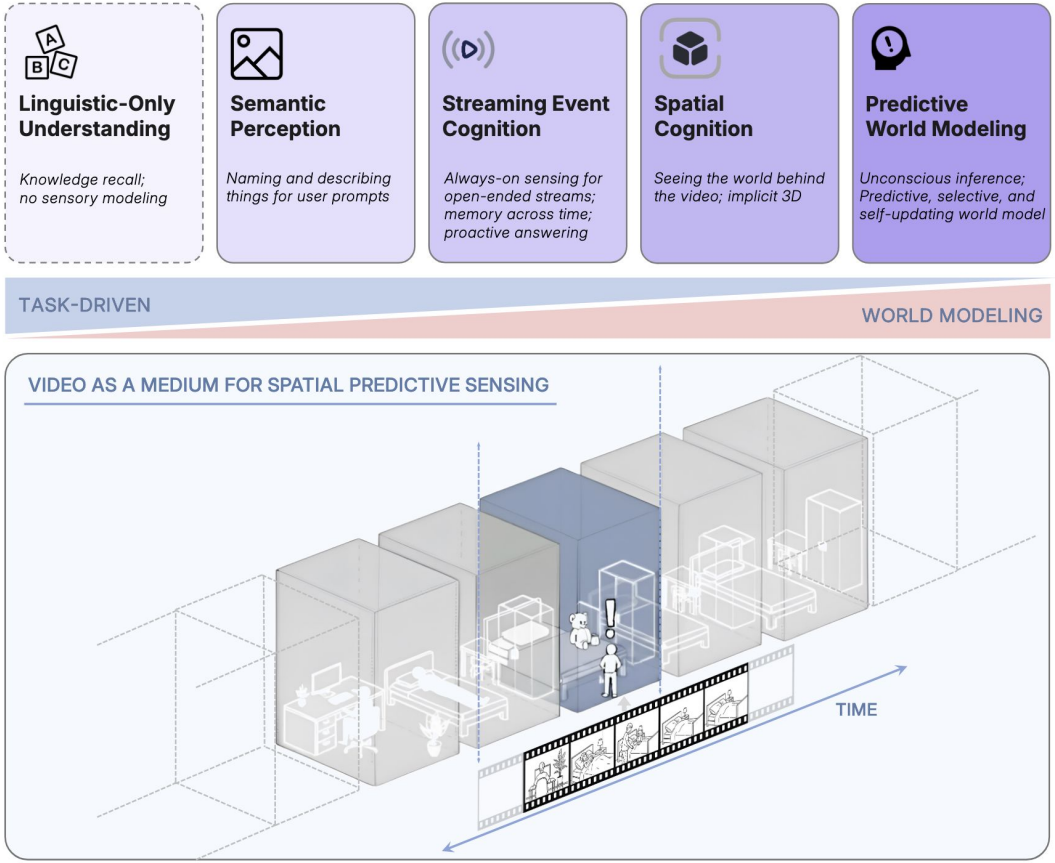
Cambrian-S: Towards Spatial Supersensing in Video

Yann LeCun, Saining Xie's Team and Li-Fei-Fei's Lab

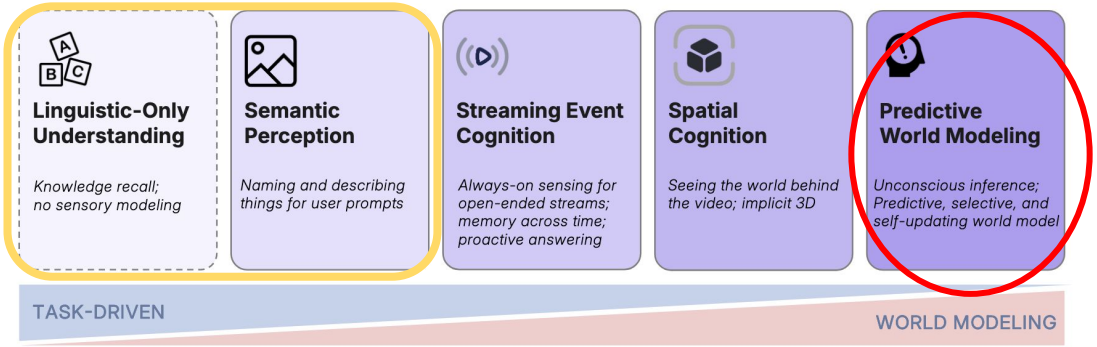
(published in Nov. 2025.)

