CVPR MUSI Workshop 06.26

Towards Multimodal World Understanding from Video

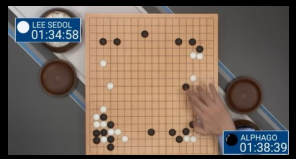
Saining Xie
New York University
AMI Labs

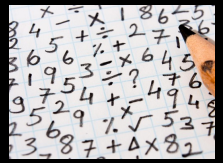
cambrian-mllm.github.io

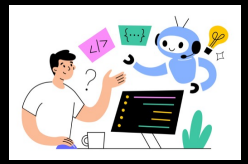
“what’s hard” and “what’s easy”

Hard for Humans

- Chess, Go
- Language fluency
- Math
- Coding

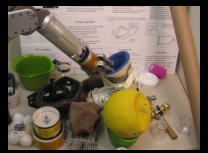


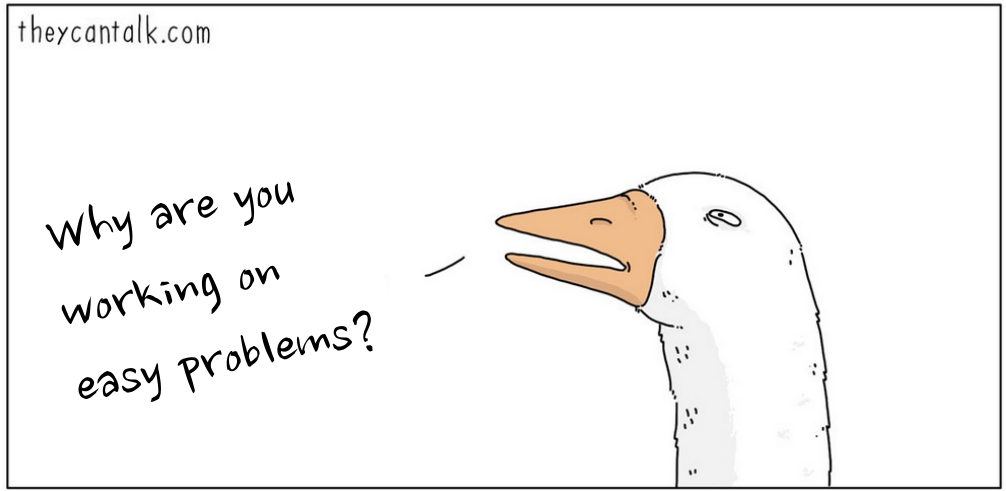




Easy for Humans

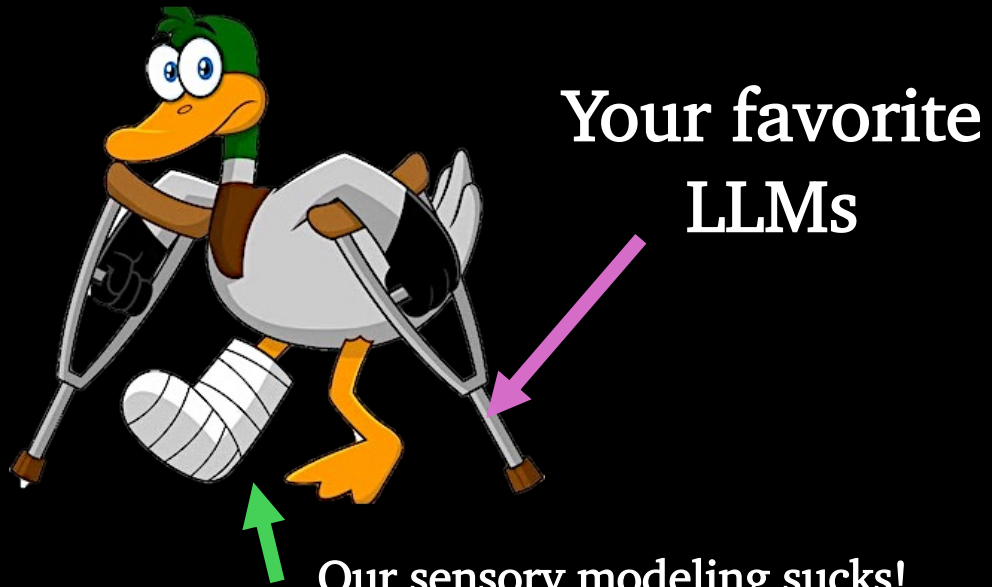
- Picking up objects
- Understanding a messy space
- Learning from real-world video
- Predictive world modeling




theycantalk.com
Why are you working on easy problems?


WHY ARE YOU WORKING ON EASY PROBLEMS?
CV Researcher
Moravec Paradox

Bet: relying too heavily too early on language can act as a shortcut, compensating for the deficiencies in real world modeling and understanding.



Towards Real World Understanding

Linguistic-Only Understanding

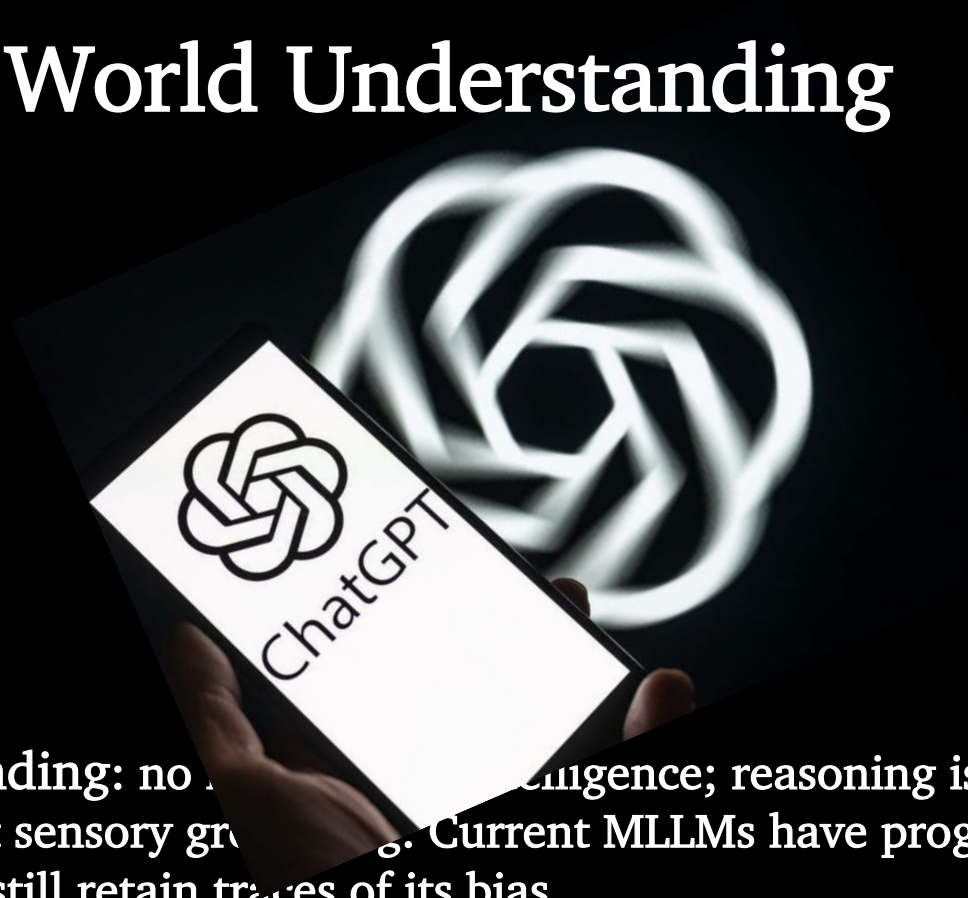
Knowledge recall; no sensory modeling

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 Real World Understanding

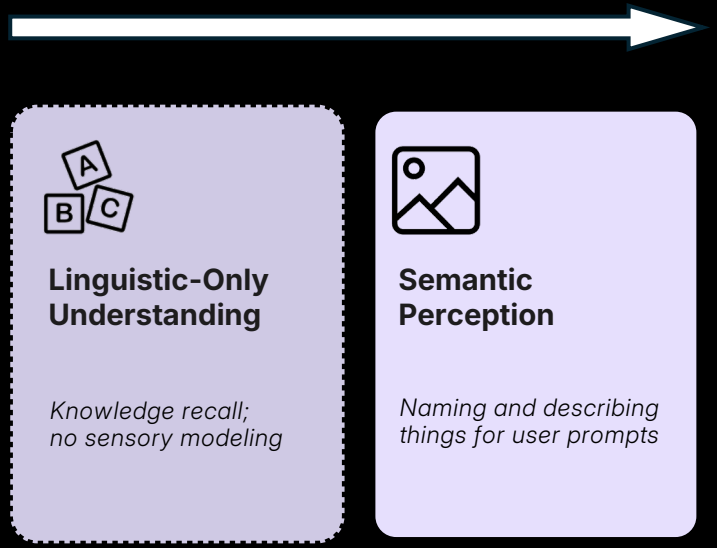
Linguistic-Only Understanding

Knowledge recall; no sensory modeling



Linguistic-only understanding: no [illegible]ligence; reasoning is confined to text and symbols without sensory gr[illegible]g. Current MLLMs have progressed beyond this stage, yet they still retain traces of its bias.

Towards Real World Understanding

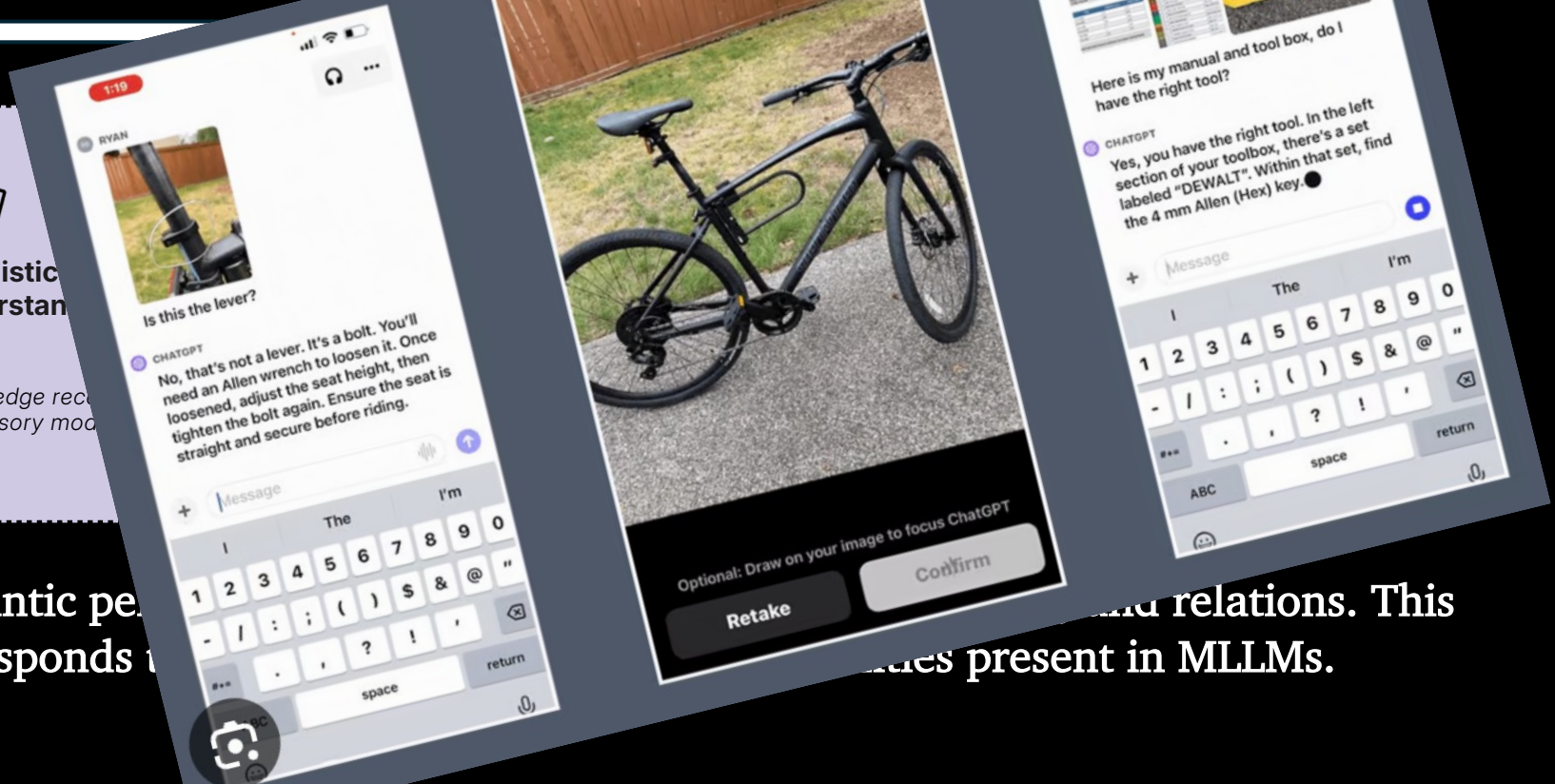


Semantic perception: parsing pixels into objects, attributes, and relations. This corresponds to the strong “show and tell” capabilities present in MLLMs.

