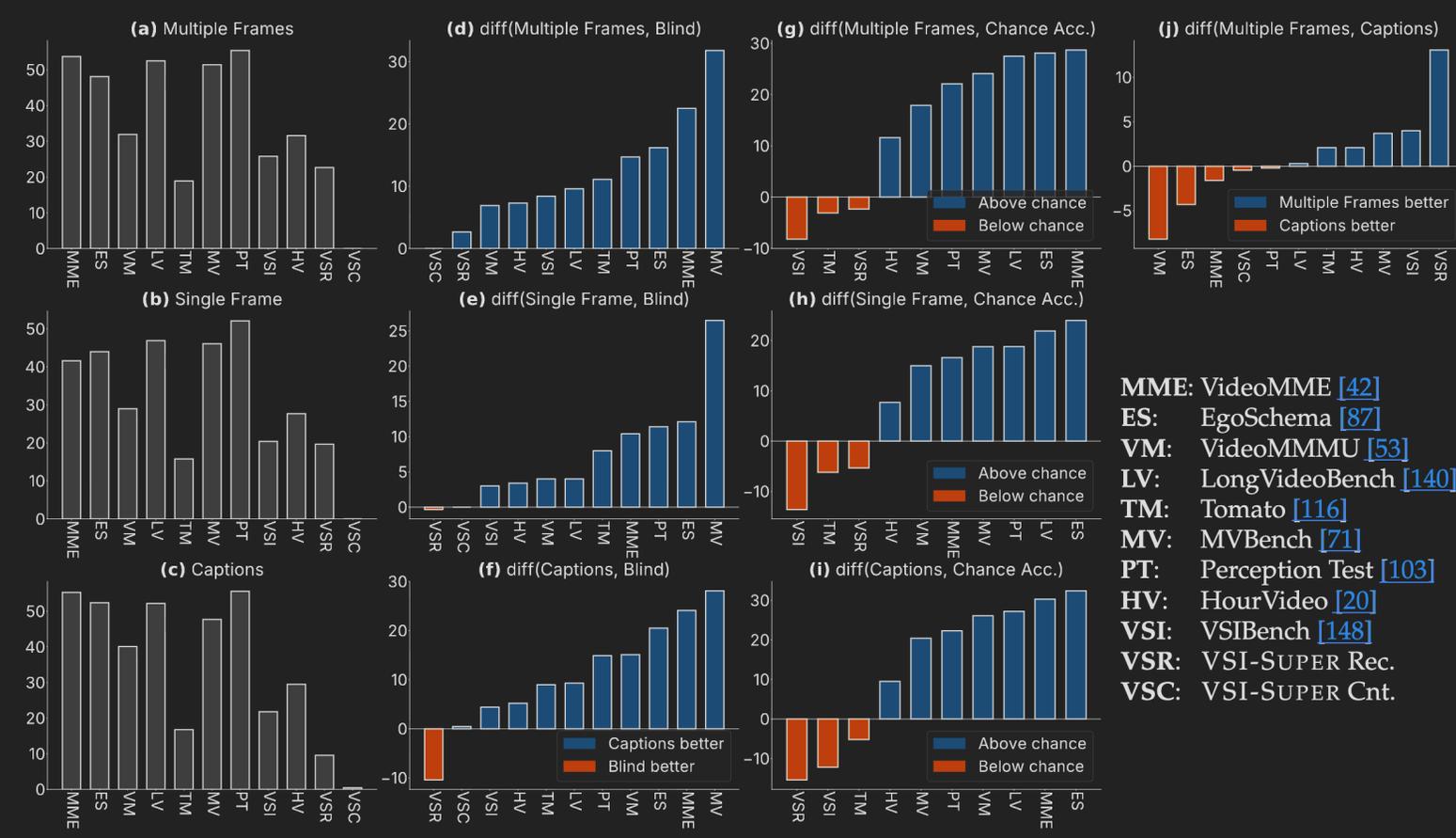
Single frame / caption perform well on many benchmarks.

Current benchmarks are NOT READY



Most benchmark focus on semantic and knowledge recall

Current benchmarks are NOT READY



Only a few benchmarks probe spatial sensing

Current benchmarks are NOT READY



Evaluation is NOT READY!

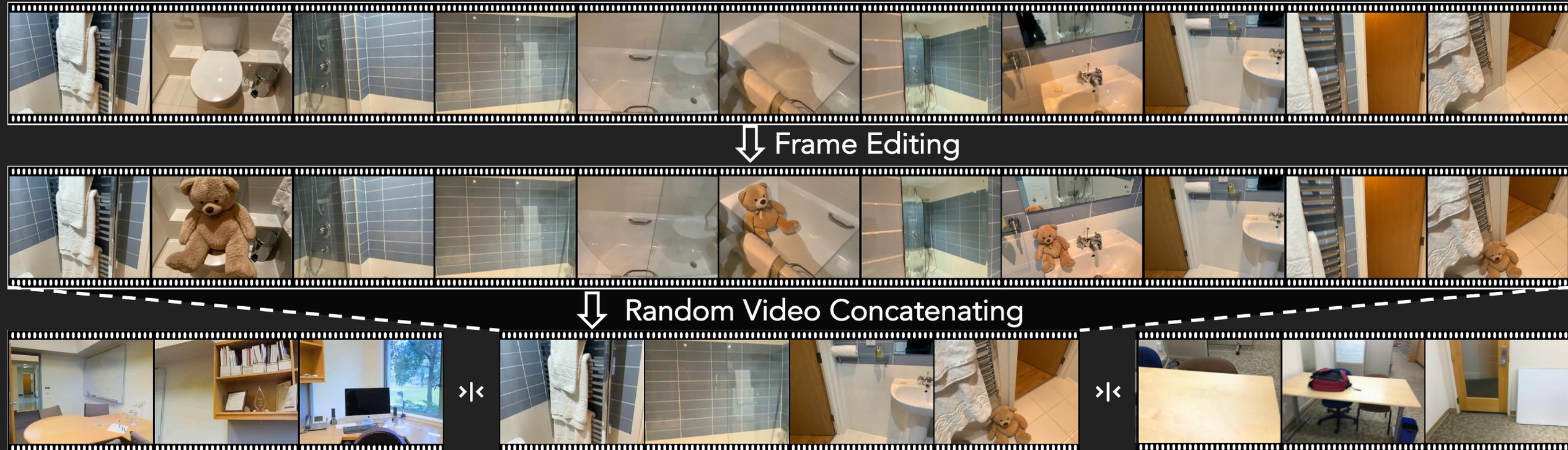
VSI-SUPER

A two-part, long-horizon benchmark for “supersensing”

- Combines concatenated video sequences with online Q&A
- Like Needle-in-a-Haystack tasks but more realistic and contextually grounded
- Designed to be resistant to brute-force context expansion, emphasizing true spatial sensing

VSI-SUPER Recall

Long-horizon spatial observation and recall



Which of the following correctly represents the order in which the Teddy Bear appeared in the video?

- A. Toilet, Bathtub, Sink, Floor
- C. Toilet, Sink, Floor, Bathtub

- B. Bathtub, Toilet, Sink, Floor
- D. Floor, Toilet, Bathtub, Sink

VSI-SUPER Count

Continual counting under changing viewpoints and scenes

Num. of Chairs: 3 1 16



Q: How many different chair(s) are there in this video? A: 20

Num. of Chairs: 3 1 16



Streaming Questions:

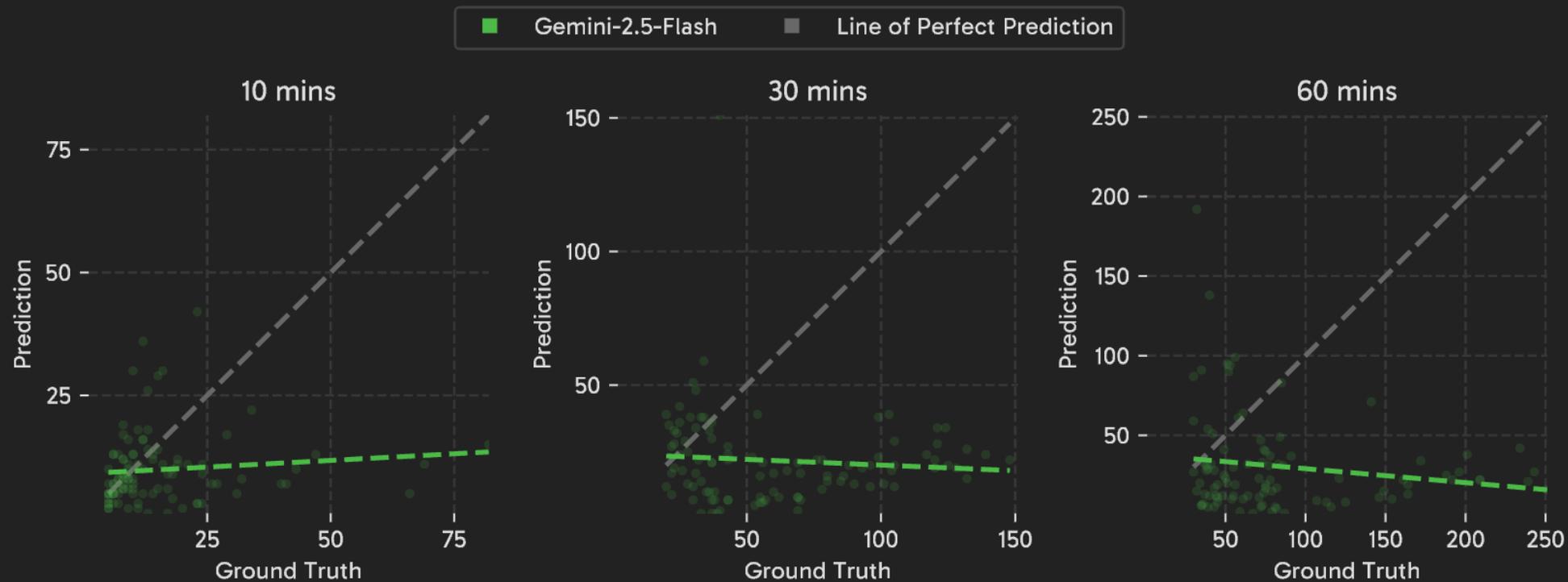
Q: How many different chair(s) are there in this video? A: 2 A: 3 A: 3 A: 4 A: 20