From pixel to predictive mind

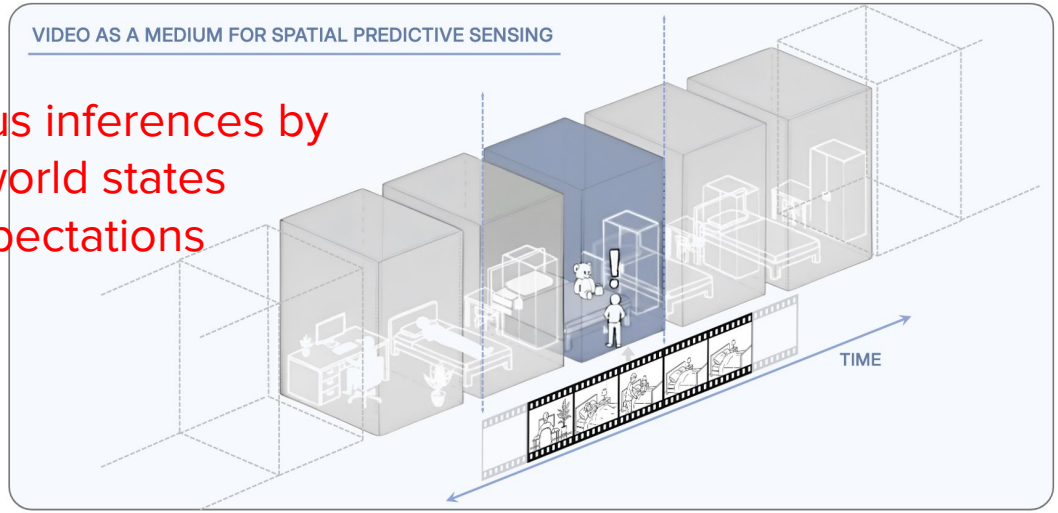


From pixel to predictive mind

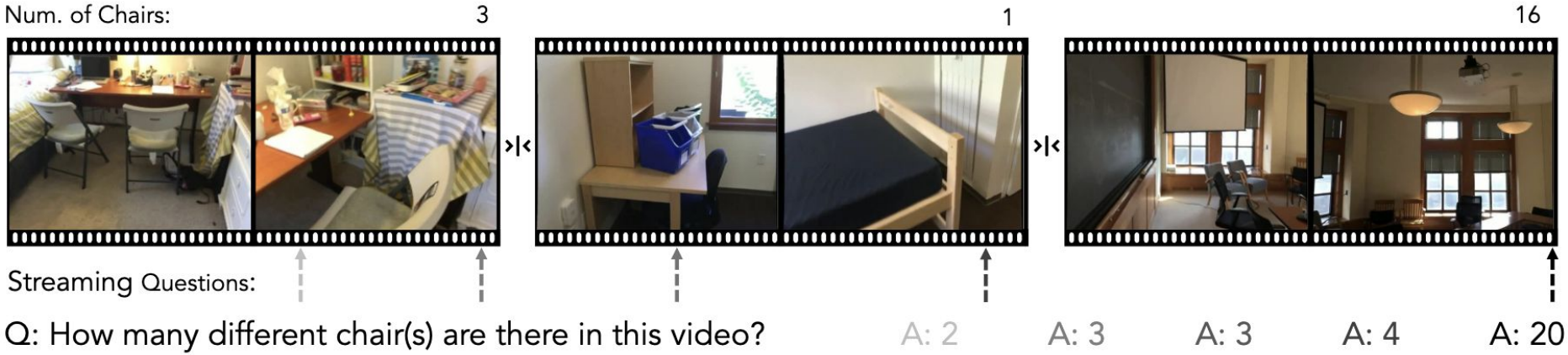
We are still in here now



Makes unconscious inferences by predicting latent world states based on prior expectations



What do you think of VLM's capabilities?



