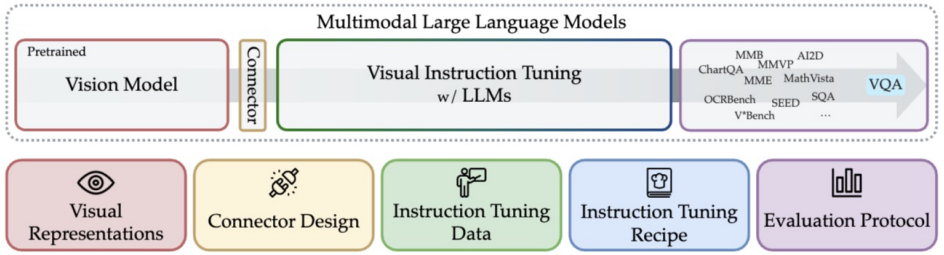
Cambrian-s

Towards Spatial Supersensing in Video

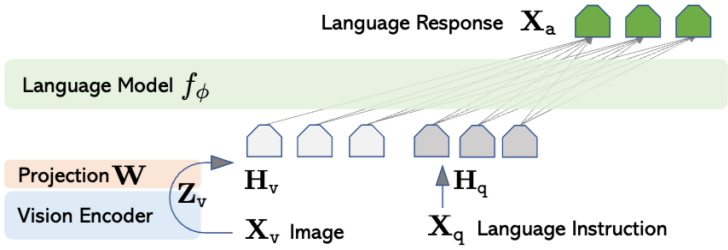
251210

Motivation

- MLLM Architecture



Protocol of MLLM



LLaVA network architecture

- **MLLM = LLM + Visual Extractor.**
- **Current MLLM considers video as sparse frames, does not fully express spatial structure and dynamics, and relies heavily on textual recall.**

Motivation

- Task Example (VQA)



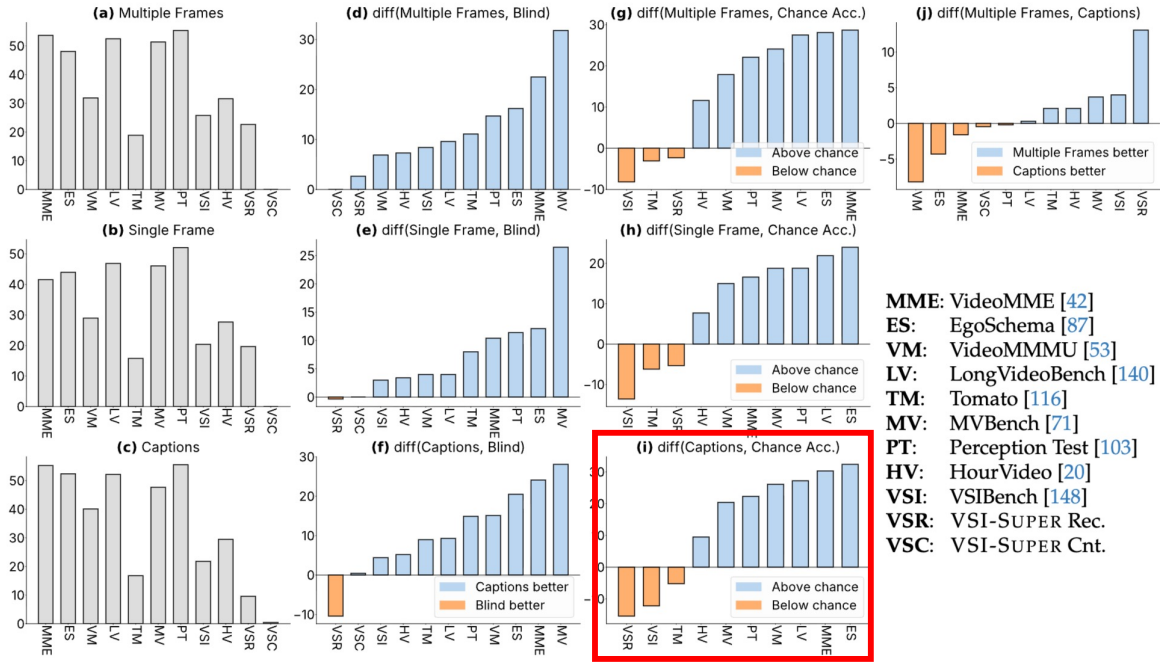
Which of the following correctly represents the order in which the Hello Kitty appeared in the video?