Towards Real World U

Linguistic
Understan

Knowledge rec
no sensory mod



Semantic pe… …nd relations. This corresponds … …ities present in MLLMs.

Towards Real World Understanding

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 Real World Understanding

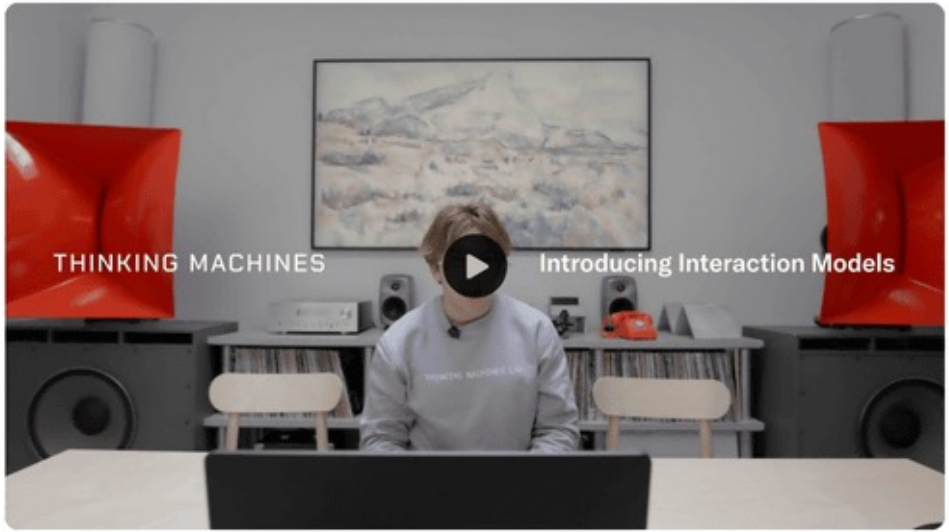
Linguistic-Only Understanding

Knowledge recall; no sensory modeling

Interaction Models: A Scalable Approach to Human-AI Collaboration

Thinking Machines

May 11, 2026



Streaming ev while proactively interpreting and responding to ongoing events. This aligns with efforts to make MLLMs real-time assistants.

Towards Real World Understanding

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 Re U ding

Linguistic-Only Understanding

Knowledge recall; no sensory modeling

Semantic Perception

Naming and things for us

Implicit 3D spatial cogn f a 3D world. Agents must know what is pr how configurations change over time. Today's video i nited here.

Towards Real World Understanding

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 Model

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.

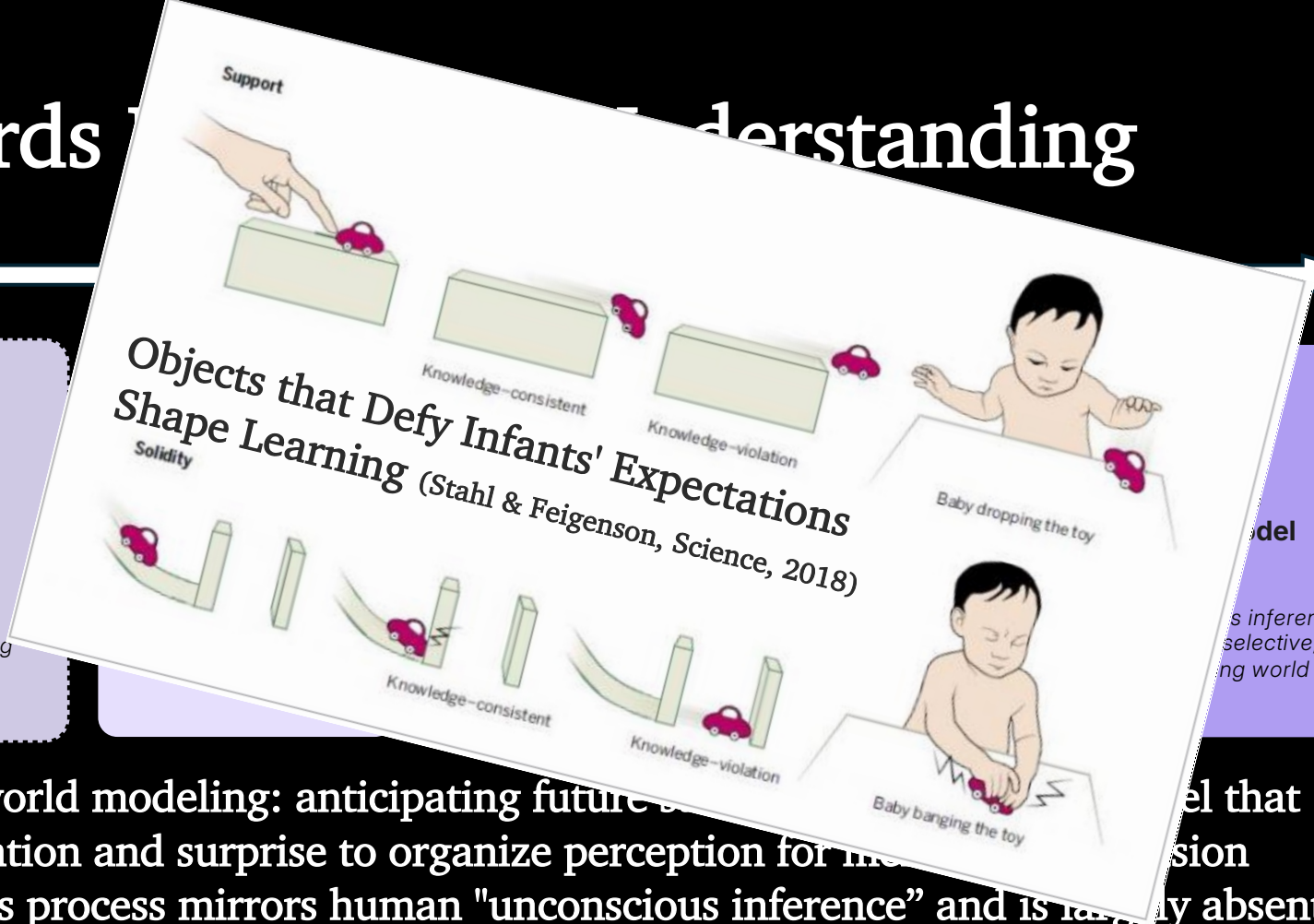
Towards derstanding

Linguistic-Only Understanding

Knowledge recall; no sensory modeling

del

s inference; selective, and ng world



Objects that Defy Infants' Expectations
Shape Learning (Stahl & Feigenson, Science, 2018)

Predictive world modeling: anticipating future … el that uses expectation and surprise to organize perception for … sion making. This process mirrors human "unconscious inference" and is … absent in current systems.

Towards Real World Understanding

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 Model

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

TASK-DRIVEN

WORLD MODELING

Towards multimodal world modeling: where are we now?

Evaluation is not ready

Data is not ready

Architecture is not ready[1]

Evaluation is not ready

Data is not ready

Architecture is not ready

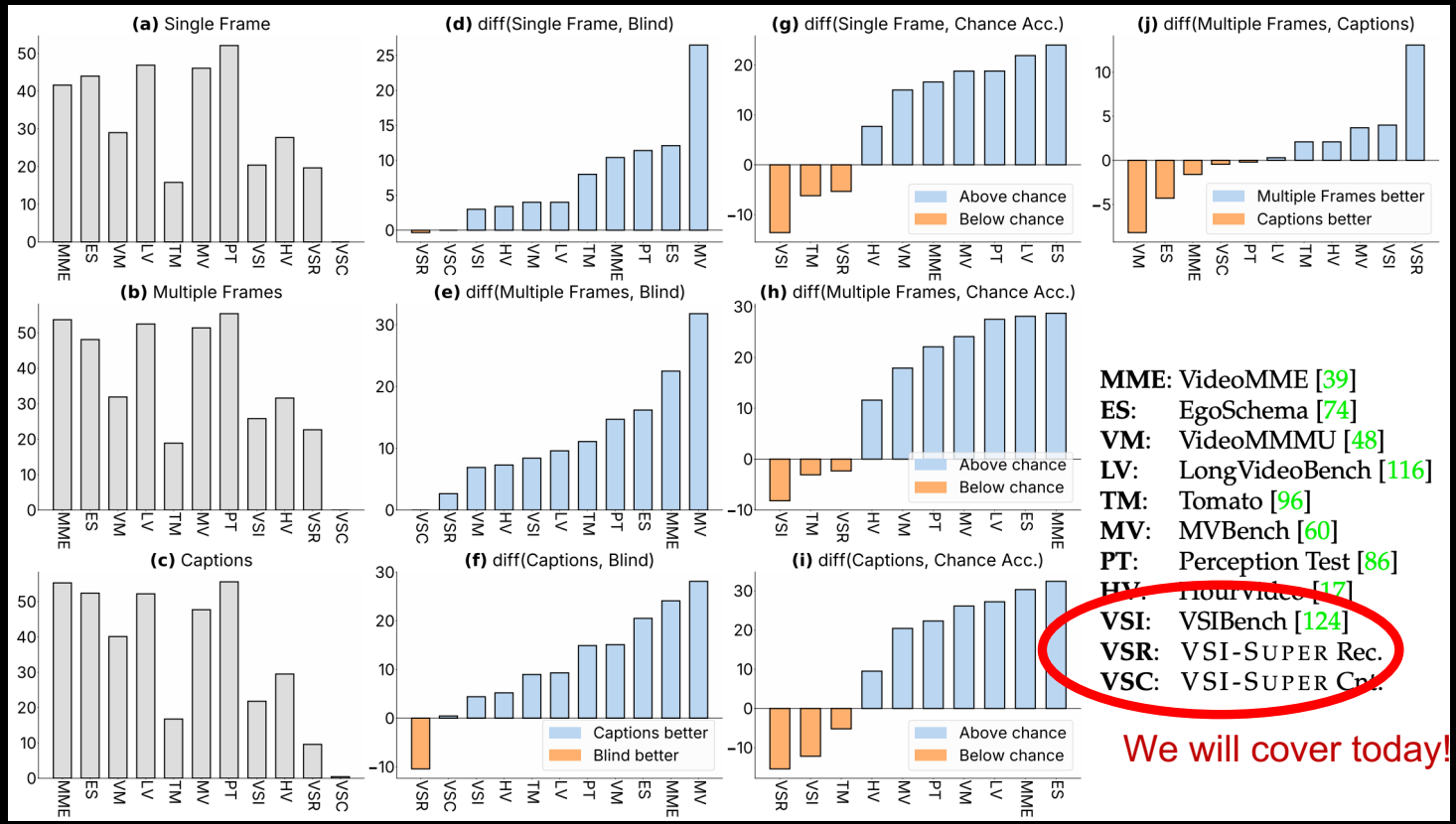
Current evals are not ready

- Video is the ultimate medium. But not all videos are the same. Without the right benchmarks, we risk taking the easy path instead of the right one.

	Previous SoTA	Humans	Gemini 2.5 Flash Preview 04/17*	Gemini 2.5 Pro Preview 05/06*
EVALUATIONS WITH VISUAL INPUTS				
EgoTempo (test set) 0-shot open-ended VideoQA	40.3 (GPT 4.1*)	63.2	36.5	**43.7**
LVBench (test set) 0-shot 4-choice VideoQA	60.1 (GPT 4.1*)	94.4	60.9	**68.2**
Perception Test (test set) 0-shot 5-choice VideoQA	71.4 (Oryx)	91.4	71.2	**77.3**
QVHighlights (val set) 4-shot Video Moment Retrieval	**76.1 (Mr BLIP)**	–	70.2	72.6
VideoMMMU (test set) 0-shot 5-choice VideoQA	76.7 (Kimi-k1.6)	74.4	71.9	**81.3**
1H-VideoQA (test set) **0-shot 5-choice VideoQA**	72.2 (Gemini 1.5 Pro)	–	64.3	**76.2**
EVALUATIONS WITH AUDIO-VISUAL INPUTS				
VideoMME (test set, long subset) 0-shot 4-choice VideoQA	72.0 (GPT 4.1)	–	77.8	**82.0**
YouCook2 Cap (val set) 4-shot Video Clip Captioning	**198.8 (VAST)**	–	185.3	198.0
YouCook2 DenseCap (val set) 4-shot Dense Video Captioning	67.2 (Vid2Seq)	–	67.6	**69.3**
EVALUATIONS WITH VISUAL-SUBTITLES INPUTS				
Minerva (test set) 0-shot 5-choice VideoQA	54.0 (GPT 4.1*)	92.5	61.9	**63.5**
Neptune (test set) 0-shot 5-choice VideoQA	85.1 (GPT 4.1*)	–	84.5	**85.4**
EVALUATIONS WITH AUDIO-VISUAL-SUBTITLES INPUTS				
VideoMME (test set) 0-shot 4-choice VideoQA	81.3 (Gemini 1.5 Pro)	–	79.3	**85.2**

Evaluation of Gemini 2.5 vs. prior models on video understanding benchmarks.
Performance is measured by string-match accuracy for multiple-choice VideoQA, LLM-based accuracy for EgoTempo, R1@0.5 for QVHighlights and CIDEr for YouCook2. *Videos were processed at 1fps and linearly subsampled to a maximum of 256 frames, except for 1H-VideoQA (7200 frames).