Current models fail when the question depends on accumulated evidence or long-horizon state updates

Benchmark issue?

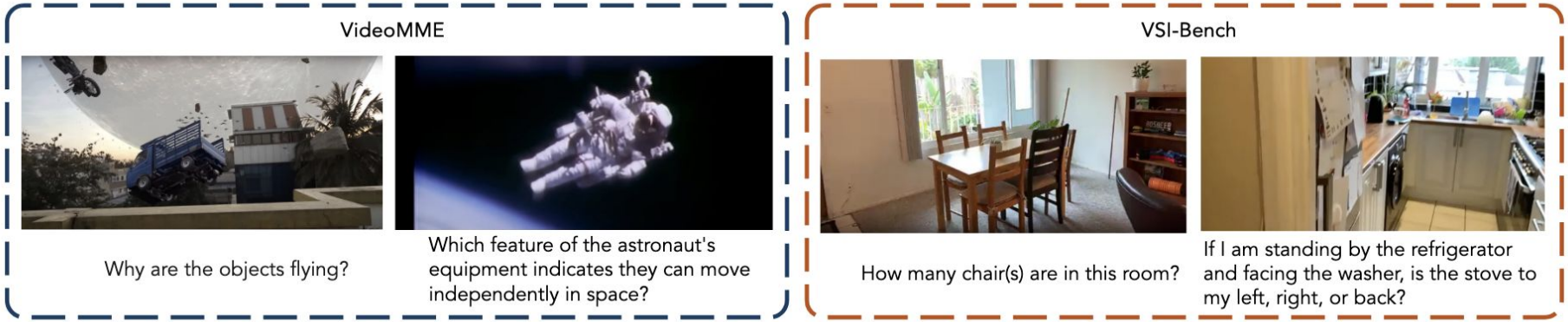


Figure 3 | **Illustrations of how spatial sensing is conceptualized in current video benchmarks**. The left panel features examples from the "spatial reasoning" subcategory of VideoMME [42], including a question regarding gravity from Shutter Authority's "`What if the Moon Crashed into the Earth?`" and a question regarding astronaut gear from NASA's "`Astronaut Bruce McCandless II Floats Free in Space.`" In contrast, the right panel shows samples from VSI-Bench [148], which highlight visual-spatial reasoning tasks such as object counting, identifying relative directions, route planning, and more.

New benchmark tasks

- VSR (VSI-Super Recall):
  - Long-horizon visual spatial recall over arbitrarily long video
- VSC (VSI-Super Counting):
  - Continual visual spatial counting across changing scenes and viewpoints

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

Is it just a data problem?

- Hypothesis: maybe current models are weak simply because they lack enough high-quality spatially grounded training data.

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.

Table 1 | **Gemini-2.5-Flash results.** As a state-of-the-art video understanding model with long-context capabilities, Gemini demonstrates strong performance on general video benchmarks but shows clear limitations towards spatial supersensing.

VSI-590K dataset

- Question types:
  - count, size, direction, ...
  - measure distance, ...
  - spatiotemporal (route planning, ...)