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

The main purpose is to maximize the ability to **view and answer** questions properly

Motivation

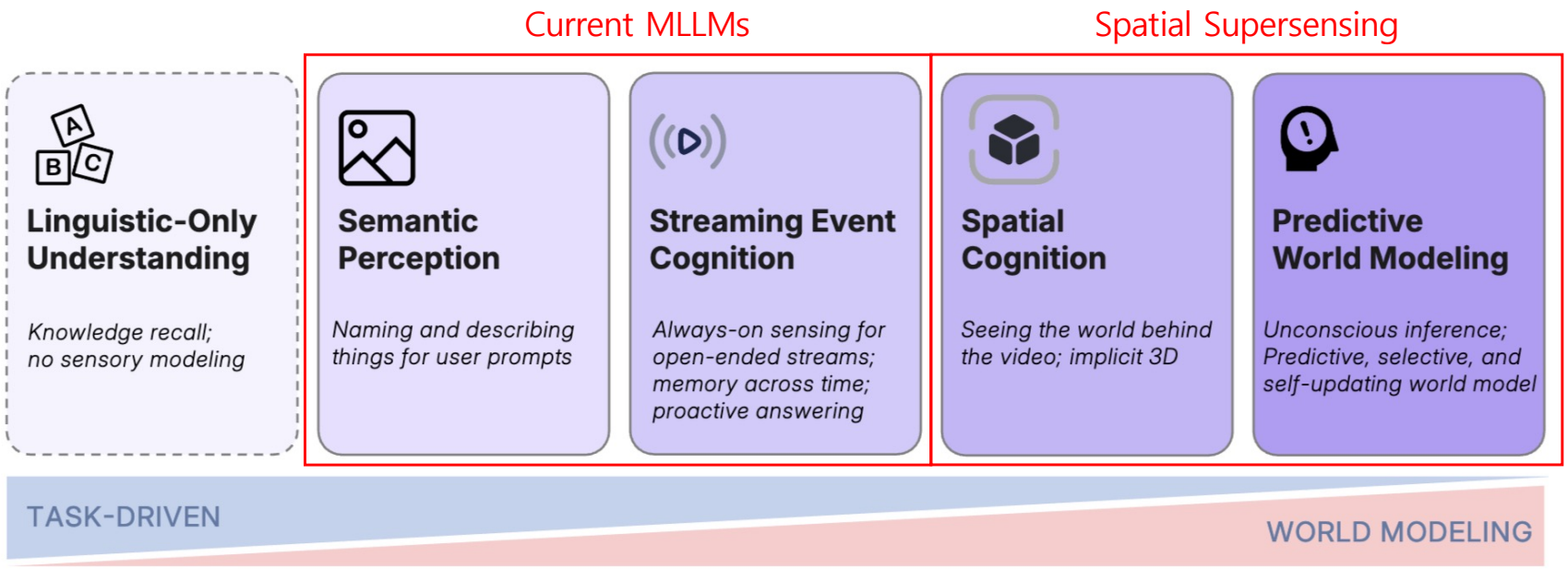
Problem 1 - Inadequacy of Benchmarks



Benchmarks may not significantly depend on visual input and rather heavily rely on the base LLM

Motivation

Problem 2 -Limitations of "Reactive" Modeling



Goal: Need to move from **Level 2 (Streaming Event)** to **Level 4 (Predictive World Modeling)**