Gemini-2.5 on VSI-SUPER

Model	VideoMME[42]	VideoMMMU[53]	VSI-Bench[148]	VSR		VSC	
				60 min	120 min	60 min	120 min
Gemini-2.5-Flash	81.5	79.2	45.7	41.5	Out of Ctx.	10.9	Out of Ctx.

Frontier models can perform pretty good on semantic or knowledge focused video benchmarks but fail at spatial sensing and supersensing tasks.

Gemini-2.5 on VSI-SUPER



Gemini struggle with counting; scaling data and context length alone fails to improve generalization.

Evaluation is NOT READY!

Data is not READY!

Architecture is not READY!

LLM Data: Blessing of the Internet

Common
Crawl

Colossal Clean
Crawled Corpus (C4)

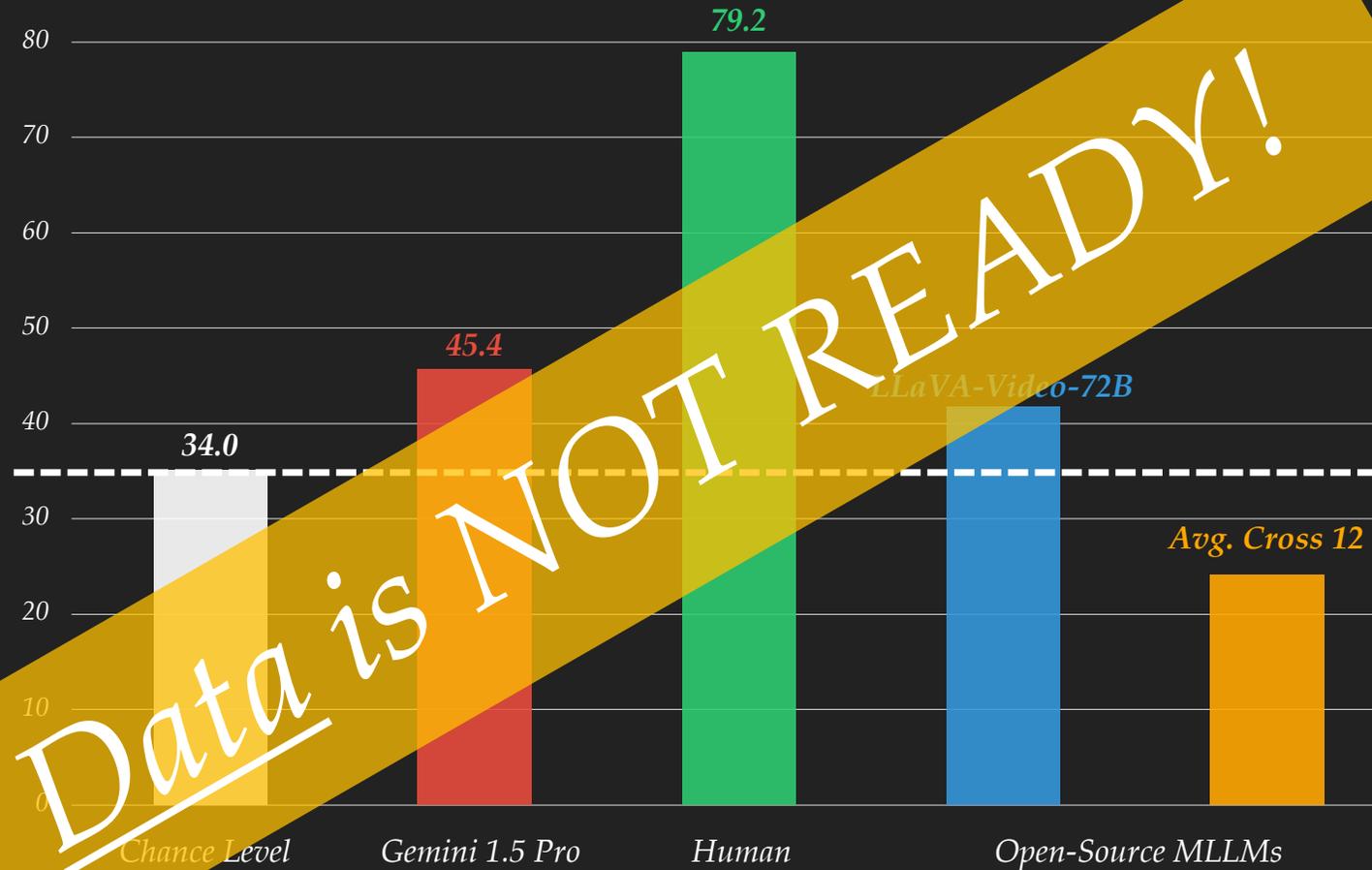


arXiv

StackExchange



Frontier models fail at spatial sensing



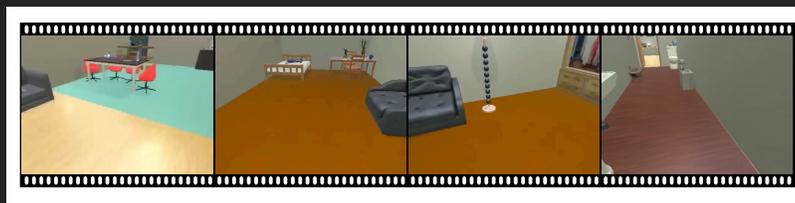
Models' performance on VSI-Bench

VSI-590K

Is spatial sensing a data problem?



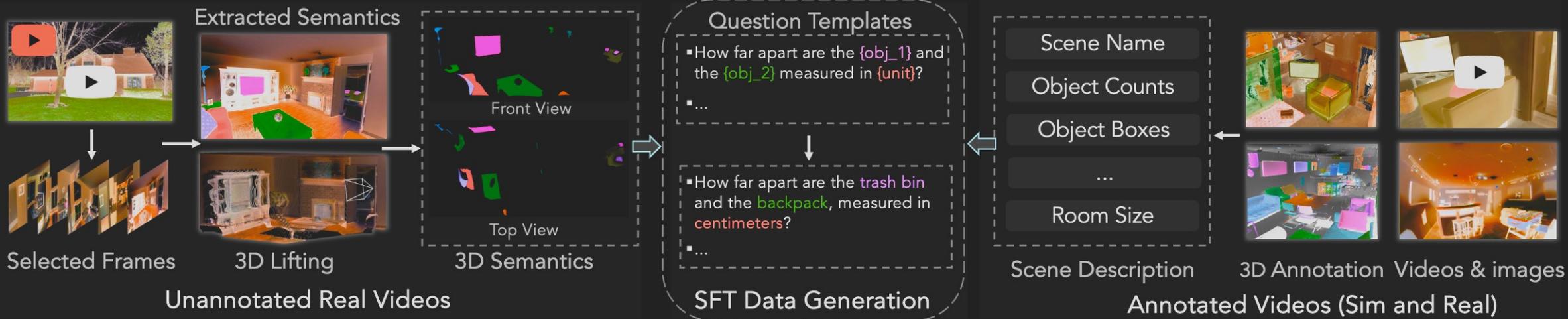
Unannotated web videos



Synthetic Videos



Annotated Real videos



VSI-590K

Is spatial sensing a data problem?

Dataset	# Videos	# Images	# QA Pairs
<i>Annotated Real Videos</i>			
S3DIS [4]	199	-	5,187
Aria Digital Twin [102]	183	-	60,207
ScanNet [33]	1,201	-	92,145
ScanNet++ V2 [153]	856	-	138,701
ARKitScenes [12]	2,899	-	57,816
<i>Simulated Data</i>			
ProcTHOR [36]	625	-	20,092
Hypersim [113]	-	5,113	176,774
<i>Unannotated Real Videos</i>			
YouTube Room Tour	-	20,100	20,100
Open X-Embodiment [100]	-	14,801	14,801
AgiBot-World [16]	-	4,844	4,844
Total	5,963	44,858	590,667

VSI-590K video sources

VSI-590K

Is spatial sensing a data problem?