Dataset	# Videos	# Images	# QA Pairs
Annotated Real Videos			
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
Simulated Data			
ProcTHOR [36]	625	-	20,092
Hypersim [113]	-	5,113	176,774
Unannotated Real Videos			
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

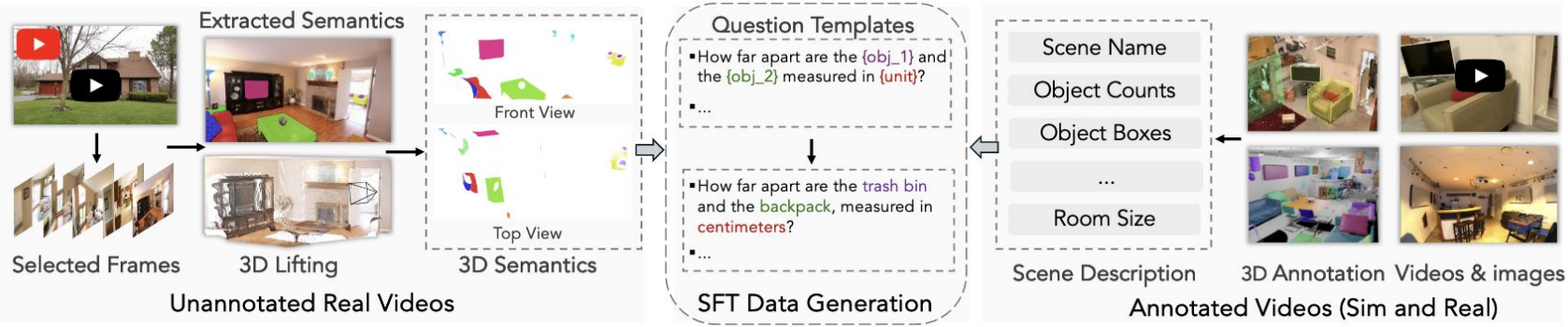


Figure 7 | **VSI-590K data curation pipeline.** We collect data from 3D-annotated real and simulated video sources, as well as from pseudo-annotated frames extracted from web videos. We then use diverse templates to automatically generate question–answer pairs for instruction tuning.

Data scaling findings

Model	VSI-Bench	VideoMME	EgoSchema	Perception Test
Different Base Models				
A1 (*w/o.* I-IT, *i.e.* QwenLM)	21.4	44.2	42.9	44.5
A2 (A1 + I-IT, *i.e.* 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 *w/.* VSI-590K				
from A1	57.2	40.3	38.7	52.3
from A2	66.8	46.7	47.2	52.3
from A3	68.8	52.3	48.4	55.8
from A4	**69.2**	**54.1**	**55.2**	**59.2**
SFT *w/.* 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**

Data scaling findings

Methods	Avg.	Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
		Numerical Answer				Multiple-Choice Answer			
Statistics									
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
Proprietary Models (API)									
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
Open-source Models									
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

Data scaling findings

Are the performance improvements simply due to **broader data coverage** (including more diverse visual configurations and question–answer pairs), or has the model actually **developed stronger spatial cognition**?

		Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
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
Open-source Models									
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

Test on VSR and VSC

Table 7 | **Cambrian-*S*-7B results on VSI-SUPER.** Despite strong performance on VSI-Bench, accuracy on VSR drops sharply from 38.3% (10 min) to 0.0% (>60 min), and VSC completely fails. Note that VSI-SUPER focuses on continual, streaming evaluation, where uniform sampling 128 frames across the entire video does not align with the online setting; results shown in gray are provided for reference only.