Spatial Supersensing = Understand, predict, and actively regulate cognitive processes within 3D spaces

Summary

- Core Question
  - Can the current MLLM paradigm reach **'True multimodal intelligence'**?
  - **How to achieve 'Spatial Supersensing'?**

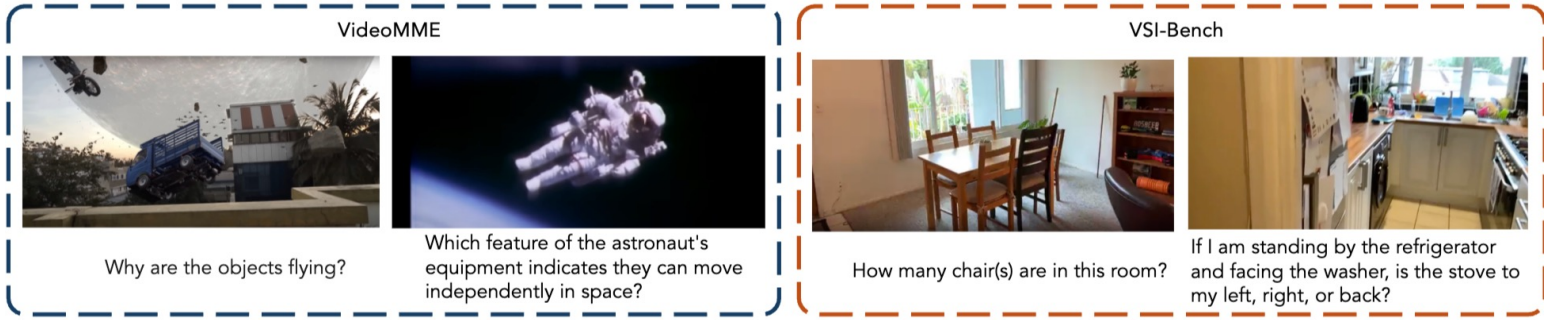
- Contributions
  - Conceptualization of **Spatial Supersensing** and Hierarchy Presentation
  - Development of **VSI-SUPER** Benchmark
  - Demonstrate the limitations of the current paradigm (via Cambrian-S)
  - Proposition of **Predictive Sensing**

Contents

- Benchmarking Spatial Supersensing
  - To define and establish what true **"spatial supersensing"** actually entails and to create the necessary tools **(VSI-SUPER)** to accurately measure it,
- Spatial Sensing Under the Current Paradigm
  - To rigorously **test the limits of current MLLM** capabilities (data/model scaling) in spatial sensing
- Predictive Sensing as a New Paradigm
  - To propose and provide **proof-of-concept** for a novel, "world model"-inspired approach

Benchmarking Spatial Supersensing

- VSI-SUPER: Towards Benchmarking Spatial Supersensing in MM-LLMs



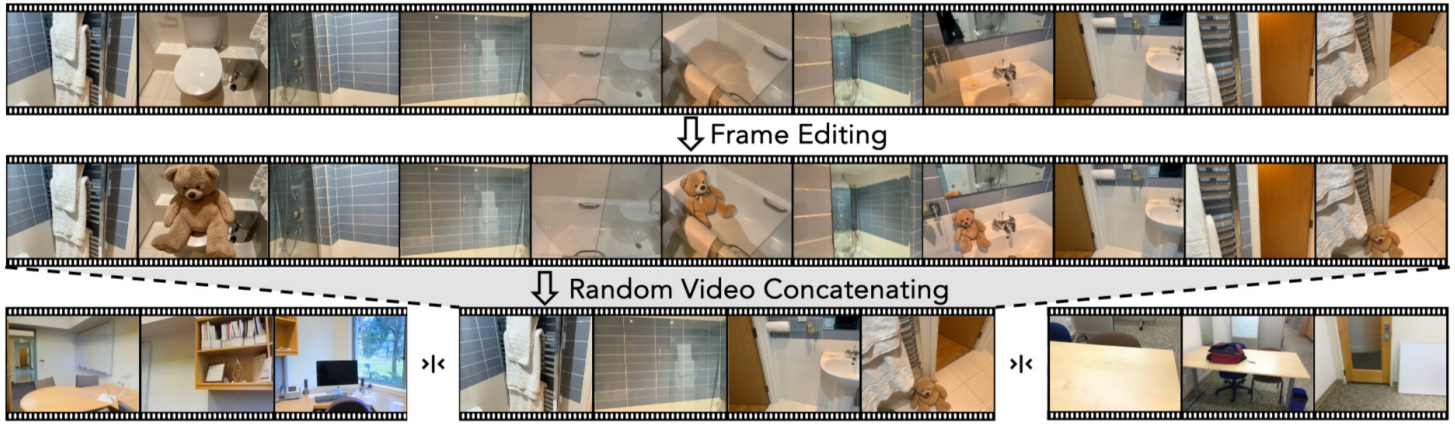
Illustrations of how spatial sensing is conceptualized in current video benchmarks.