Types	Rel./Abs.	Perspective	Modality	Group	Example template
Size	Rel.	—	V & I	Configuration	"Between $\{object_1\}$ and $\{object_2\}$, which is larger?"
	Abs.	—	V & I	Measurement	"What is the height of the $\{object\}$ in $\{unit\}$?"
	Abs.	—	V & I	Measurement	"What is the room's size in $\{unit\}$?"
Direction	Rel.	Camera	I	Configuration	"From the camera's perspective, is the $\{object\}$ on the left or the right?"
	Rel.	Object	V & I	Configuration	"Facing the $\{object_1\}$ from the $\{object_2\}$, would the $\{object_3\}$ be placed left, right, or back?"
	Abs.	Object	V & I	Measurement	"Standing at $\{object_1\}$, facing toward $\{object_2\}$, how far clockwise do I rotate (in degrees) to see the $\{object_3\}$?"
Count	Rel.	—	V & I	Configuration	"Are there fewer $\{object_1\}$ than $\{object_2\}$?"
	Abs.	—	V & I	Measurement	"How many $\{object\}$ are present?"
Distance	Rel.	Camera	I	Configuration	"Which object is closer to the camera, the $\{object_1\}$ or the $\{object_2\}$?"
	Rel.	Object	V & I	Configuration	"Which is nearer to the $\{object_3\}$, the $\{object_1\}$ or the $\{object_2\}$?"
	Abs.	Object	V & I	Measurement	"What is the distance between the $\{object_1\}$ and the $\{object_2\}$ in $\{unit\}$?"
Appr. Order	—	—	V	Spatiotemporal	"Determine how $\{object_1\}$, $\{object_2\}$, $\{object_3\}$, and $\{object_4\}$ are ordered by their initial appearances in the video"

VSI-590K Taxonomy

VSI-590K

Is spatial sensing a data problem?

Real and synthetic data together provide rich sources that boost spatial understanding

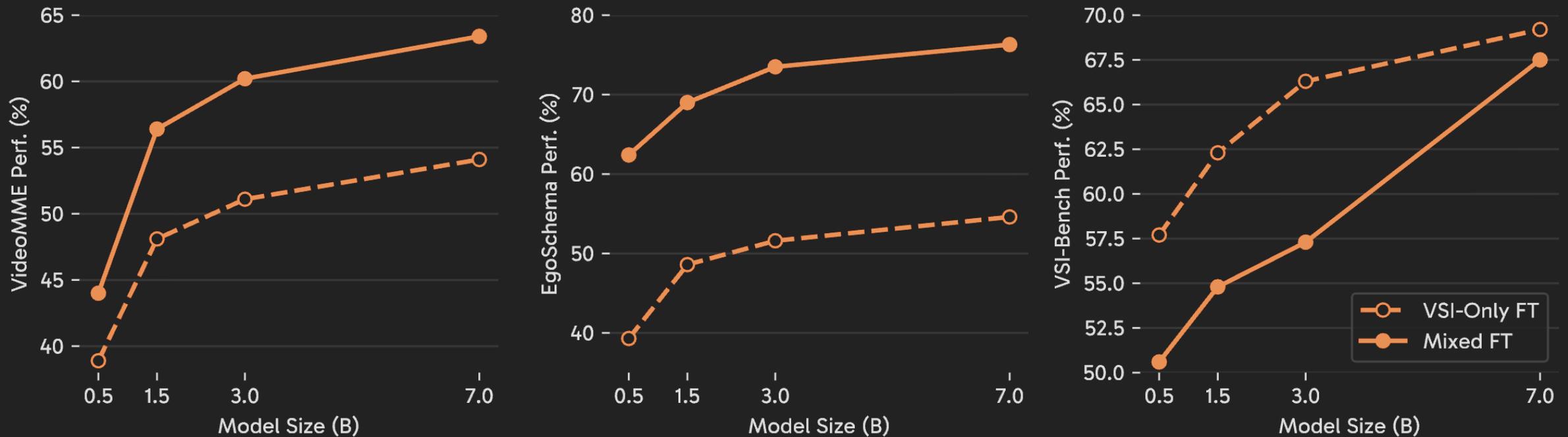
	Image			VSI-Bench (Video)								
	MMVP	3DSR	CV-B	Avg	Obj Ct	Abs Dst	Obj Sz	Rm Sz	Rel Dst	Rel Dir	Rte Pln	Ap Ord
<i>VSI Data Mixture</i>												
Baseline	52.7	54.5	73.5	28.5	18.9	20.0	36.0	22.2	42.9	31.3	21.6	33.0
+ S3DIS	54.0	54.9	75.3	41.6	63.8	21.0	44.9	37.0	43.8	47.4	34.0	41.1
+ ADT	59.8	56.5	77.5	41.0	51.9	19.9	51.5	49.2	42.0	59.3	34.0	39.8
+ ARKitScenes	50.0	56.7	77.3	51.0	70.2	32.7	64.5	60.0	55.1	45.2	37.1	43.5
+ ScanNet	54.7	57.7	77.5	56.3	70.9	37.9	67.5	59.3	57.0	46.7	35.1	76.1
+ ScanNet++ V2	52.7	57.3	77.5	56.3	72.5	40.7	65.7	56.9	59.7	47.1	31.4	76.2
<i>Simulated Videos</i>												
+ ProcThor	53.3	55.7	74.9	36.4	21.0	29.7	49.3	3.8	52.3	45.7	30.4	58.7
+ HyperSim	52.0	56.0	79.7	45.6	67.8	32.0	59.3	36.4	53.2	47.0	32.5	36.6
<i>Pseudo-Annotated Images</i>												
+ YTB RoomTour	55.3	52.6	75.0	32.5	43.4	25.8	24.2	27.3	38.7	31.4	28.4	40.9
+ OXE & AGIBot	56.0	54.4	72.5	30.6	40.3	23.1	27.9	26.6	38.0	22.8	32.0	33.8
Full Mix	54.7	54.0	77.9	63.2	73.5	49.4	71.4	70.1	66.9	61.5	36.6	76.6

Pretraining is important

Model	VSI-Bench	VideoMME	EgoSchema	Perception Test
Different Base Models				
A1 (<i>w/o.</i> I-IT, <i>i.e.</i> QwenLM)	21.4	44.2	42.9	44.5
A2 (A1 + I-IT, <i>i.e.</i> Cambrian-1)	25.8	53.7	48.1	55.4
A3 (A2 + V-IT, 429K data)	28.9	61.2	50.3	66.3
A4 (A2 + V-IT, 3M data)	35.7	62.6	77.0	70.9
SFT <i>w/.</i> VSI-590K				
from A1	57.2	49.3	38.7	52.3
from A2	66.8	46.7	47.2	52.3
from A3	69.8	52.3	49.4	59.9
from A4	69.2	54.1	55.2	59.2
SFT <i>w/.</i> VSI-590K & general V-IT data mixture				
from A1	61.3	60.5	52.8	65.0
from A2	63.2	62.6	52.9	65.6
from A3	64.0	61.0	54.9	66.8
from A4	65.1	61.9	77.3	71.2

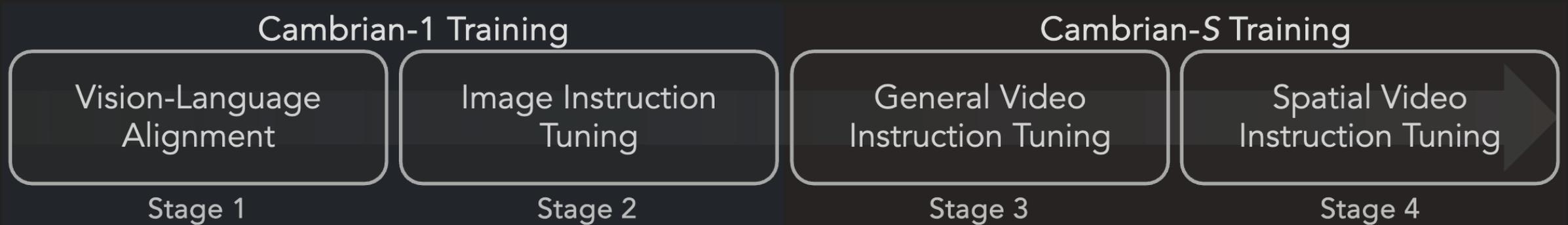
The quality of ***multimodal*** pre-training strongly influences post-training effectiveness.