	VSR					VSC			
Eval Setup	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

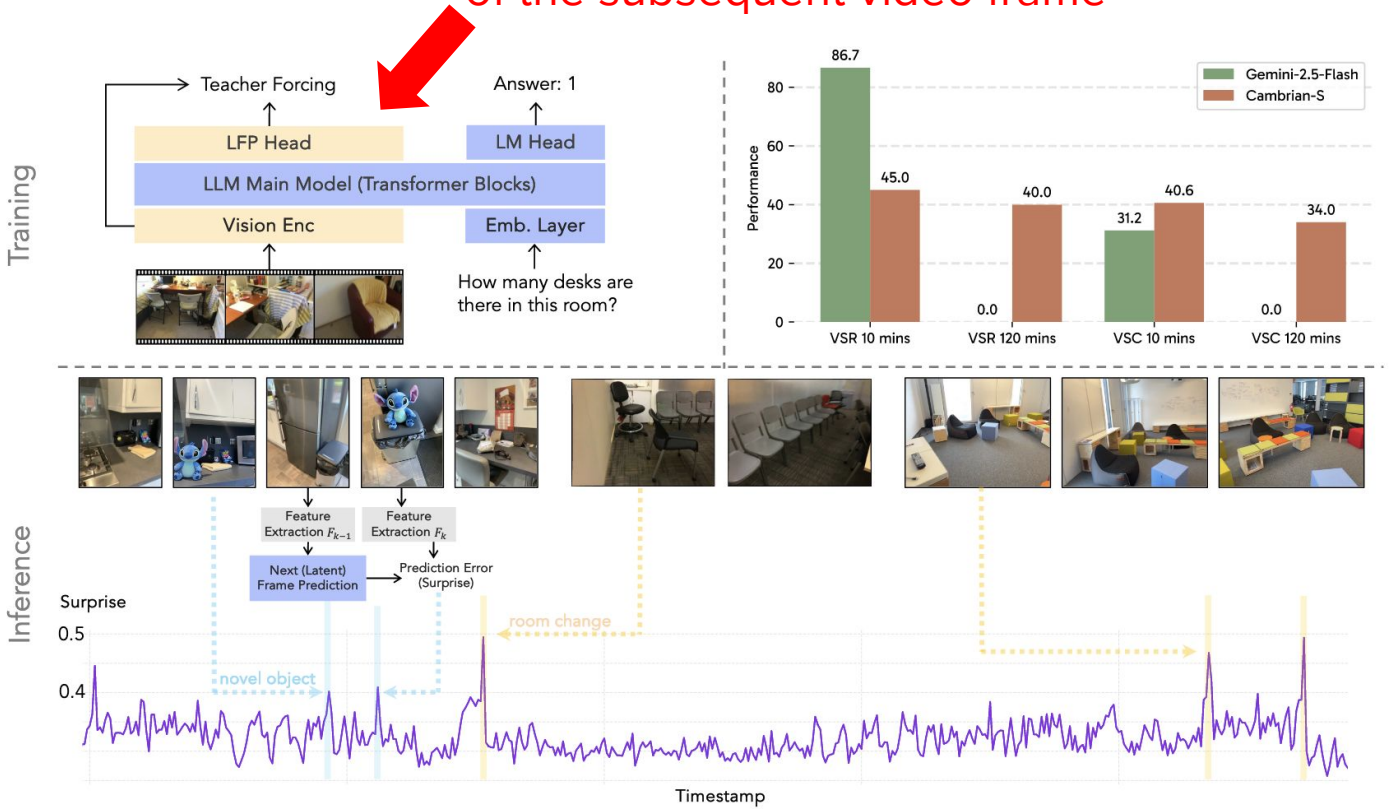
Not only a data issue!