- Limitation of existing Video Benchmark:
  - QAs can be answered without visual information

- **VSI-SUPER**: a two-part benchmark for **continual spatial sensing (VSR + VSC)**
- Require **selective filtering** and **structured accumulation of visual information** across unbounded spatial videos
  - **Resistant to brute-force context expansion**

Benchmarking Spatial Supersensing

- VSR Benchmark

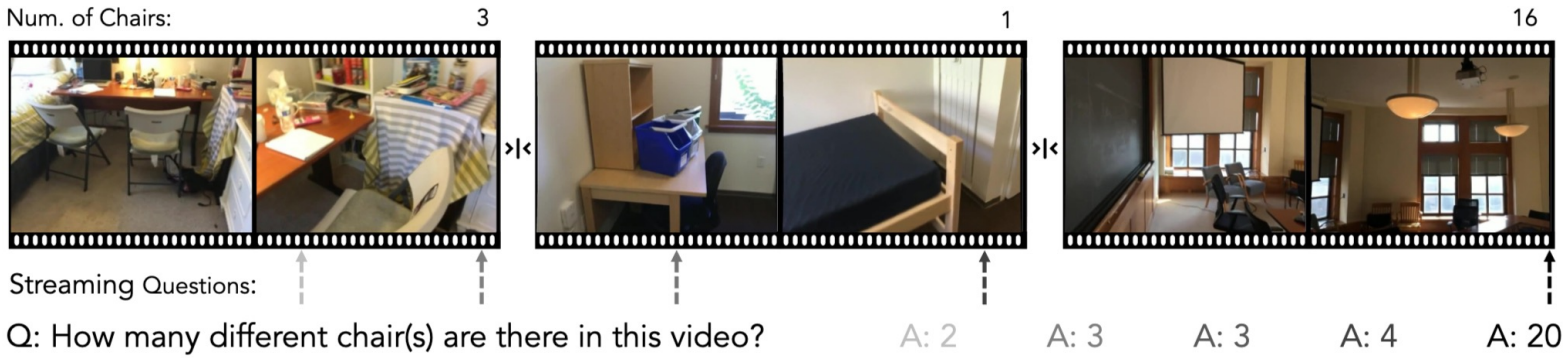


"Recall and remember spatial information in chronological order"