Mixed-data training helps preserve ability



Mixed-data training helps preserve general video understanding capability

Cambrian-S: Spatially-grounded MLLMs



Alignment → Pure Image IT → General Video IT →
Spatial Video IT

Cambrian-S: Spatially-grounded MLLMs

Model	Base LM	Video										Image		
		VSI-Bench	VSI-Bench ^{Debiased}	Tomato	HourVideo	Video ^{MME}	EgoSchema	Video ^{MMU}	LongVBench	MVBench	Percept. Test	MMVP	3DSR	CV-Bench
<i>Proprietary Models</i>														
Claude-3.5-sonnet	UNK.	-	-	27.8	-	62.9	-	65.8	-	-	-	-	48.2	-
GPT-4o	UNK.	34.0	-	37.7	37.2	71.9	-	61.2	66.7	-	-	66.0	44.2	-
Gemini-1.5-Pro	UNK.	45.4	40.1	36.1	37.3	75.0	72.2	53.9	64.0	-	-	-	-	-
Gemini-2.5 Pro	UNK.	51.5	49.1	-	-	-	-	83.6	67.4	-	-	51.3	-	-
<i>Open-Source Models</i>														
LLaVA-Video-7B	Qwen2-7B	35.6	30.7	22.5	28.6	63.3	57.3	36.1	58.2	58.6	67.9	-	-	75.7
LLaVA-One-Vision-7B	Qwen2-7B	32.4	28.5	25.5	28.3	58.2	60.1	33.9	56.4	56.7	57.1	54.7	-	74.3
Qwen-VL-2.5-7B	Qwen2.5-7B	33.5	29.6	-	-	65.1	65.0	47.4	56.0	69.6	-	56.7	48.4	-
InternVL2.5-8B	InternLM2.5-7B	34.6	24.9	-	-	64.2	50.6	-	60.0	72.0	-	55.3	50.9	-
InternVL3.5-8B	Qwen3-8B	56.3	49.7	-	-	66.0	61.2	49.0	62.1	72.1	-	56.0	-	-
Cambrian-S-7B	Qwen2.5-7B	67.5	59.9	27.0	36.5	63.4	76.8	38.6	59.4	64.5	69.9	60.0	54.8	76.9

Competitive on general image / video benchmark

Cambrian-S: Spatially-grounded MLLMs

Methods	Avg.	<i>Obj. Count</i>	<i>Abs. Dist.</i>	<i>Obj. Size</i>	<i>Room Size</i>	<i>Rel. Dist.</i>	<i>Rel. Dir.</i>	<i>Route Plan</i>	<i>Appr. Order</i>
		Numerical Answer				Multiple-Choice Answer			
<i>Statistics</i>									
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
<i>Proprietary Models (API)</i>									
GPT-4o	34.0	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5
Gemini-1.5 Flash	42.1	49.8	30.8	53.5	54.4	37.7	41.0	31.5	37.8
Gemini-1.5 Pro	45.4	56.2	30.9	64.1	43.6	51.3	46.3	36.0	34.6
Gemini-2.5 Pro	51.5	43.8	34.9	64.3	42.8	61.1	47.8	45.9	71.3
<i>Open-source Models</i>									
Cambrian-S-7B	67.5	73.2	50.5	74.9	72.2	71.1	76.2	41.8	80.1
Cambrian-S-3B	57.3	70.7	40.6	68.0	46.3	64.8	61.9	27.3	78.8
Cambrian-S-1.5B	54.8	68.4	40.0	61.5	50.1	62.4	48.9	29.9	77.5
Cambrian-S-0.5B	50.6	67.9	35.4	52.2	52.5	52.3	46.5	25.8	72.2

Leading performance on spatial sensing (VSI-Bench)

Cambrian-S fail at VSI-SUPER

Eval Setup	VSR					VSC			
	10 min	30 min	60 min	120 min	240 min	10 mins	30 min	60 min	120 min
Uni. Sampling, 128F	26.7	21.7	23.3	30.0	28.2	16.0	0.0	0.0	0.0
FPS Sampling, 1FPS	38.3	35.0	6.0	0.0	0.0	0.6	0.0	0.0	0.0

Scaling up data benefits spatial sensing, but only scaling data cannot solve supersensing!

Evaluation is NOT READY!

Data is not READY!

Architecture is not READY!



Which of the following correctly represents the order in which the Hello Kitty appeared in the video?

- A. Nightstand, Bed, Crib, Blue bench
- B. Blue bench, Crib, Nightstand, Bed
- C. Bed, Nightstand, Blue bench, Crib
- D. Blue bench, Bed, Crib, Nightstand

Num. of Chairs: 3



16



Q: How many different chair(s) are there in this video?

A: 20

VSI-SUPER Recall

VSI-SUPER Count

Frontier Models



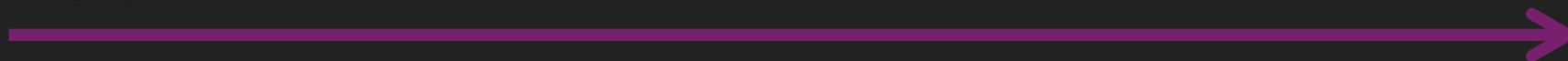
Video Duration

10mins

30mins

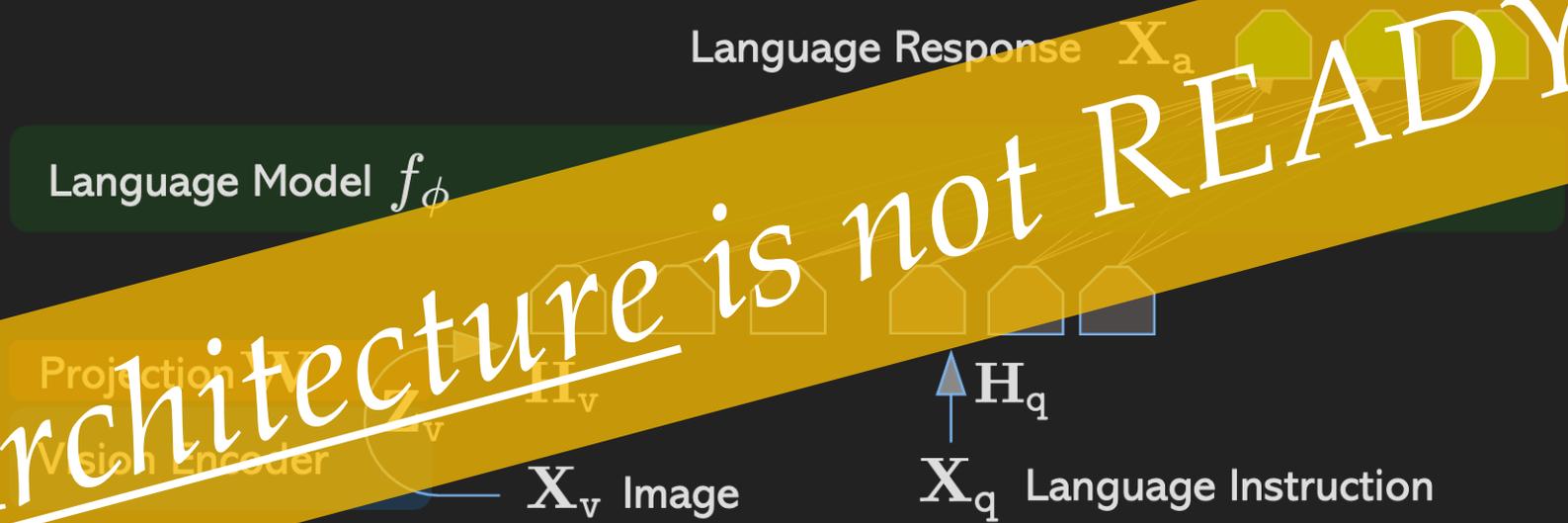
60mins

120mins



Current architectures are not ready

Architecture is not READY!