Cambrian-S model

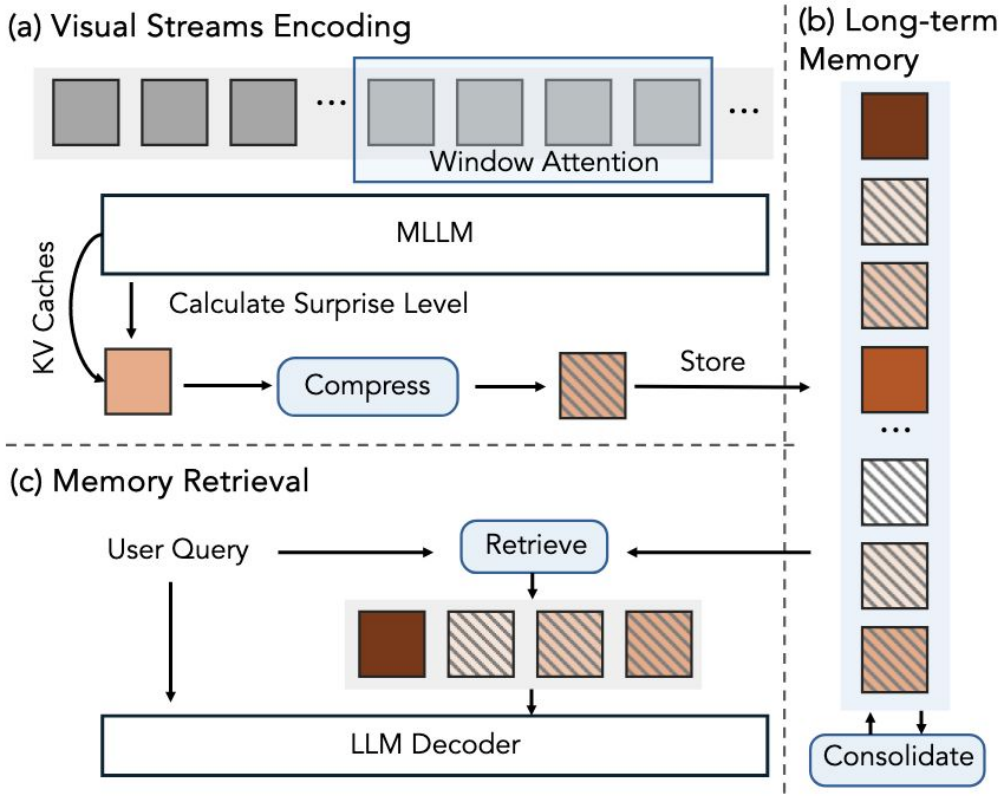
- They propose “predictive sensing” as a path forward, where models learn to anticipate their input and construct internal world models to handle unbounded visual streams.

Cambrian-S

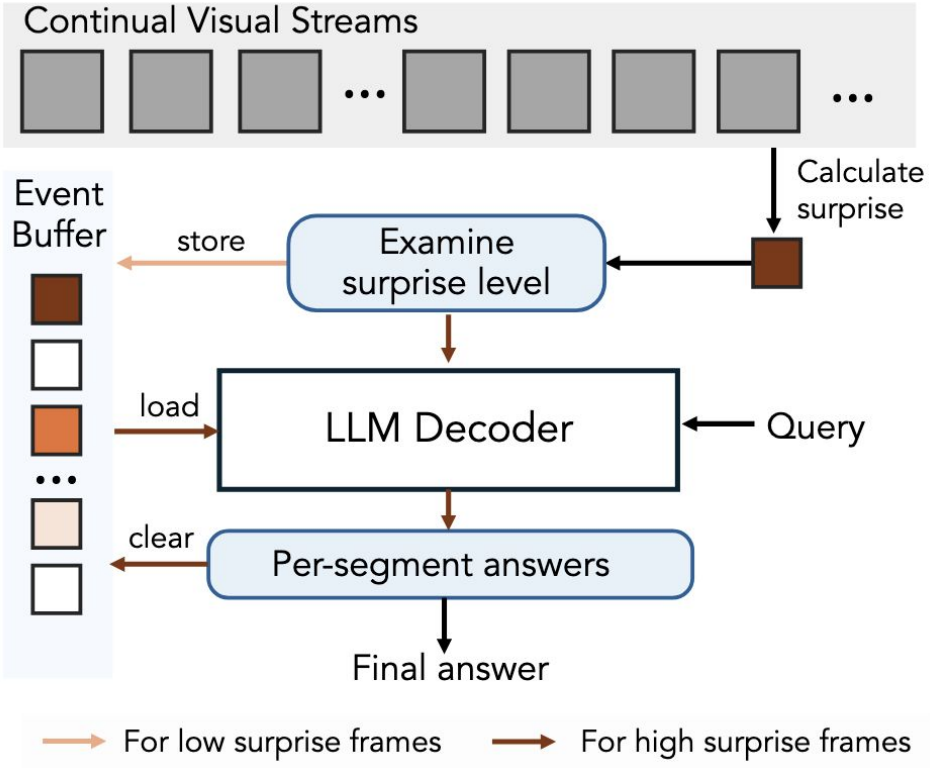
Predict the latent representation of the subsequent video frame