Deconstructing Existing Video 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



CVPR 2025

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

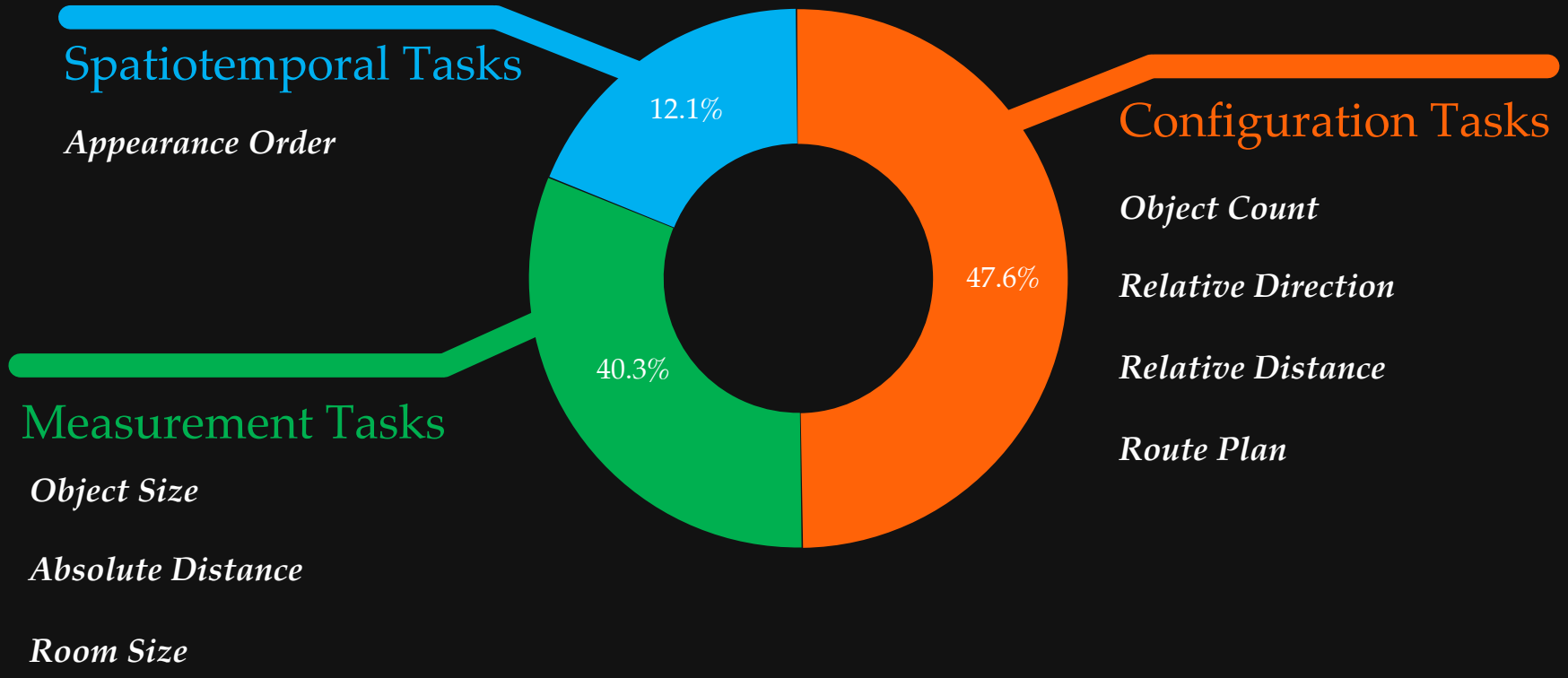
Gemini-1.5 Pro Answer: Right

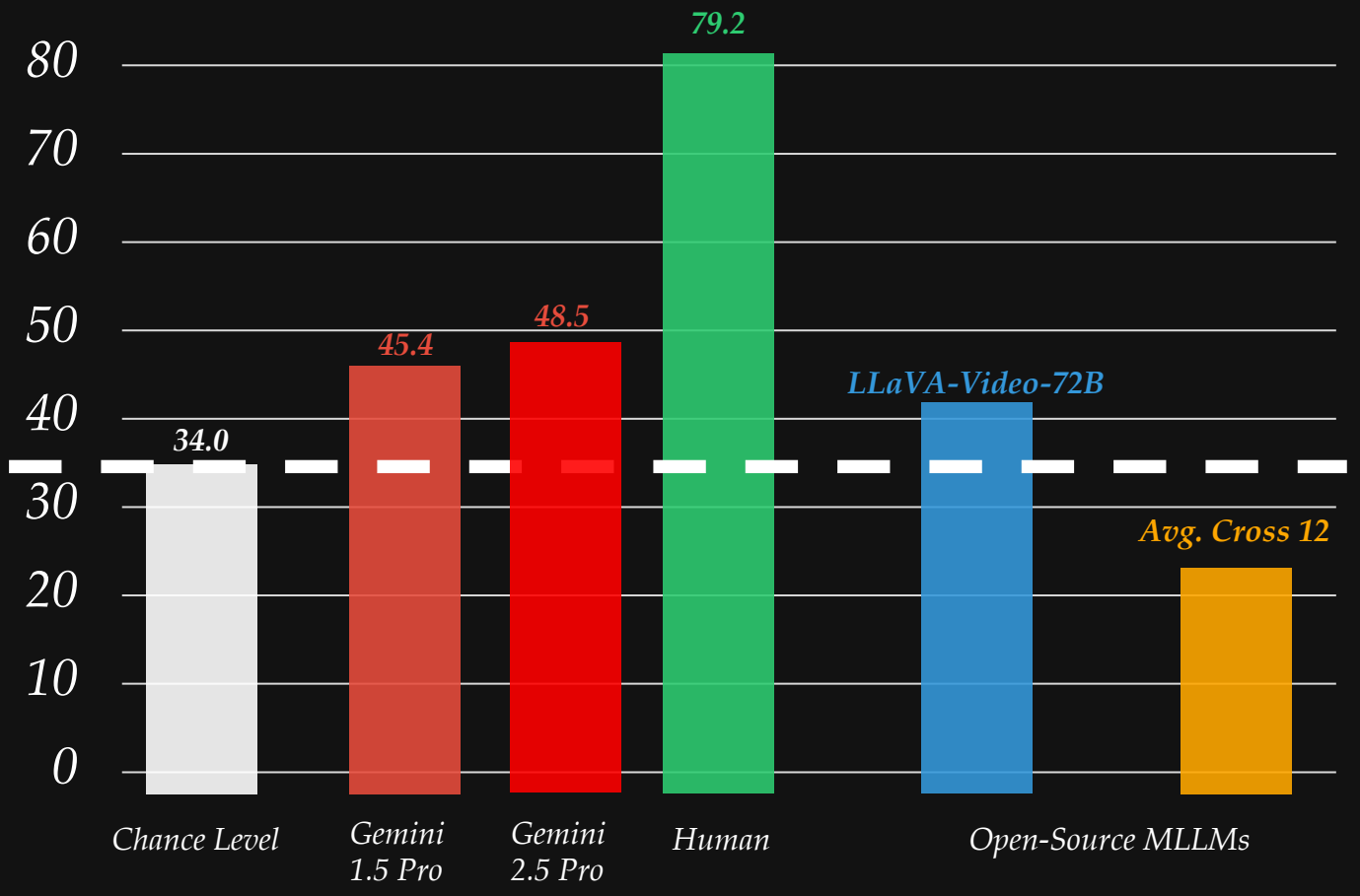


How do humans do this?
Can models do this? How?




Task Definition




80
70
60
50
40
30
20
10
0
79.2
48.5
45.4
34.0
LLaVA-Video-72B
Avg. Cross 12
Chance Level
Gemini 1.5 Pro
Gemini 2.5 Pro
Human
Open-Source MLLMs

Analysis by Self-Explanation

Error Breakdown

Visual Perception

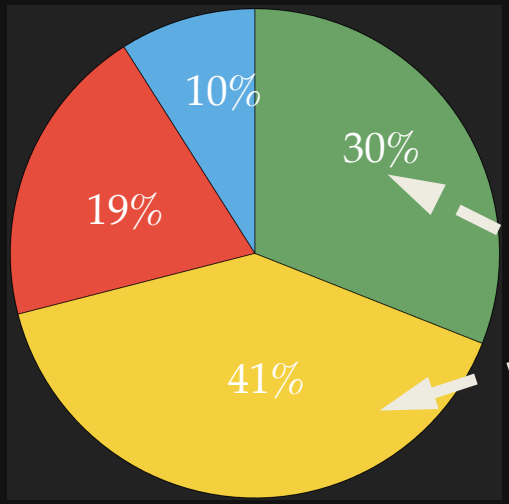
Linguistic Intelligence

Relational Reasoning

Egocentric-Allocentric

Spatial Reasoning

From 163 incorrect samples

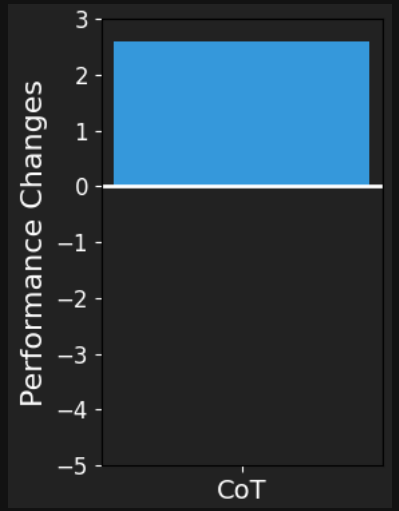


71% spatial reasoning errors

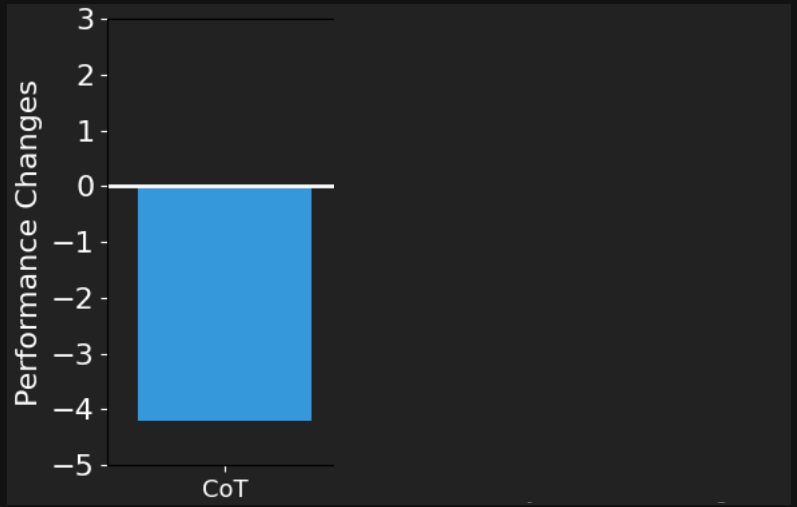
Spatial reasoning is the main bottleneck for MLLMs on VSI-Bench

Scaling Linguistic Reasoning

Chain-of-thought (CoT)



On Video-MME



On VSI-Bench

Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models.