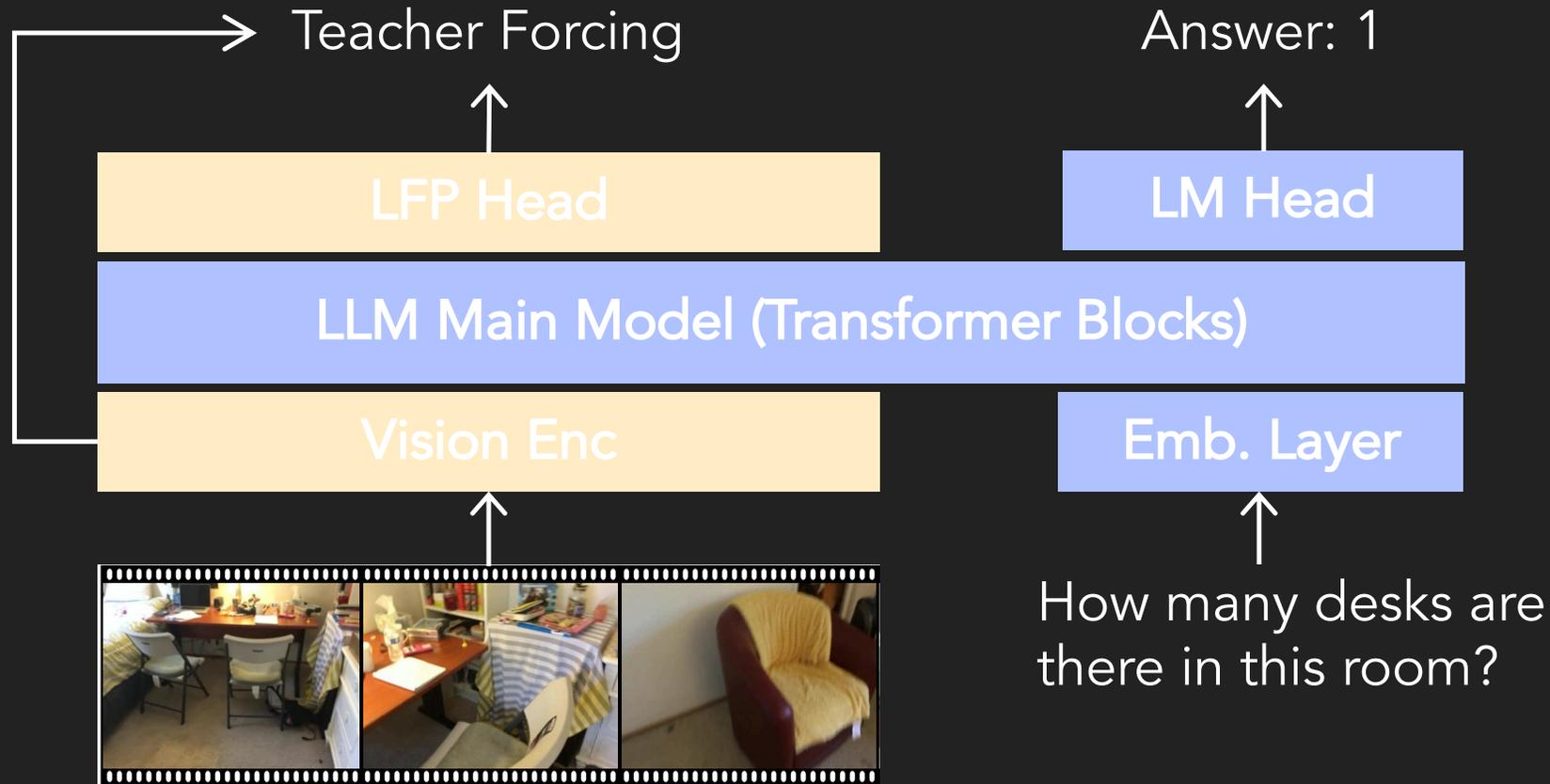


Typical Image/Video MLLM framework

Current architectures are not ready

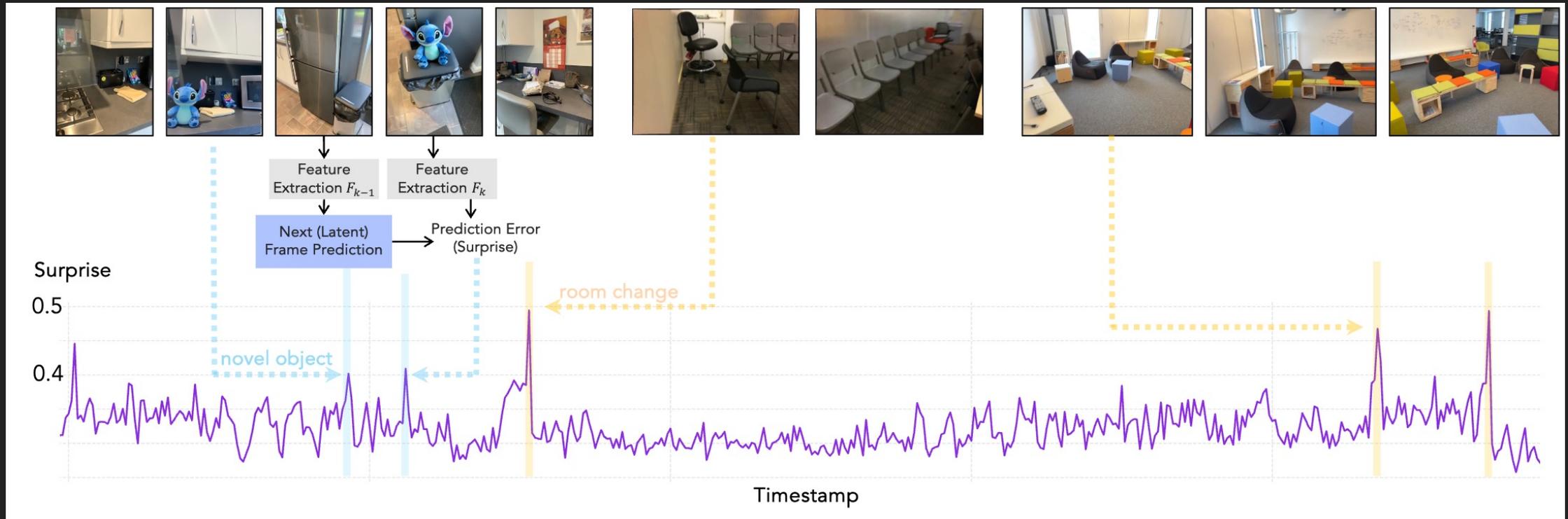
- What makes spatial supersensing unique?
- Infinite tokens in, infinite tokens out. Streaming, and always-on
- Our real-world experience isn't meant to be processed token by token

Prototyping: Predictive sensing



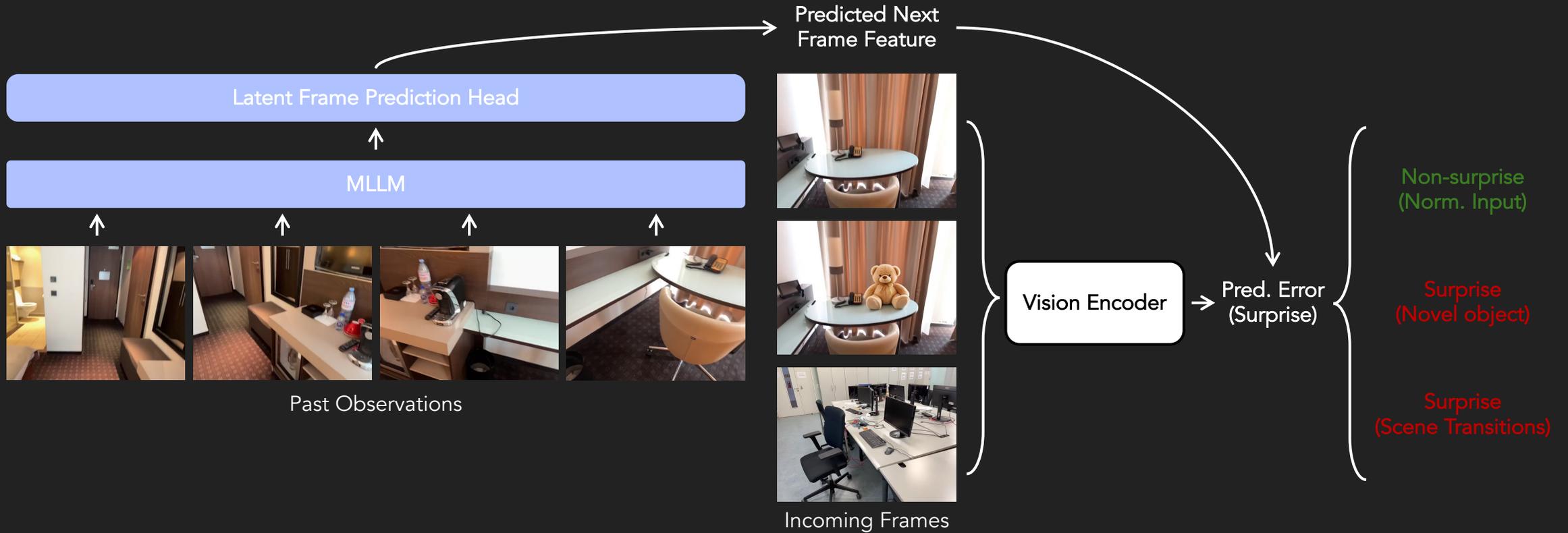
Building internal world model through latent-frame prediction in training

Violation-of-Expectation (or simply, surprises!): how humans regulate what information they take in.