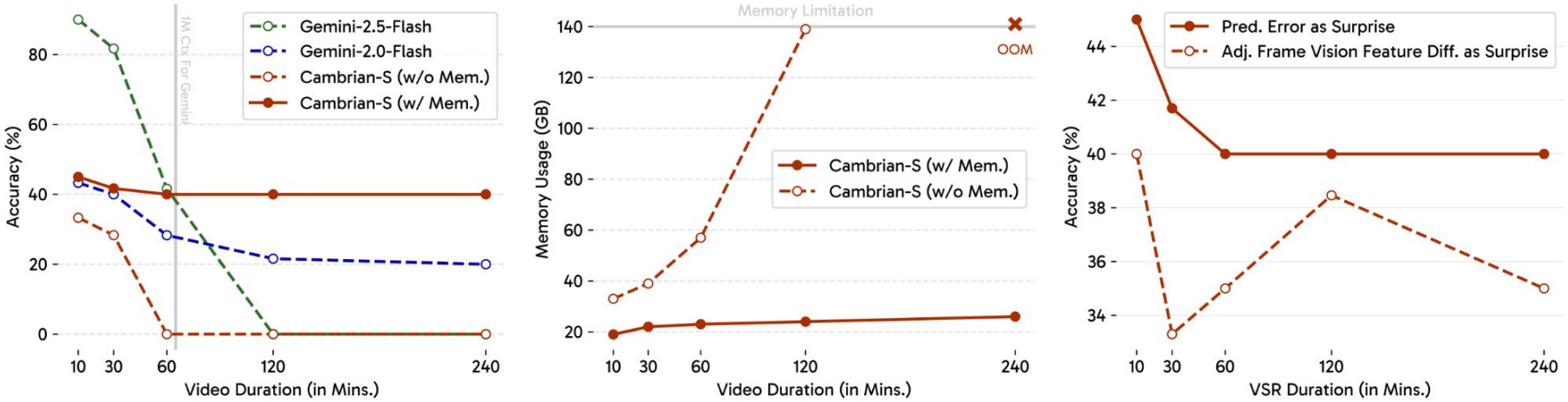
Case study: Surprise-driven memory management system (VSR)



Case study: Surprise-driven continual video segment (VSC)



VSR Results



(a) VSR results (b) GPU memory usage (c) Surprise comparison

Figure 11 | **Performance analysis of surprise-driven memory on VSR.** (a) Surprise-driven memory allows Cambrian-*S* to maintain strong performance as video length increases. (b) Surprise-driven memory maintains a stable GPU memory footprint as video length increases. (c) Ablation: Using LFP prediction error as the surprise signal is more robust and consistently outperforms using adjacent-frame similarity.