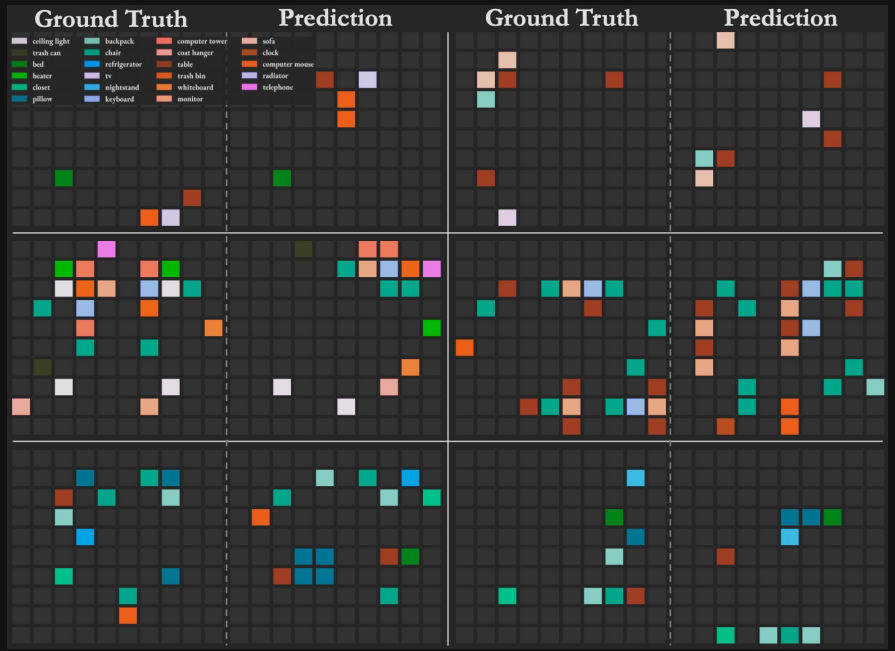
Analysis by Visualizing Cognitive Map

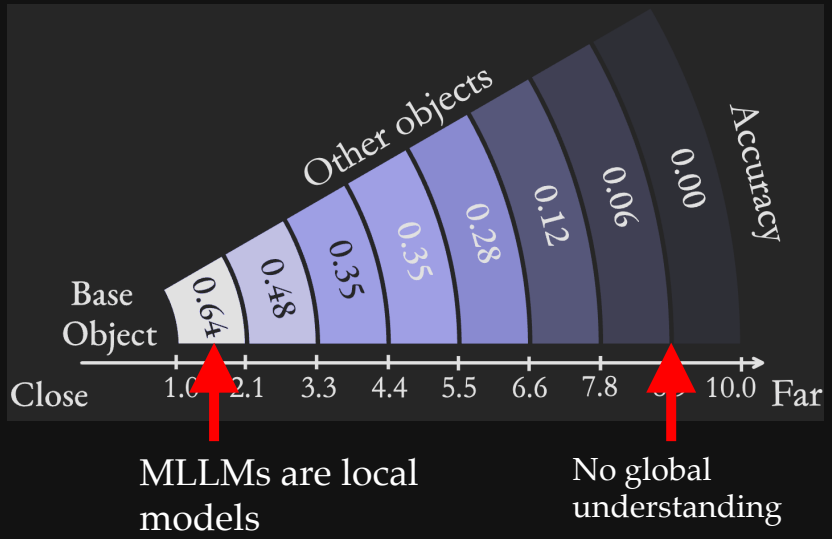
MLLM

10 by 10 cognitive map

Object center positions

Quantitatively Assess Cognitive Map

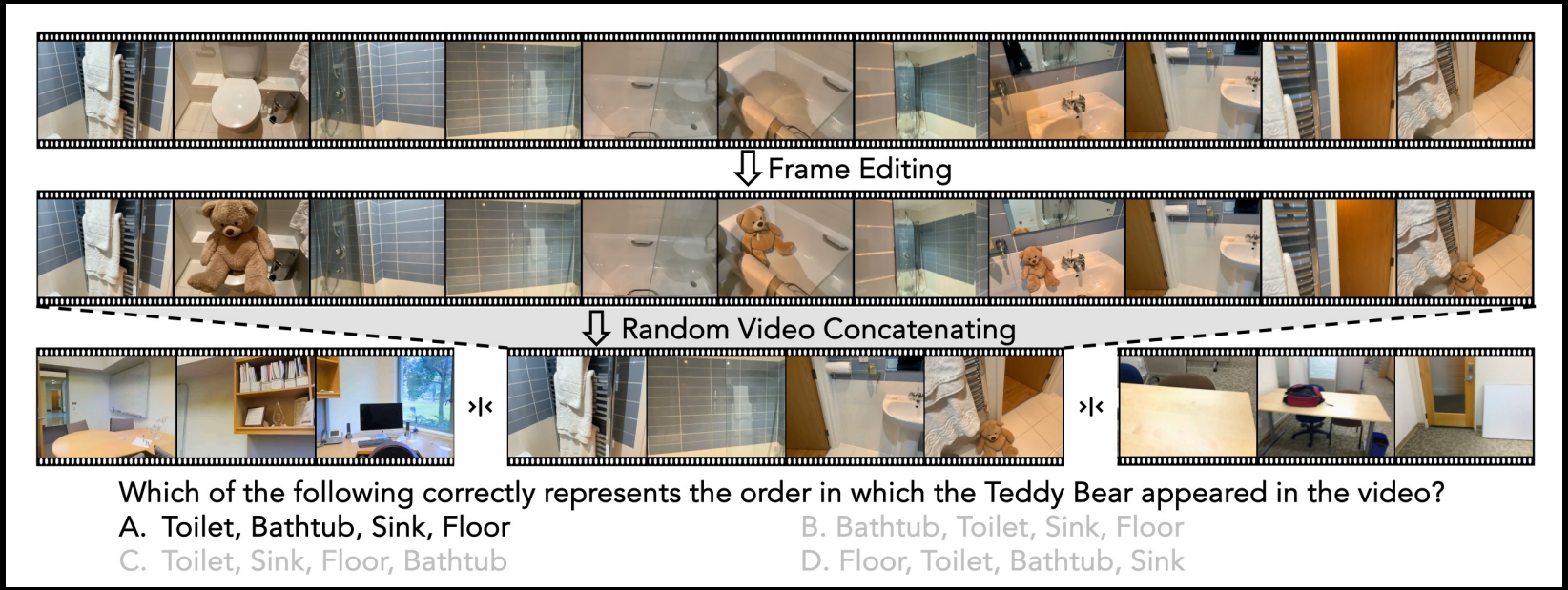





Cambrian-S
TOWARDS
SPATIAL SUPERSENSING

VSI-SUPER Recall:

Long-horizon spatial observation and recall



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

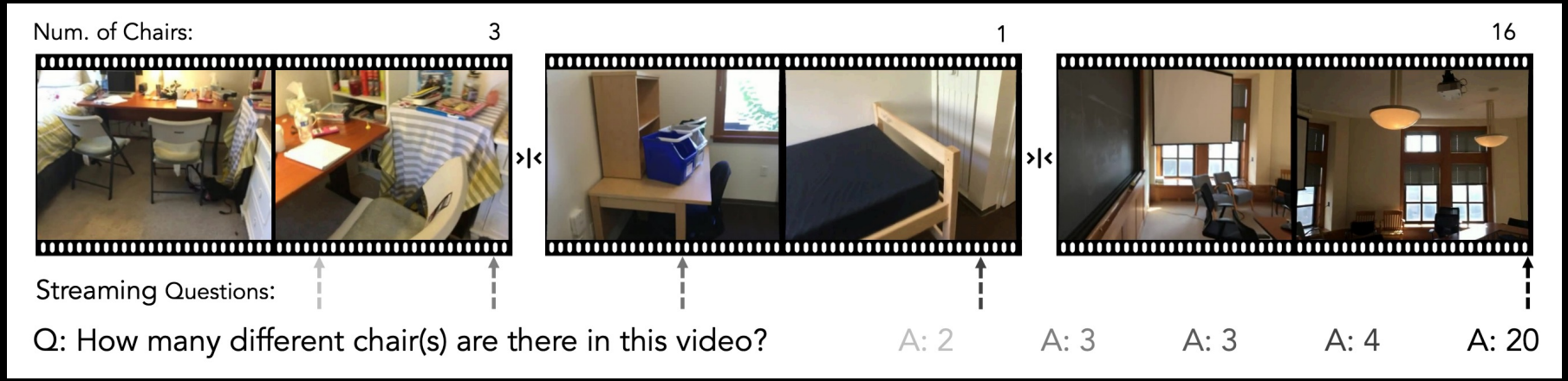
A. Stove, Trash bin, Refrigerator, Counter
B. Trash bin, Refrigerator, Counter, Stove
C. Stove, Counter, Refrigerator, Trash bin
D. Trash bin, Stove, Counter, Refrigerator

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

VSI-SUPER Count:

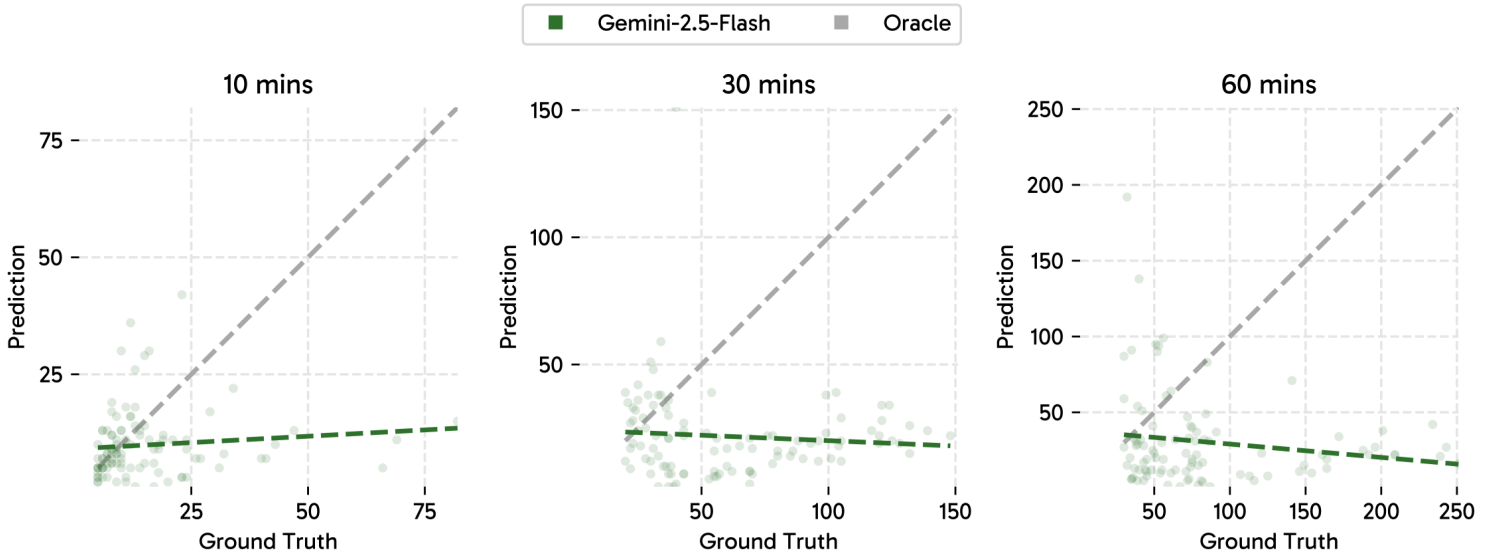
Continual counting under changing viewpoints and scenes.



Easy for humans,
yet extremely difficult for current models!

Gemini on VSI-SUPER

Model	VideoMME[39]	VideoMMMU[48]	VSI-Bench[124]	VSR		VSC	
				60 mins	120 mins	60 mins	120 mins
Gemini-2.5-Flash	81.5	79.2	45.7	41.5	Out of Ctx.	10.9	Out of Ctx.



“All benchmarks are wrong, but some are useful…”

Correcting errors in 3D annotations

ReVSI: Rebuilding Visual Spatial Intelligence Evaluation for Accurate Assessment of VLM 3D Reasoning

Yiming Zhang [*1] Jiacheng Chen [*1] Jiaqi Tan [1] Yongsen Mao [2] Wenhu Chen [3] Angel X. Chang [1 4]

Project Page GitHub Hugging Face

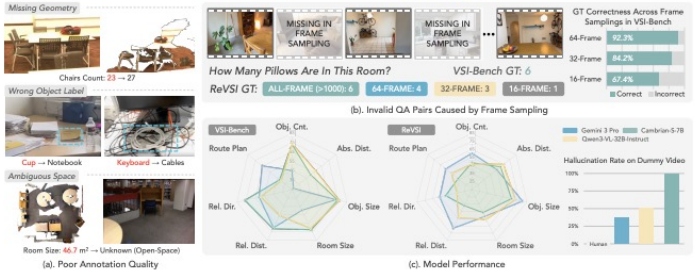


Figure 1. We revisit VSI-Bench (Yang et al., 2025a), a widely used benchmark for spatial reasoning, by systematically revealing issues in annotation correctness and bias (a) and pointing out that answers to questions should be sensitive to the input frames provided to the model (b). We develop a more accurate benchmark for assessing visual intelligence by thoroughly re-annotating, debiasing, and establishing new frame-aware protocols. We additionally construct *dummy-videos* by removing frames containing the queried objects, enabling controlled analysis of how models rely on visual evidence. Surprisingly, we find that proprietary models are under-assessed by VSI-Bench (*e.g.*, on object counting), while fine-tuned models show high hallucination rate under the dummy-videos setting (c).

Reducing the non-visual solvability from language biases

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

Abstract

Robust benchmarks are crucial for accurately evaluating Multimodal Large Language Models (MLLMs). However, we find that models can ace many multimodal benchmarks *without* strong visual understanding by exploiting biases, linguistic priors, and superficial patterns. This is particularly problematic for *vision-centric* benchmarks, which explicitly aim to require visual inputs to be solved. We introduce a diagnostic principle for robust benchmark design: if a benchmark *can* be gamed, it *will* be. Therefore, designers should proactively try to “game” their own benchmarks first as a key step in the development lifecycle—adopting rigorous diagnostic and debiasing procedures to systematically identify, quantify, and mitigate non-visual biases. We demonstrate that effective diagnosis of these issues *must* involve directly “training on the test set”—i.e., probing the *specific test set* being released for its intrinsic, exploitable patterns.