Measure surprise (or violation-of-expectation)
during inference time

Predictive Sensing



Use Case #1: Memory Management

Infinitely-long input requires memory system, and memory is built for:

- **Store useful information**
- **Delete unused / duplicated information**
- **Retrieve useful information when being queried**

Use Case #1: Memory Management

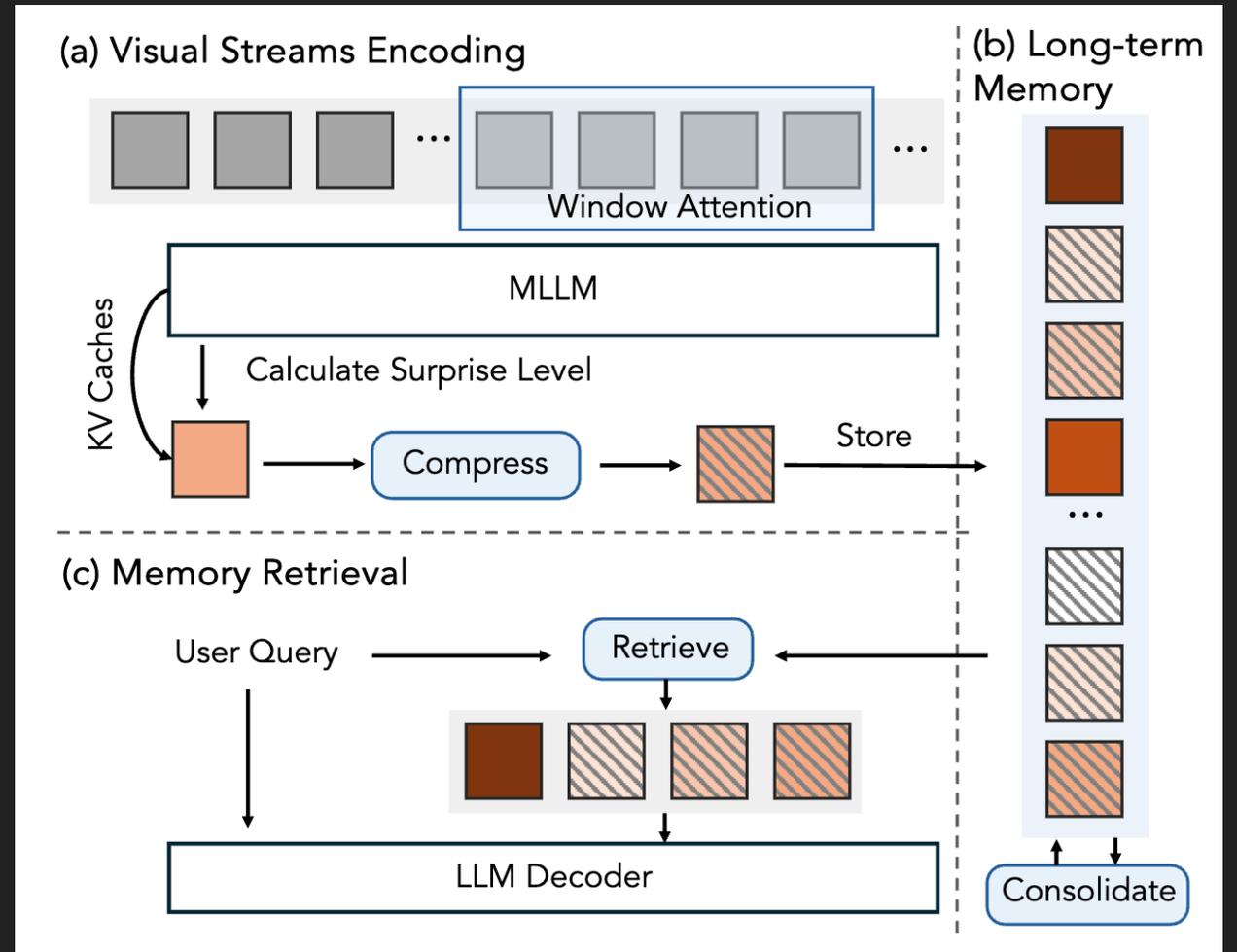
One of the key problem of infinitely long input:

Deciding what to keep and what to discard depends on its potential use for future tasks.

Surprise (high-entropy / unpredictable) info may be more important for future use!

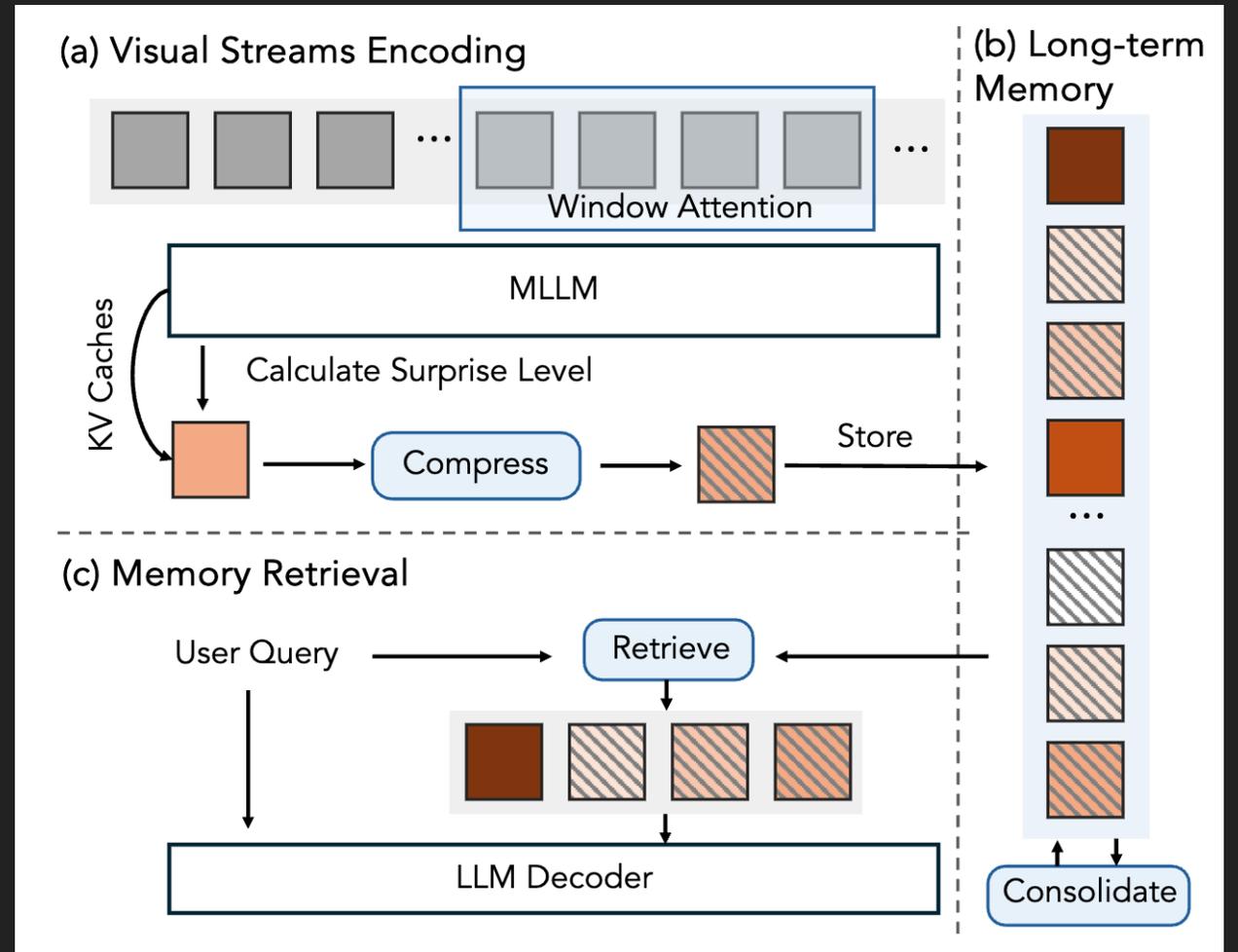
Surprise-driven Memory Management

- KV-cache as memory
- Sliding sensory memory
- Long term memory
- Memory consolidation
- Memory retrieval

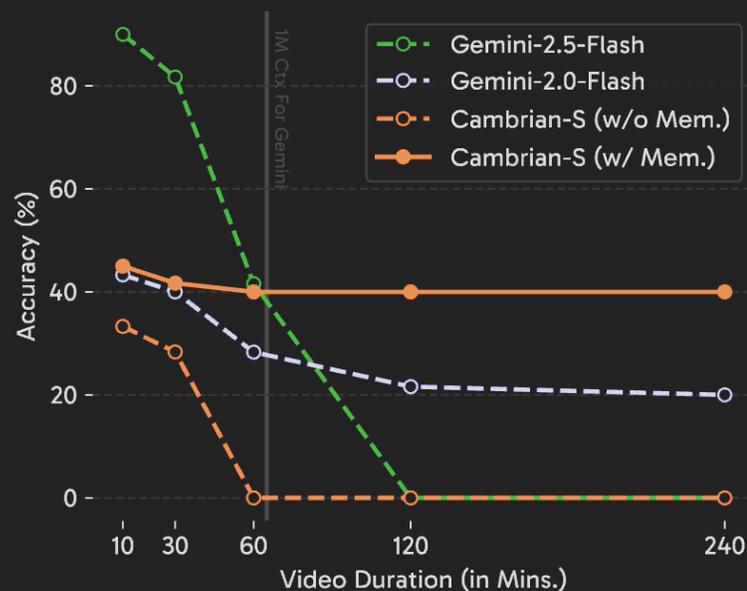


Surprise-driven Memory Management

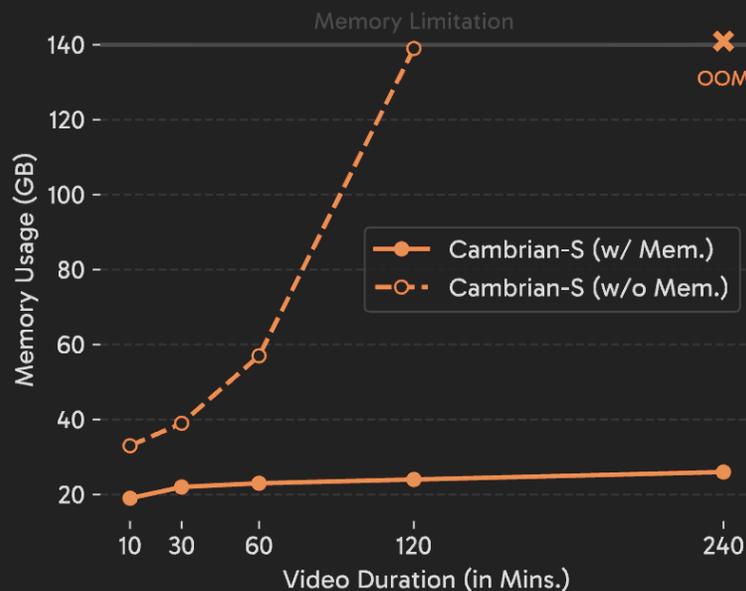
Compress and drop/merge low-surprise memory!