VSC Results

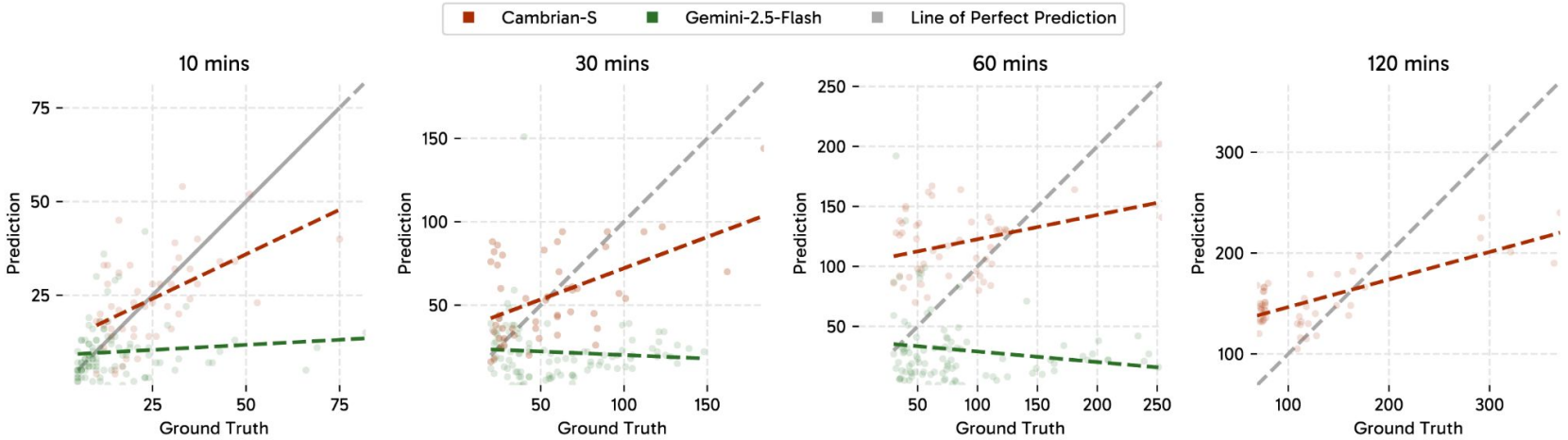


Figure 14 | **Cambrian-*S* scales to higher ground truth object counts whereas Gemini saturates.** Predicted counts are plotted against ground-truth counts for videos of different lengths (10, 30, 60, and 120 minutes). Using surprise-driven segmentation, Cambrian-*S*'s predicted counts grow approximately linearly with the ground-truth, tracking the $y = x$ perfect-count line (gray dashed), whereas Gemini-2.5-Flash's predicted counts remain clustered near small values and fail to increase with ground-truth count, indicating early saturation and poor extrapolation to larger counts.

Related Work

- https://cambrian-mllm.github.io/
- World Models Reading List

Publication Nov 7, 2025	Cambrian-S	Towards Spatial Supersensing in Video
Publication Nov 7, 2025	Test-set Stress-Test	Benchmark Designers Should "Train on the Test Set" to Expose Exploitable Non-Visual Shortcuts
Publication Nov 7, 2025	SIMS-V	Simulated Instruction-Tuning for Spatial Video Understanding
Publication Dec 18, 2024	Thinking in Space	How Multimodal Large Language Models See, Remember and Recall Spaces
Publication Jun 24, 2024	Cambrian-1	A Fully Open, Vision-Centric Exploration of Multimodal LLMs

LeWorldModel: Stable End-to-End Joint-Embedding Predictive Architecture from Pixels

Lucas Maes[*1] **Quentin Le Lidec**[*2] **Damien Scieur**[1,3] **Yann LeCun**[2] **Randall Balestriero**[4]

[1]Mila & Université de Montréal [2]New York University [3]Samsung SAIL [4]Brown University

Website Code

Abstract

Joint Embedding Predictive Architectures (JEPAs) offer a compelling framework for learning world models in compact latent spaces, yet existing methods remain fragile, relying on complex multi-term losses, exponential moving averages, pre-trained encoders, or auxiliary supervision to avoid representation collapse. In this work, we introduce LeWorldModel (LeWM), the first JEPA that trains stably end-to-end from raw pixels using only two loss terms: a next-embedding prediction loss and a regularizer enforcing Gaussian-distributed latent embeddings. This reduces tunable loss hyperparameters from six to one compared to the only existing end-to-end alternative. With 15M parameters trainable on a single GPU in a few hours, LeWM plans up to 48× faster than foundation-model-based world models while remaining competitive across diverse 2D and 3D control tasks. Beyond control, we show that LeWM's latent space encodes meaningful physical structure through probing of physical quantities. Surprise evaluation confirms that the model reliably detects physically implausible events.