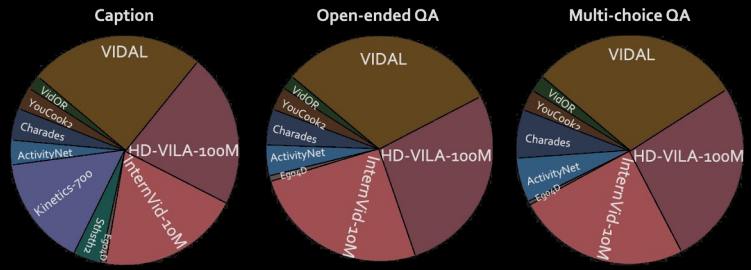
Benchmark is not ready

Data is not ready

Architecture is not ready

LLM data:
the Internet's greatest blessing

Spatial sensing data remain underrepresented in today's datasets.

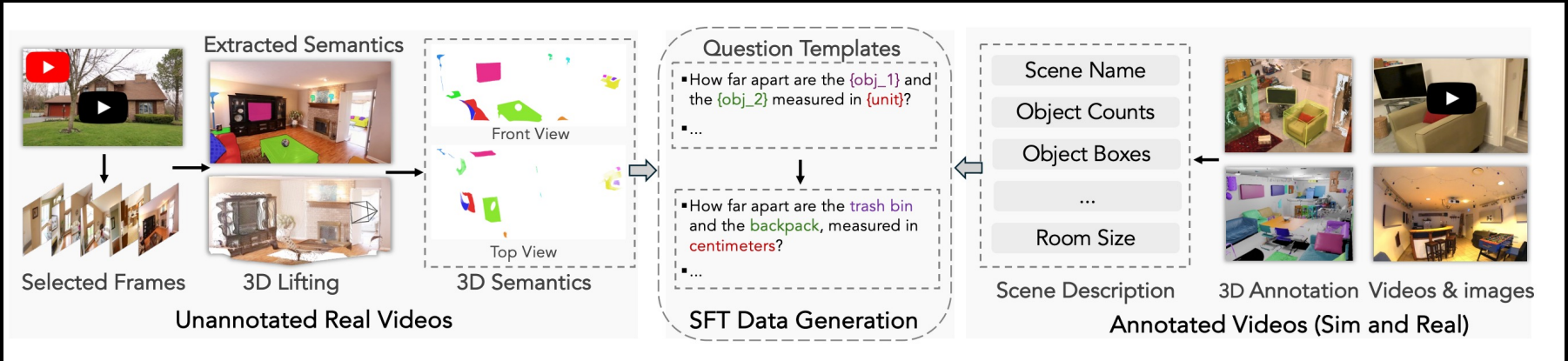


Dataset	#Caption	#Open-Ended	#Multi-Choice	Dataset	#Caption	#Open-Ended	#Multi-Choice
VidOR	4,018	19,875	4,773	Sthsth2	8,700	0	0
YouCook2	7,411	32,143	5,776	Ego4D	1,065	5,912	520
Charades	9,803	48,187	13,401	InternVid-10M	45,000	245,840	48,246
ActivityNet	7,953	44,100	12,771	HD-VILA-100M	48,260	263,652	51,743
Kinetics-700	34,998	0	0	VIDAL	55,000	300,472	58,968

Figure 5: Distribution of data across different datasets and question types (Caption, Open-ended, and Multi-Choice).

Llava-video-178K

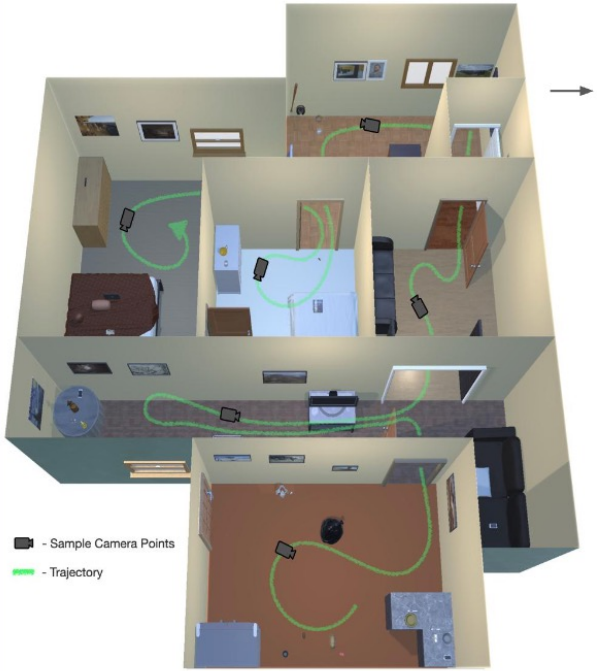
VSI-590K: a large-scale spatial sensing dataset



Data composition and sources

Dataset	# Videos	# Images	# QA Pairs
Annotated Real Videos			
S3DIS [3]	199	-	5,187
Aria Digital Twin [85]	183	-	60,207
ScanNet [31]	1,201	-	92,145
ScanNet++ V2 [129]	856	-	138,701
ARKitScenes [11]	2,899	-	57,816
Simulated Data			
ProcTHOR [34]	625	-	20,092
Hypersim [94]	-	5,113	176,774
Unannotated Real Videos			
YouTube Room Tour	-	20,100	20,100
Open X-Embodiment [83]	-	14,801	14,801
AgiBot-World [15]	-	4,844	4,844
Total	5,963	44,858	590,667

Simulating 3D-consistent spatial reasoning video training data…



Spatial

Obs. Level

Targeted QA Generation

… improves ***real*** video spatial performance

VSI-Bench

+ 8.4% LLaVA-Vid

+ 5.4% LLaVA-OV

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

and on ***out-of-domain*** benchmarks as well

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: Simulated Instruction-Tuning for Multimodal Spatial-Understanding Ellis Brown et al.

Data contributions

VSI Data Mixture	Image			VSI-Bench (Video)								
	RWQA[1]	3DSR	CV-B	Avg	Obj Ct	Abs Dst	Obj Sz	Rm Sz	Rel Dst	Rel Dir	Rte Pln	Ap Ord
Baseline	64.2	54.5	73.5	28.5	18.1	20.0	36.0	22.2	42.9	31.3	24.6	33.0
Real Videos												
+ S3DIS	65.1	54.9	76.3	41.6	43.8	21.0	44.9	37.0	43.8	47.4	34.0	41.1
+ ARKitScenes	66.8	56.7	77.3	51.0	70.2	32.7	64.5	60.0	55.1	45.2	**37.1**	43.5
+ ScanNet	67.5	57.7	77.5	54.1	70.9	37.8	65.5	58.2	57.9	46.7	35.1	56.1
+ ScanNet++ V2	66.1	57.3	77.5	56.3	72.5	40.7	63.7	56.9	59.7	47.1	31.4	70.2
Simulated Videos												
+ ProcThor	62.4	55.7	74.9	46.4	71.5	29.7	59.8	58.3	52.3	45.7	30.4	58.7
+ HyperSim	67.2	56.0	**79.7**	45.6	67.8	32.0	59.3	36.4	53.2	47.0	32.5	36.6
Pseudo-Annotated Images												
+ YTB RoomTour	62.2	52.6	75.0	32.5	43.4	25.8	24.2	27.3	38.7	31.4	28.4	40.9
+ OXE & AGIBot	64.4	54.4	72.5	30.6	40.3	23.1	27.9	26.6	38.0	22.8	32.0	33.8
All-in-One	60.8	54.0	77.9	**63.2**	**73.5**	**49.4**	**71.4**	**70.1**	**66.9**	**61.5**	36.6	**76.6**

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

+ 30% on VSI-Bench

Benchmark is not ready

Data is not ready

Architecture is not ready

Current architecture is not ready

What makes spatial sensing unique?

- ∞ tokens in, ∞ tokens out!

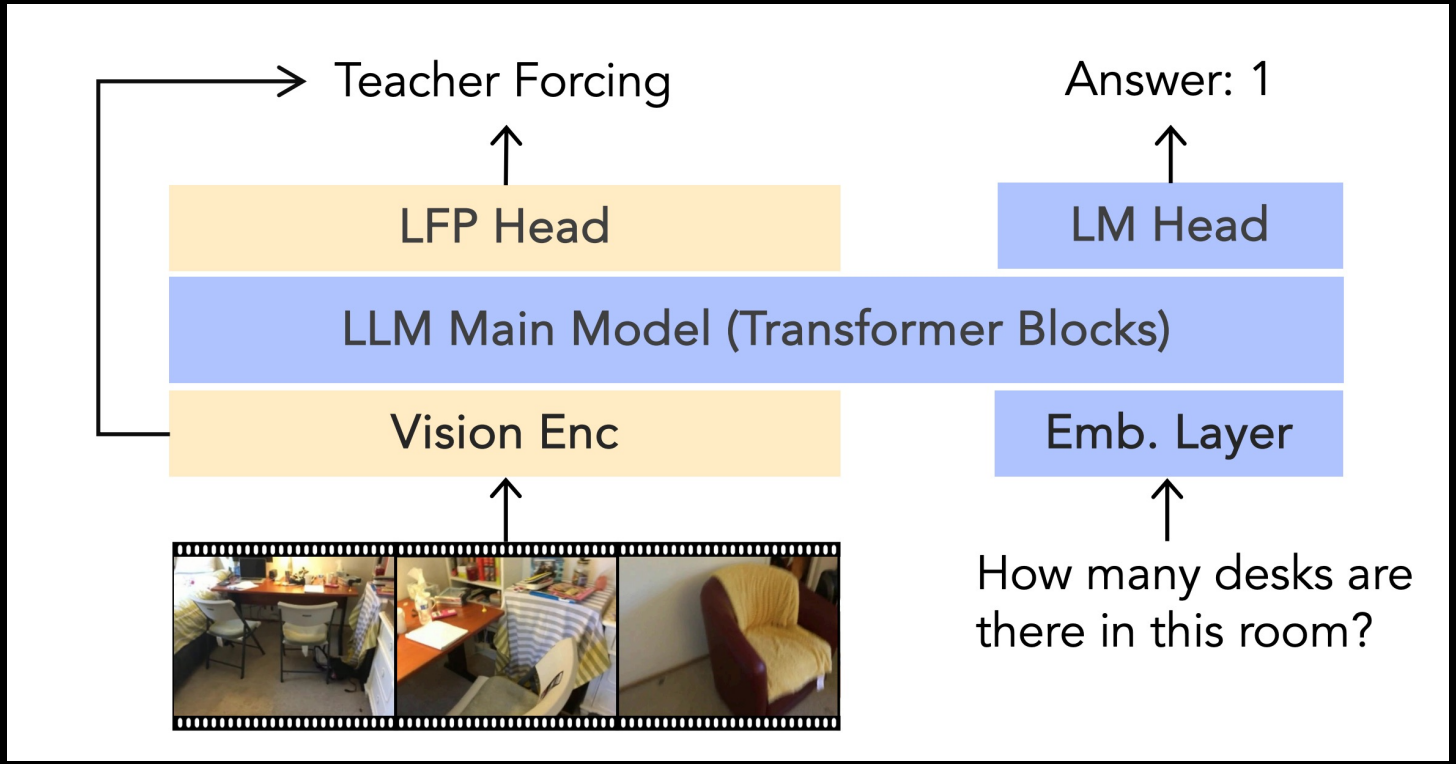
- our real-world experience isn’t meant to be processed token by token.