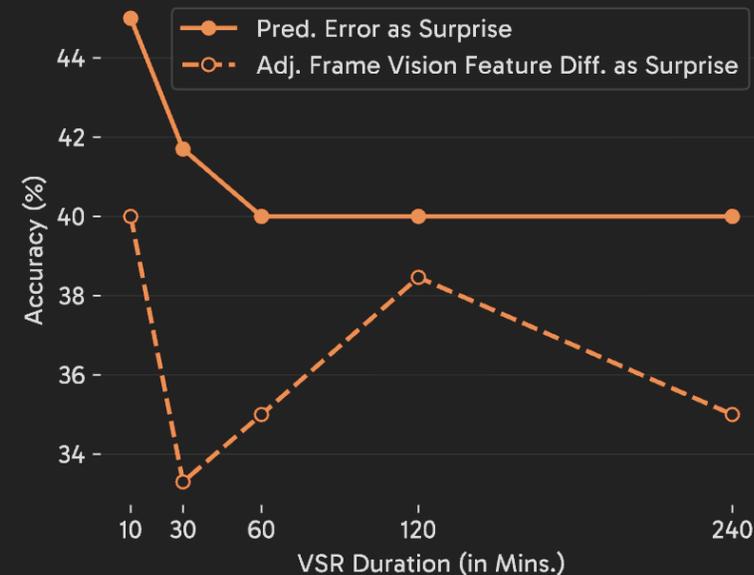
Use Case #1: Memory Management



(a) VSR results



(b) GPU memory usage



(c) Surprise comparison

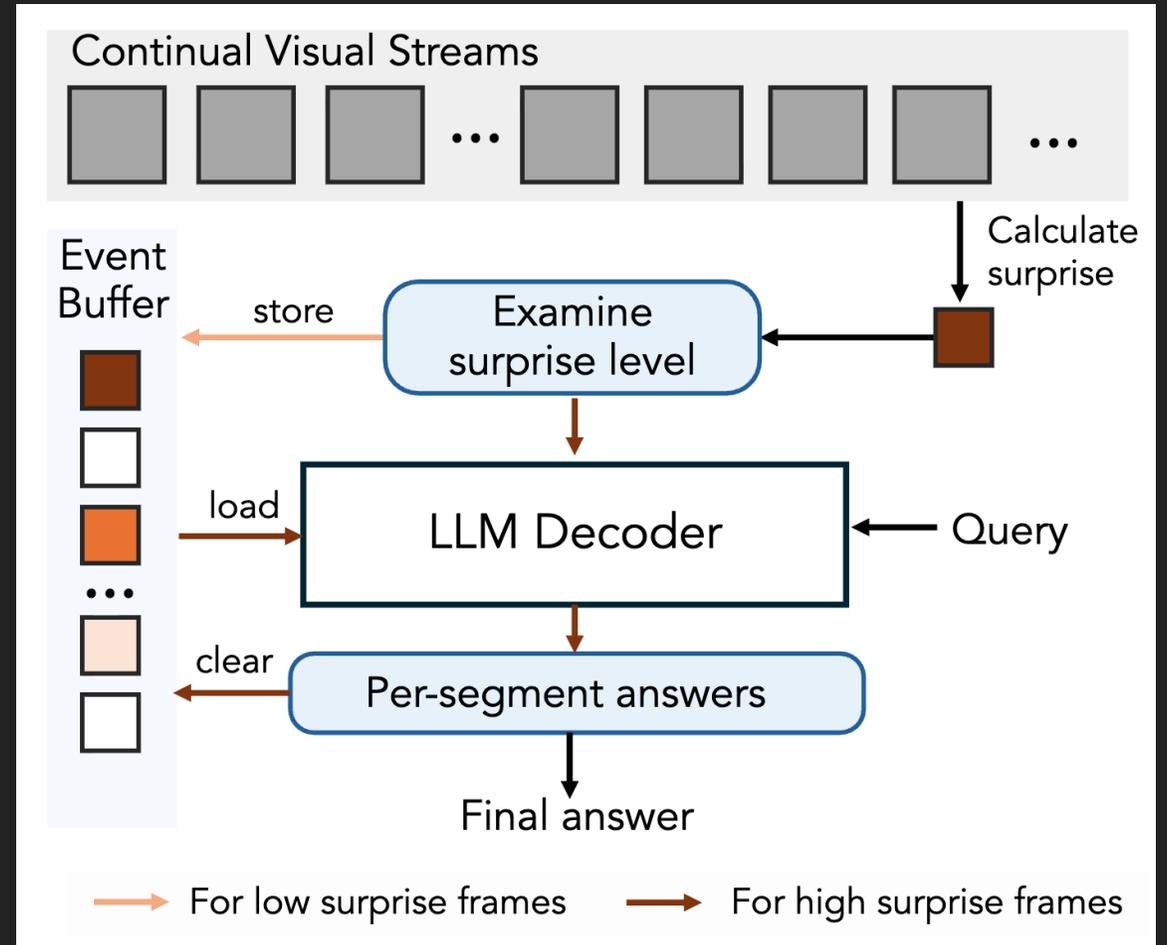
Better performance on VSI-SUPER Recall than Gemini models with consistent memory footprint

Use Case #2: Scene/Event Segmentation

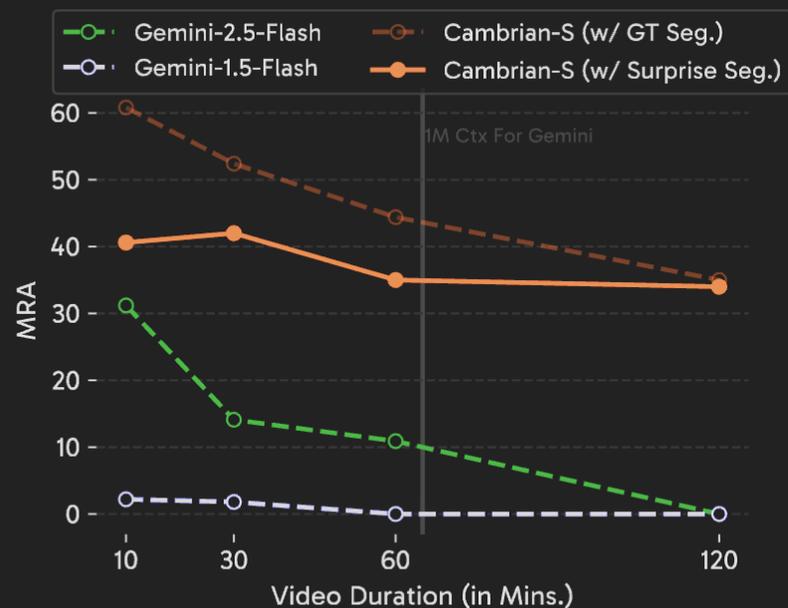
Surprise also helps human to organize memory structures
(like build episodic or event memory)

Surprise -driven Scene Segmentation

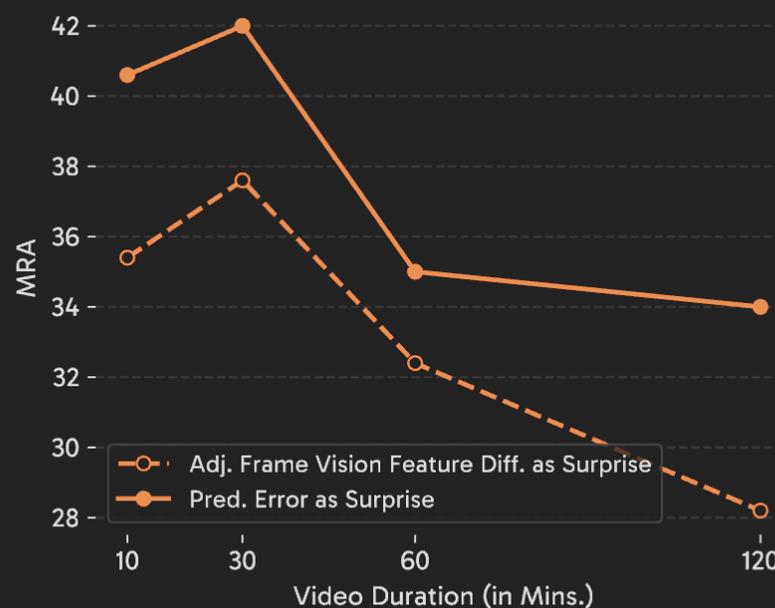
- Divide-and-conquer framework for VSI-SUPER Count
- Surprise for scene segmentation