Video Length: 10 / 30 / 60 / 120 / 240 min

Benchmarking Spatial Supersensing

- VSC Benchmark



Test Accumulative Counting Ability on long video

Video Length: 10 / 30 / 60 / 120 min

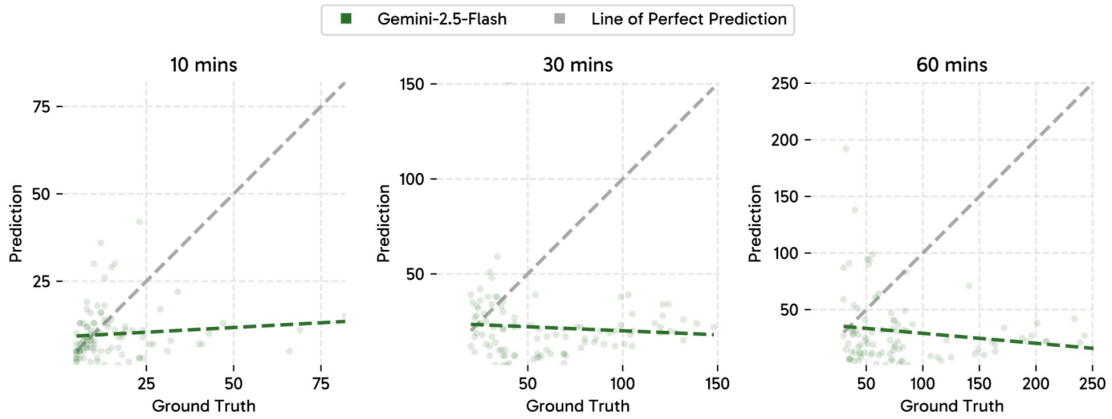
It consists of tasks that require **space division** and **prevention of duplicate counting**

Benchmarking Spatial Supersensing

- Result (Gemini-2.5-Flash)

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.

Gemini-2.5-Flash results



Visualization of Gemini-2.5-Flash's predictions v.s. ground truth on VSC

"Scaling Context Window is NOT the solution for infinite streams."

Benchmarking Spatial Supersensing

- How VSI-SUPER challenges the current paradigm

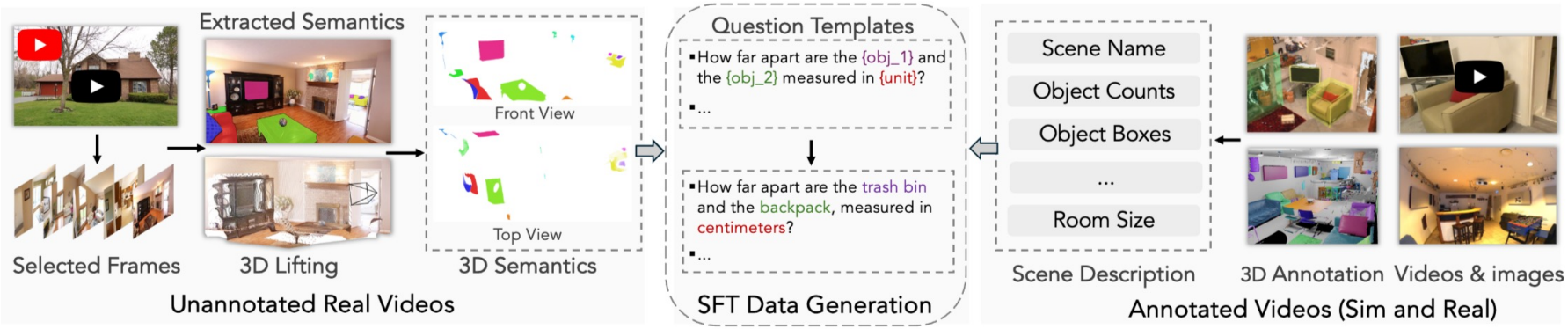
1. Limitations of Scaling
2. Need for **selective cognition**: Only remember information that is subconsciously important
    - Need for predictive and Screening mechanism
3. **True Generalization: "Process" over "Pattern”**
    - VSI-SUPER needs to understand the "Process" itself to keep adding numbers no matter how long the video gets.

Spatial Sensing Under the Current Paradigm

- VSI-590K (New Training Dataset)

Why build VSI-590K:

- **The Data Gap:** Existing video instruction datasets focus on general descriptions, lacking high-quality **3D spatial annotations** (e.g., distance, size, coordinates) .
- **Goal:** To enable conventional MLLMs to ground their linguistic capabilities in the physical 3D world.