Current architecture is not ready

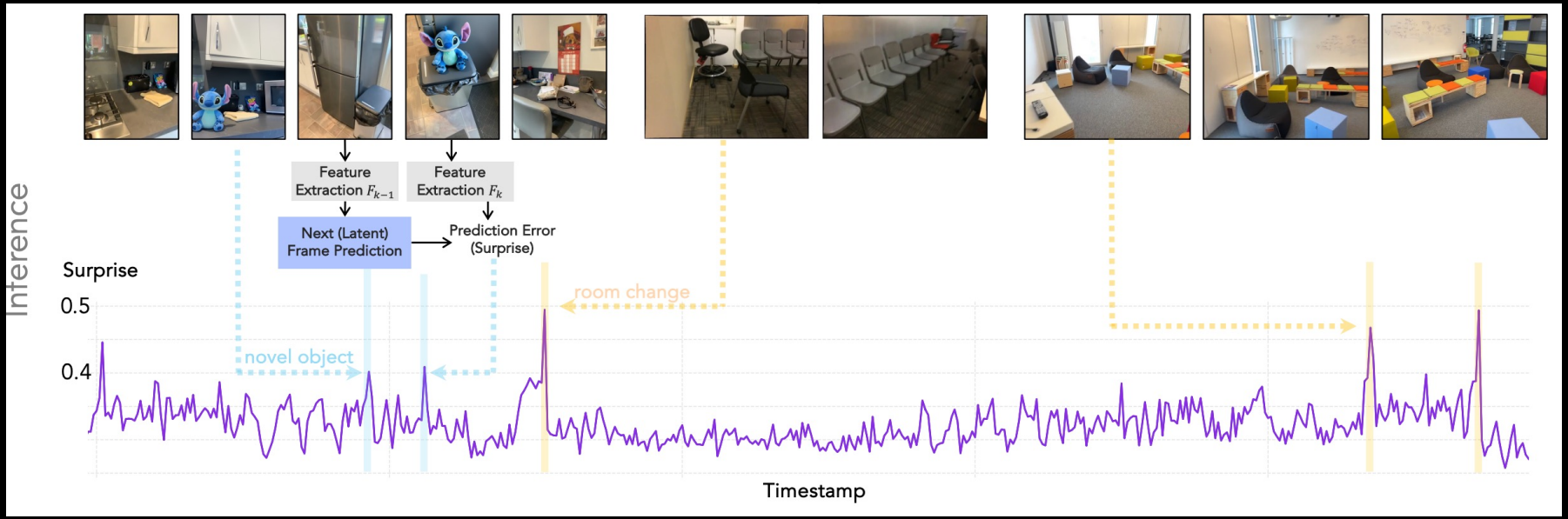
🧠 Human Visual Stream = Extremely High Bandwidth
👁️ Retina → Brain: ~10 million bits/sec
🌐 All sensory input (mostly vision): up to 1 billion bits/sec
🧩 Conscious awareness: only ~10 bits/sec

Most visual data is filtered and compressed before reaching perception. How?

Prototype: Predictive Sensing

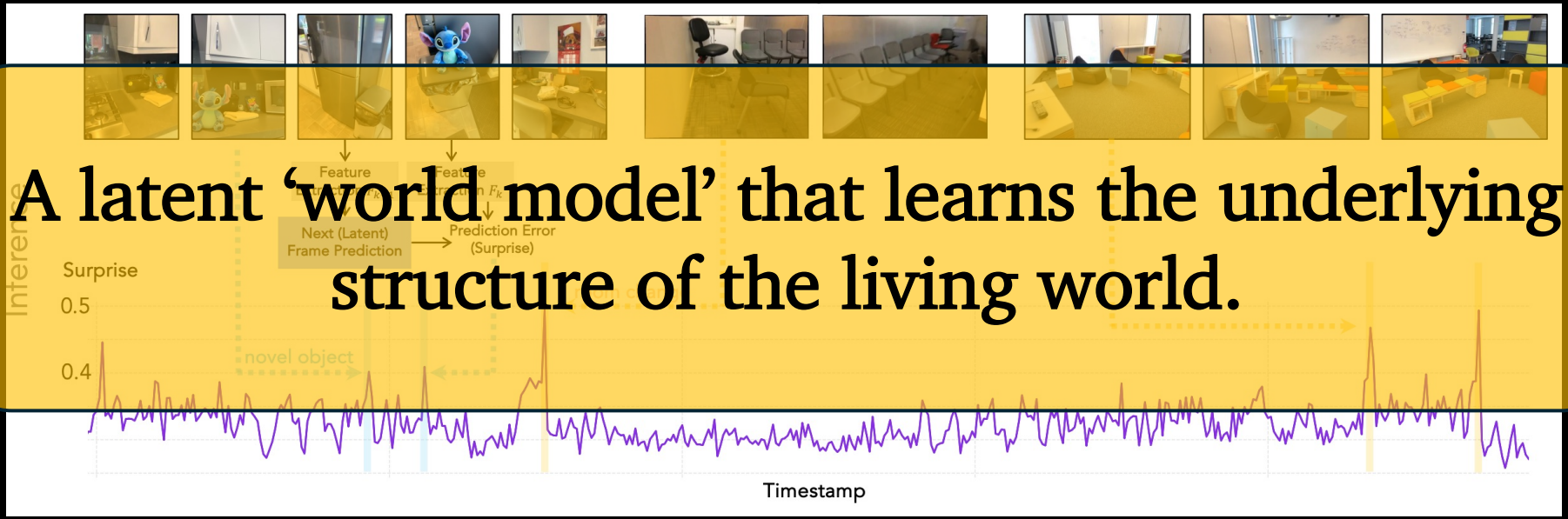


Violation-of-Expectation (or simply, surprises!):
how humans regulate what information they take in.



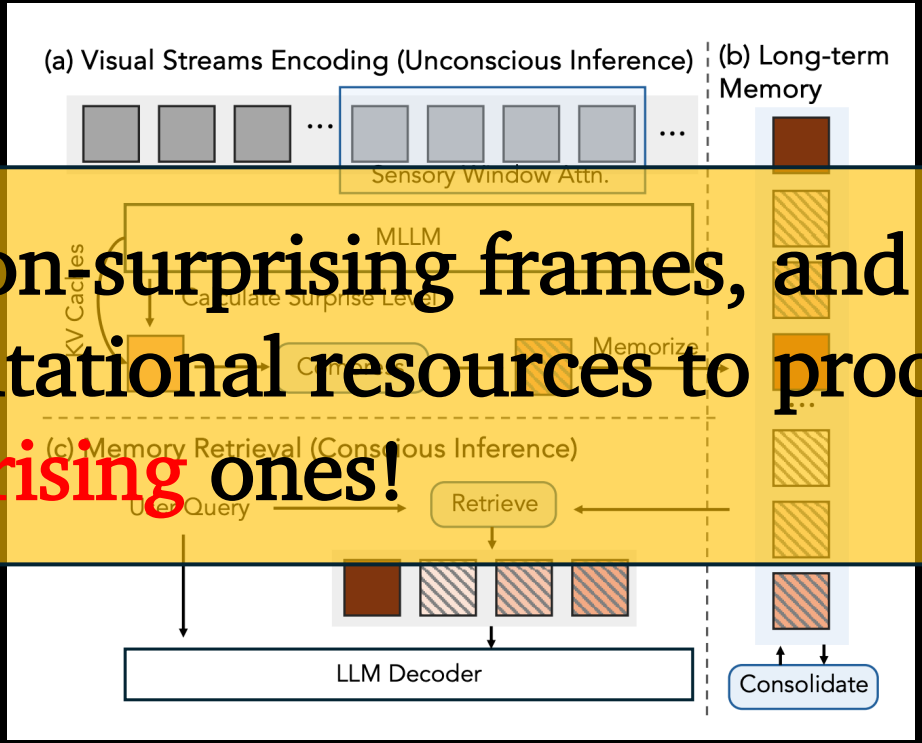
Violation-of-Expectation (or simply, surprises!):

how humans regulate what information they take in.



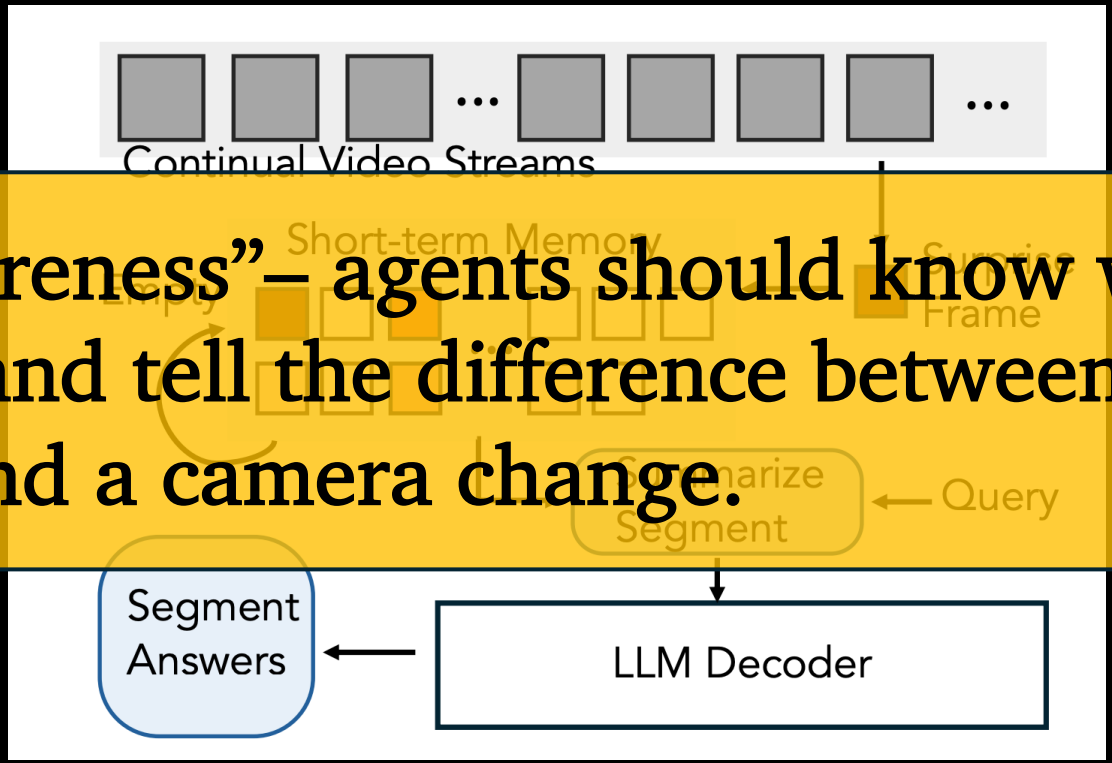
A latent ‘world model’ that learns the underlying structure of the living world.

Use Case #1: Memory Management

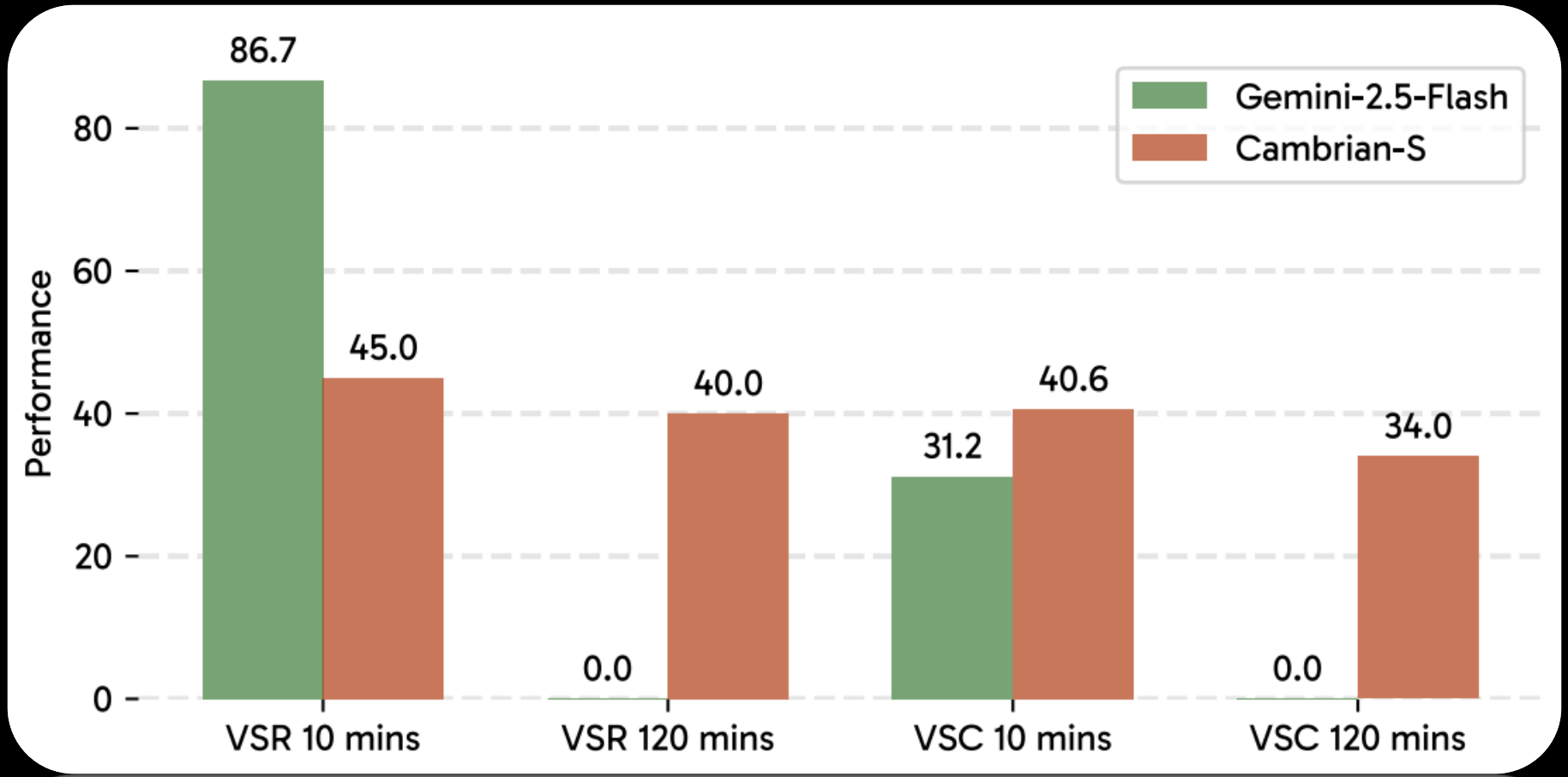


Compress non-surprising frames, and allocate more computational resources to processing and storing surprising ones!