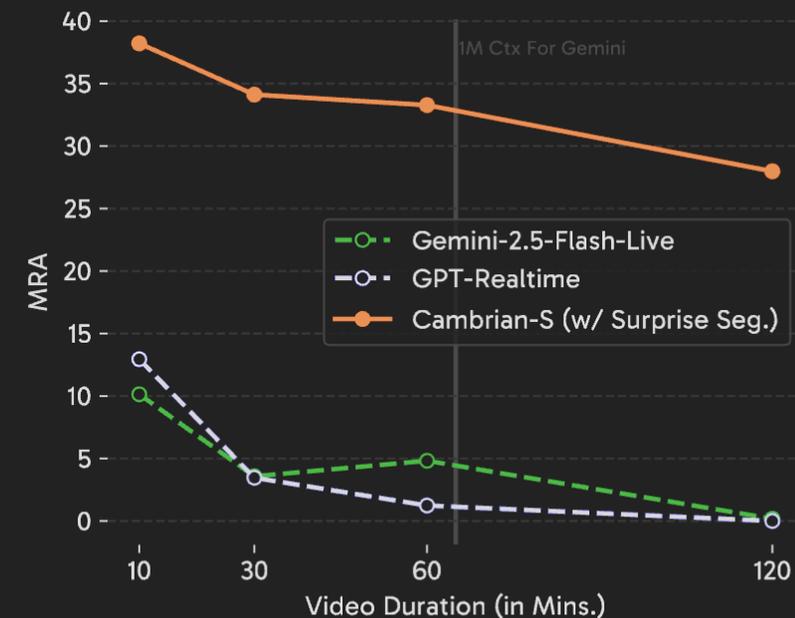
Use Case #2: Scene/Event Segmentation



(a) VSC results



(b) Different surprise measurement

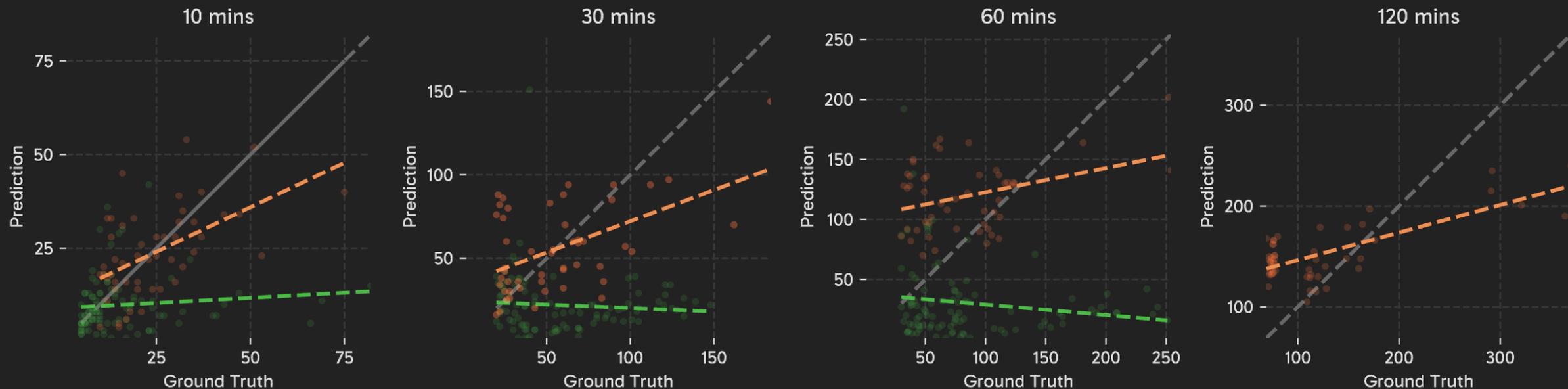


(c) Streaming evaluation

Better results on VSI-SUPER Count and its streaming version.

Use Case #2: Scene/Event Segmentation

■ Cambrian-S ■ Gemini-2.5-Flash ■ Line of Perfect Prediction



Divide-and-conquer framework generalize well to larger ground truth numbers

Summary

Supersensing as our ultimate target



Linguistic-Only Understanding

*Knowledge recall;
no sensory modeling*



Semantic Perception

*Naming and describing things
for user prompts*



Streaming Event Cognition

*Always-on sensing for open-
ended streams; memory
across time; proactive
answering*



Spatial Cognition

*Seeing the world behind the
video; implicit 3D*



Predictive World Modeling

*Unconscious inference;
Predictive, selective, and self-
updating world model*

world understanding and modeling in
infinite-long video context

Benchmark is NOT READY

- Existing benchmark is not ready for probing supersensing
- VSI-SUPER as an initial step
 - VSI-SUPER Recall
 - VSI-SUPER Count

We need more efforts on building real task!

Data is NOT READY

- Existing video data primarily focus on semantic parsing
- VSI-590K as our initial step toward better spatial sensing
- Cambrian-S models shows promising spatial sensing results but limited VSI-SUPER benchmark

**Scaling data alone does not solve
supersensing!**

Architecture is NOT READY

- Existing architecture is not ready for infinitely-long video context
- Prototype predictive sensing
- Violation-of-Expectation (surprise) for memory management or scene transition detection

Infinitely-long video modeling is a world modeling problem!

Details and resources are available at

- Project Page: <https://cambrian-mlm.github.io/>
- Website: <https://cambrian-mlm.github.io/cambrian-s/>
- arXiv: <https://arxiv.org/abs/2511.04670>
- Training and Inference code: <https://github.com/cambrian-mlm/cambrian-s>
- Model checkpoints: <https://huggingface.co/collections/nyu-visionx/cambrian-s-models>
- VSI-590K data: <https://hf.co/datasets/nyu-visionx/vsi-590k>
- VSI-SUPER benchmark: <https://hf.co/collections/nyu-visionx/vsi-super>

Cambridges

Thank you for listening!

**TOWARDS
SPATIAL SUPERSENSING**

Whats the right representation for WMs?

- RAE → much better
unified representation
- Scaling RAE → T2I
 - let you do diffusion in the same semantic latent space used for understanding
 - bridges understanding and generation into a single representation

DIFFUSION TRANSFORMERS WITH REPRESENTATION AUTOENCODERS

Boyang Zheng Nanye Ma Shengbang Tong Saining Xie
New York University

Scaling Text-to-Image Diffusion Transformers with Representation Autoencoders

Shengbang Tong*, Boyang Zheng*, Ziteng Wang*, Bingda Tang, Nanye Ma,
Ellis Brown, Jihan Yang, Rob Fergus, Yann LeCun, Saining Xie
New York University

 Website  Code  Models  Data

Beyond Language Modeling: An Exploration of Multimodal Pretraining

Shengbang Tong^{1,2,*}, David Fan^{1,*}, John Nguyen^{1,*},
Ellis Brown^{1,2}, Gaoyue Zhou^{1,2}, Shengyi Qian¹, Boyang Zheng², Théophile Vallaes¹, Junlin Han¹,
Rob Fergus^{1,2}, Naila Murray¹, Marjan Ghazvininejad¹, Mike Lewis¹, Nicolas Ballas¹, Amir Bar¹,
Michael Rabbat¹, Jakob Verbeek¹,
Luke Zettlemoyer^{1,†}, Koustuv Sinha^{1,†}, Yann LeCun^{2,†}, Saining Xie^{2,†}

¹FAIR, Meta, ²New York University

*equal contribution, †equal advising

The visual world offers a critical axis for advancing foundation models beyond language. Despite growing interest in this direction, the design space for native multimodal models remains opaque. We provide empirical clarity through controlled, from-scratch pretraining experiments, isolating the factors that govern multimodal pretraining without interference from language pretraining. We adopt the Transfusion framework, using next-token prediction for language and diffusion for vision, to train on diverse data including text, video, image-text pairs, and even action-conditioned video. Our experiments yield four key insights: (i) Representation Autoencoder (RAE) provides an optimal unified visual representation by excelling at both visual understanding and generation; (ii) visual and language data are complementary and yield synergy for downstream capabilities; (iii) unified multimodal pretraining leads naturally to world modeling, with capabilities emerging from general training; and (iv) Mixture-of-Experts (MoE) enables efficient and effective multimodal scaling while naturally inducing modality specialization. Through IsoFLOP analysis, we compute scaling laws for both modalities and uncover a scaling asymmetry: vision is significantly more data-hungry than language. We demonstrate that the MoE architecture harmonizes this scaling asymmetry by providing the high model capacity required by language while accommodating the data-intensive nature of vision, paving the way for truly unified multimodal models.

Date: March 3, 2026

Project Page: <https://beyond-llms.github.io/>