Use Case #2: Scene/Event Segmentation



"Self-awareness"– agents should know where they are and tell the difference between a scene change and a camera change.


Gemini-2.5-Flash
Cambrian-S
Performance
0
20
40
60
80
86.7
45.0
0.0
40.0
31.2
40.6
0.0
34.0
VSR 10 mins
VSR 120 mins
VSC 10 mins
VSC 120 mins

Cambrian-*P*

Pose-Grounded Video Understanding

Jihan Yang[1*] Zifan Zhao[1*] Xichen Pan[1]

Shusheng Yang[1] Junyi Zhang[2]

Bingyi Kang[1] Hu Xu[3] Saining Xie[1]

[1]New York University [2]UC Berkeley

[3]Meta FAIR

Current architecture is not ready: Not grounded in 3D

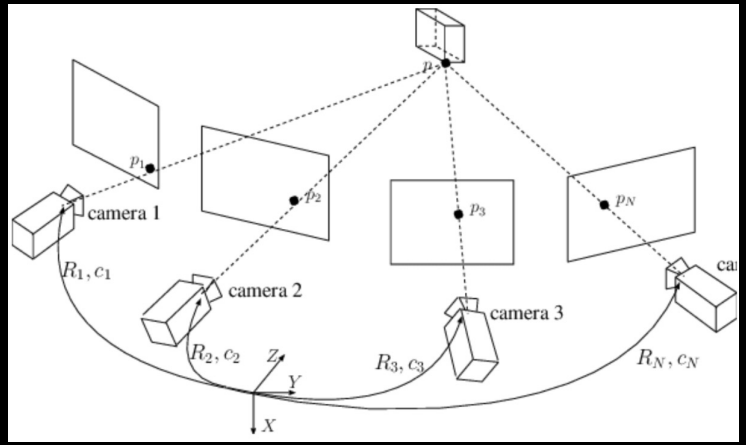
A video is a sequence of views from a moving camera 📸

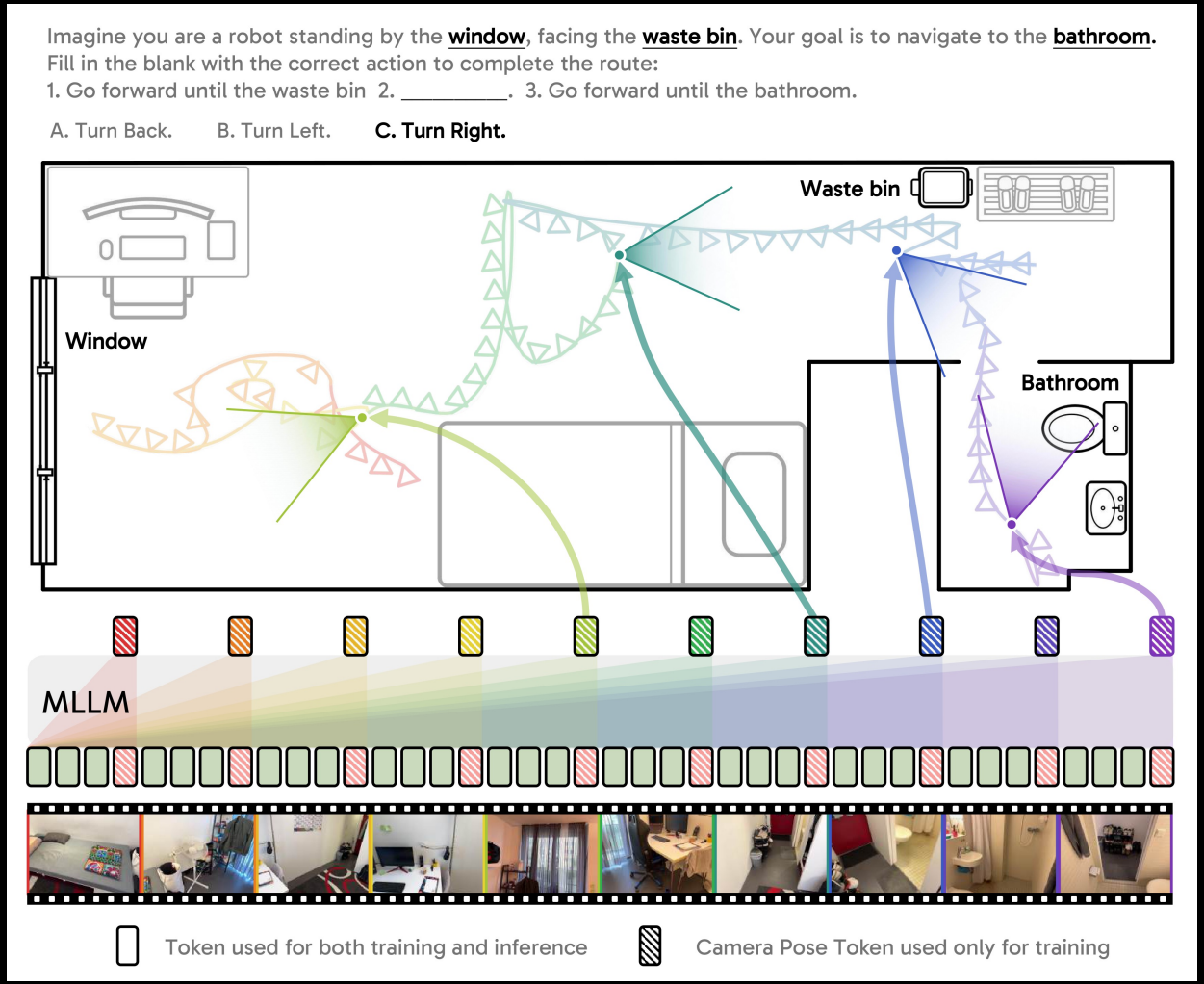
Each frame has

- Position $c \in R^3$
- Orientation $R \in SO(3)$

Pose matters:

- **Build correspondence between pixels**
- **separate ego motion from scene motion**
- **understand global relationship between frame**

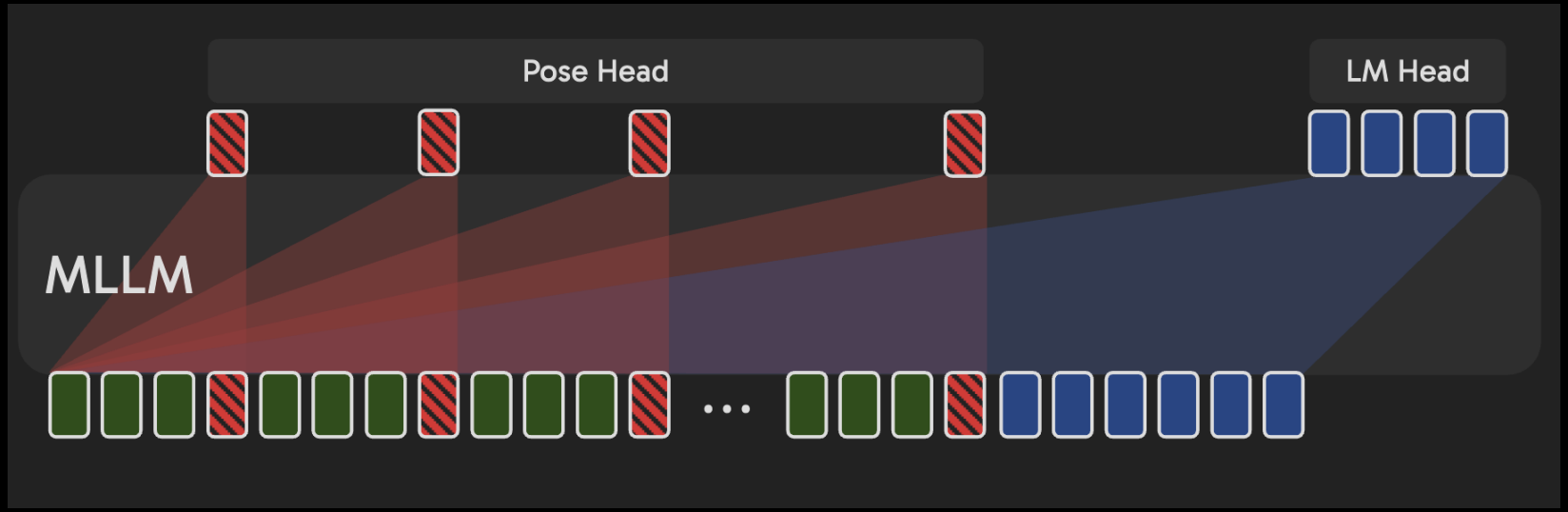




Cambrian-P:

a video multimodal LLM that jointly predict the camera's pose

Pose-grounded architecture



- Add learnable pose tokens per frame
- Tiny pose head (4 self-attn layers) regresses $g = [t, q, f]$ (absolute translation, rotation quaternion, field-of-view)
- Joint loss: $L = L_{NTP} + \lambda_{pose} \cdot L_{pose}$

VQA and Pose Estimation have conflicting training recipes

Naively adding a pose loss does not work out of the box.

	VQA training	Pose training
Frame sampling	Uniform intervals	Random start, dynamic intervals
Training duration	1 epoch	Tens of epochs
Data augmentation	None (preserve facts)	Heavy (color jitter, blur, grayscale)

The gap:
Uniform sampling ⇒ same poses each iteration ⇒ memorization shortcut.
Heavy augmentation ⇒ scrambled colors ⇒ hurts semantic understanding.

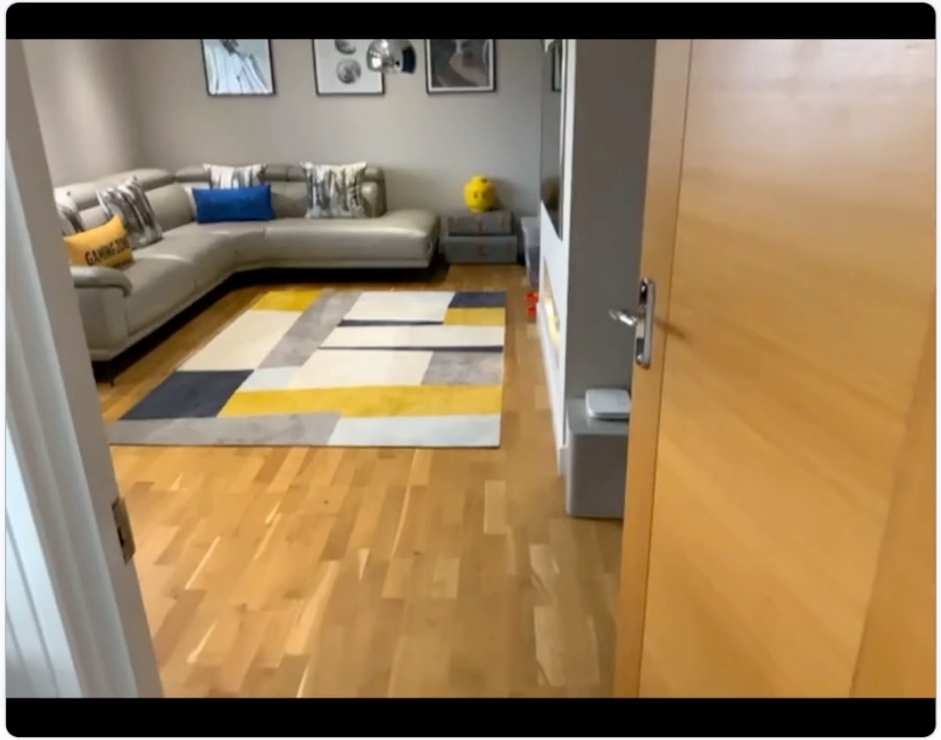
Solution: Interleaved Training + Random Jitter

Two simple ideas:

- Interleaved batches: dedicated pose-only samples use dynamic sampling + augmentation (pose loss only); VQA samples use uniform sampling (joint loss)
- Random jitter sampling: perturb VQA frame indices by $\delta \sim \mathrm{U}(-\Delta,\Delta)$, $a = 0.005$ to break pose memorization without hurting VQA

Cambrian-P pred

Mesh comes from ground truth.



Predicted poses are only for illustration and not used during VQA inference.

Table 1 | **VSI-Bench Results Comparison**. † indicates this Cambrian-*S* is fine-tuned only on VSI-590K.

Model	LM		Numerical Answer				Multiple-Choice Answer			
		Avg.	Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
Baselines										
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
General-purpose Models										
GPT-4o [41]	Unk.	34.0	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5
Gemini-2.5 Pro [25]	Unk.	51.5	43.8	34.9	64.3	42.8	61.1	47.8	45.9	71.3
Qwen2.5VL-7B [8]	Qwen2.5-7B	29.3	25.2	10.5	36.4	29.6	38.4	38.0	29.8	26.8
InternVL-3 8B [129]	Qwen2.5-7B	42.1	68.1	39.0	48.4	33.6	48.3	36.4	27.3	35.4
InternVL-3.5 8B [129]	Qwen3-8B	56.3	–	–	–	–	–	–	–	–
Qwen3-VL 8B [7]	Qwen3-8B	56.6	–	–	–	–	–	–	–	–
Spatial-specialist Models										
VST 7B [110]	Qwen2.5-7B	61.2	71.6	43.8	75.5	69.2	60.0	55.6	44.3	69.2
VLM-3R 7B [29]	Qwen2-7B	60.9	70.2	49.4	69.2	67.1	65.4	80.5	45.4	40.1
VG-LLM 8B [128]	Qwen2.5-7B	50.7	67.9	37.7	58.6	62.0	46.6	40.7	32.4	59.2
Cambrian-*S* 7B [111]	Qwen2.5-7B	67.5	73.2	50.5	74.9	72.2	71.1	76.2	41.8	80.1
SenseNova-SI 8B [17]	Qwen2.5-7B	68.7	–	–	–	–	–	–	–	–
GeoThinker 7B [48]	Qwen2.5-7B	68.5	–	–	–	–	–	–	–	–
GeoThinker 8B [48]	Qwen3-8B	72.6	–	–	–	–	–	–	–	–
Cambrian-*S*-7B† [111]	Qwen2.5-7B	69.2	73.6	53.7	75.2	74.7	71.5	82.0	38.7	84.3
Cambrian-*P*	Qwen2.5-7B	**73.7**	**74.9**	**60.1**	**76.0**	**76.9**	**74.8**	**89.5**	**52.6**	**85.0**

Table 1 | **VSI-Bench Results Comparison.** † indicates this Cambrian-*S* is fine-tuned only on VSI-590K.

Model	LM		Numerical Answer				Multiple-Choice Answer			
		Avg.	Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
Baselines										
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
General-purpose Models										
GPT-4o [41]	Unk.	34.0	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5
Gemini-2.5 Pro [25]	Unk.	51.5	43.8	34.9	64.3	42.8	61.1	47.8	45.9	71.3
Qwen2.5VL-7B [8]	Qwen2.5-7B	29.3	25.2	10.5	36.4	29.6	38.4	38.0	29.8	26.8
InternVL-3 8B [129]	Qwen2.5-7B	42.1	68.1	39.0	48.4	33.6	48.3	36.4	27.3	35.4
InternVL-3.5 8B [129]	Qwen3-8B	56.3	–	–	–	–	–	–	–	–
Qwen3-VL 8B [7]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	–	–
Spatial-specialist Models										
VST 7B [110]	Qwen2.5-7B	61.2	71.6	43.8	75.5	69.2	60.0	55.6	44.3	69.2
VLM-3R 7B [29]	Qwen2-7B	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	80.5	45.4	40.1
VG-LLM 8B [128]	Qwen2.5-7B	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	40.7	32.4	59.2
Cambrian-*S* 7B [111]	Qwen2.5-7B	67.5	73.2	50.5	74.9	72.2	[illegible]	76.2	41.8	80.1
SenseNova-SI 8B [17]	Qwen2.5-7B	68.7	–	–	–	–	–	–	–	–
GeoThinker 7B [48]	Qwen2.5-7B	68.5	–	–	–	–	–	–	–	–
GeoThinker 8B [48]	Qwen3-8B	[illegible]	–	–	–	–	–	–	–	–
Cambrian-*S*-7B† [111]	Qwen2.5-7B	69.2	73.6	53.7	75.2	74.7	71.5	82.0	38.7	84.3
Cambrian-*P*	Qwen2.5-7B	**73.7**	**74.9**	**60.1**	**76.0**	**76.9**	**74.8**	**89.5**	**52.6**	**85.0**

Largest gains on tasks requiring global understanding!

Table 3 | **Out-of-Distribution Generalization for Spatial and General VQA Benchmarks**. Cambrian-*P* is fine-tuned only on VSI-590K, without any in-distribution training data for benchmarks here.

Model	SparBench	MMSIBench	MMSIVideo	MindCube	MVBench	EgoSchema	Perception Test	Tomato
Cambrian-*P* (w/o Pose)	32.7	26.2	20.1	34.3	51.9	49.6	56.4	20.4
Cambrian-*P*	**35.9**	**28.0**	**22.9**	**38.4**	**53.5**	**52.5**	**58.4**	**26.7**

Trained only on VSI-590K.
Evaluated on benchmarks it was *never trained on*.

Pose-grounding is a general methodology.

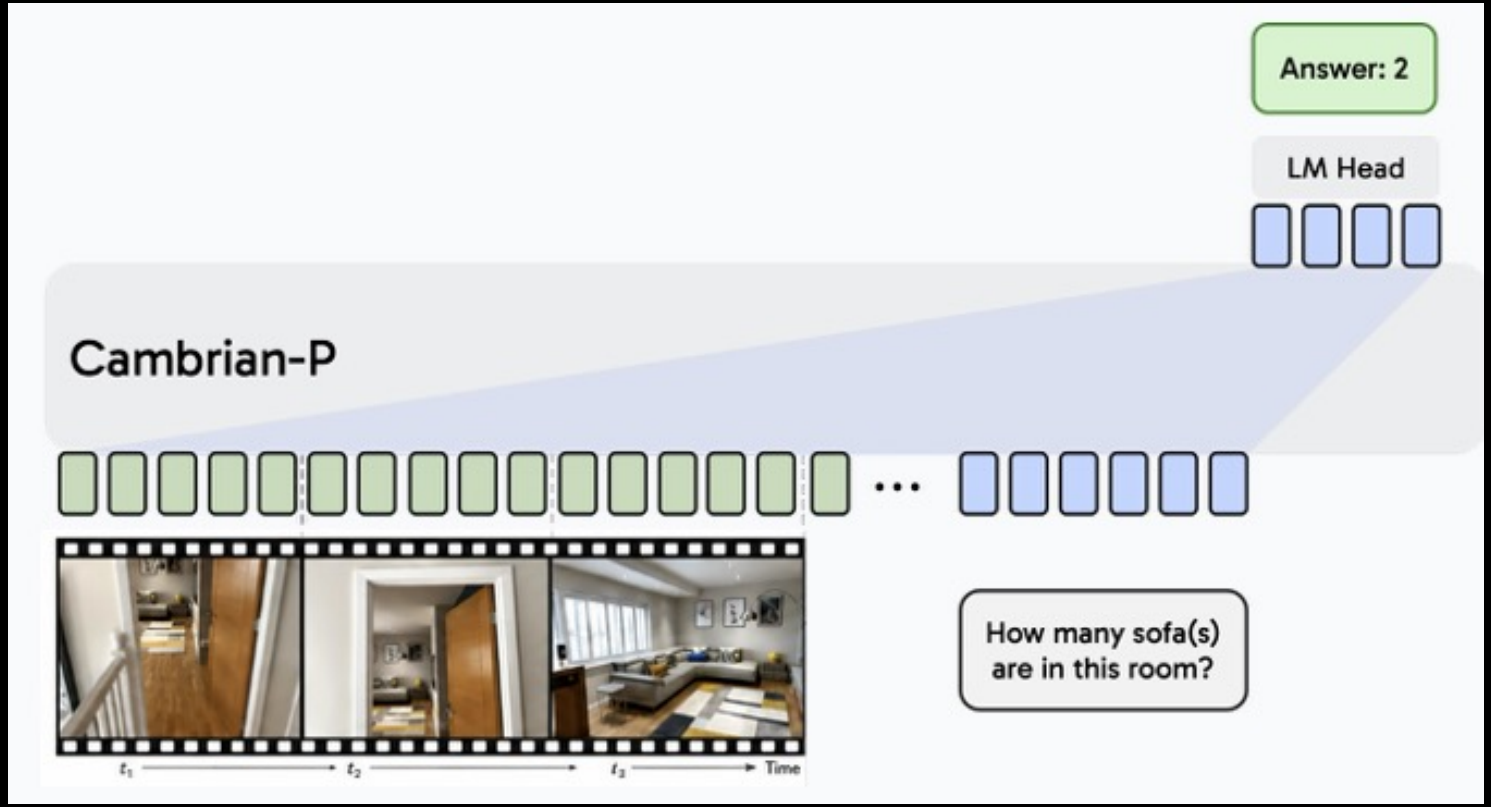
Scaling Further with Pseudo-Poses on In-the-Wild Video

Beyond GT pose data: pseudo-annotate Cambrian-$\mathcal{S}$-3M videos with VIPE.

Training Data	Pose Sup.	% Pose	VSIBench	MVBench	PercTest	EgoSchema
VSI-590K	–	0%	71.2	51.7	56.7	48.5
VSI-590K	GT	49%	73.7	53.8	58.1	51.3
VSI-590K + CamS-590K	–	0%	70.9	68.0	66.9	71.2
VSI-590K + CamS-590K	GT	25%	73.7	67.9	67.8	71.7
VSI-590K + CamS-590K	GT + Pseudo	48%	**73.9**	**69.3**	**67.9**	**73.6**

- Pseudo poses from noisy in-the-wild videos also help
- Pose supervision is scalable — no special data collection required

Standard Inference Pipeline



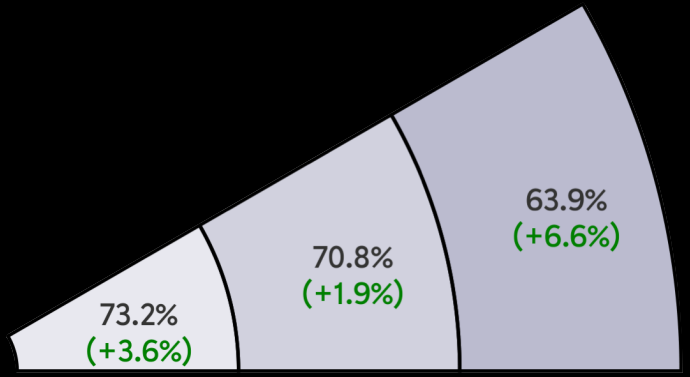
Pose Shapes the *Representation*, Not the Inference

Pose Token Train	Pose Token Infer	VSI	VSTI	MVBench	EgoSch.	PercTest	Tomato
✗	✗	67.3	55.1	51.9	49.6	56.4	20.4
✓	✓	**72.0**	56.5	**53.5**	**52.5**	58.4	**26.7**
✓	✗	**72.0**	**56.6**	53.2	**52.5**	**58.8**	**26.7**

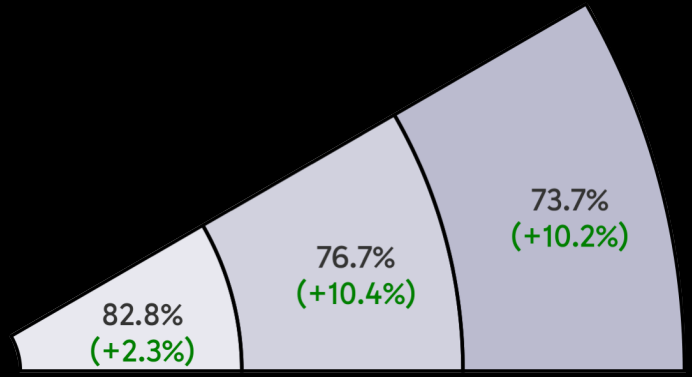
We don't need camera token during inference

- Pose tokens force the LLM's hidden states to encode cross-frame geometric structure.
- That structure remains in the representation.
- The model never explicitly "uses" predicted poses — pose is a training-time *scaffold*

Pose Enables More *Global* Spatial Reasoning



Relative Distance



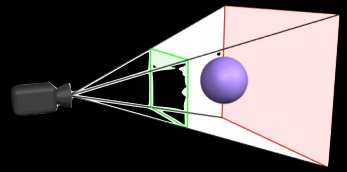
Relative Direction

- Model Performance degrades as objects get farther apart
- Camera Pose brings larger gains for distant objects than nearby ones

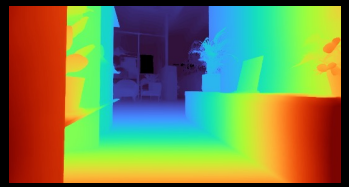
Pose > Depth: Better Global Signal

Pose	Depth	VSI-Bench ↑
✗	✗	67.3
✓	✗	**72.0**
✓	✓	71.7
✗	✓	69.4

Camera Pose: 9 numbers per frame, global



- Depth: dense per-pixel, local



Takeaway

Adding more 3D supervision (depth on top of pose) doesn't help.

Thank You!

Website: cambrian-mllm.github.io

Code: github.com/cambrian-mllm/cambrian-p

Models: huggingface.co/nyu-visionx