VSI-590K data curation pipeline

Spatial Sensing Under the Current Paradigm

- VSI-590K (New Training Dataset)

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"

Taxonomy of spatiotemporal question types in VSI-590K

"VSI-590K generates diverse 3D spatial QA pairs that span both **absolute measurements** and **relative comparisons** across key spatial attributes like size, distance, and direction, thereby addressing the lack of fine-grained spatial grounding in existing datasets."

Spatial Sensing Under the Current Paradigm

- VSI-590K (New Training Dataset)

VSI Data Mixture	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
Baseline (Cambrian-1)	52.7	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	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	50.6	56.5	77.5	41.0	51.0	29.8	52.5	40.2	42.3	38.8	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
Simulated Videos												
+ 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
Pseudo-Annotated Images												
+ 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**

Contributions of Different Data Sources in the VSI-590K Mixture

- From 10 Data sources (Real / Simulated Videos + pseudo-annotated frames)
- Effectiveness of Data source ablation
  - Annotated real videos > simulated data > pseudo-annotated images.

Spatial Sensing Under the Current Paradigm

- Post Training for Spatial Sensing

Greater Exposure to general video data

General Video Benchmark

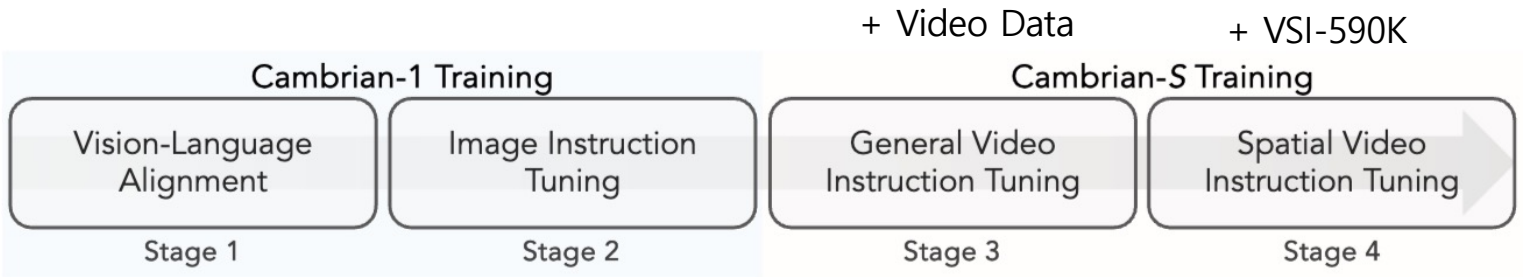
Model	VSI-Bench	VideoMME	EgoSchema	Perception Test
Different Base Models				
A1 (*w/o.* I-IT, *i.e.* QwenLM)	21.4	44.2	42.9	44.5
A2 (A1 + I-IT, *i.e.* Cambrian-1)	25.8	53.7	48.1	55.4
A3 (A2 + V-IT, 429K data)	28.9	61.2	50.3	66.3
A4 (A2 + V-IT, 3M data)	**35.7**	**62.6**	**77.0**	**70.9**
SFT *w/.* VSI-590K				
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**

Post-training exploration for spatial sensing

1. **Importance of Base Model (Stronger Base = Better Spatial Sensing)**
2. **Data Mixture Strategy (Domain-Specific + General)**

Spatial Sensing Under the Current Paradigm

- Cambrian-S: Spatially-Grounded MLLMs



Overall Cambrian-S training pipeline

Spatial Sensing Under the Current Paradigm

- Empirical Results: Improved Spatial Cognition

Model	Base LM	VSI-Bench	VSI-Bench^{Debiased}	Video								Image		
				Tomato	HourVideo	Video^{MME}	EgoSchema	Video^{MMMU}	LongVBench	MVBench	Percept. Test	MMVP	3DSR	CV-Bench
Proprietary Models														
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	-	-
Open-Source Models														
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**
VILA1.5-3B	Sheared-LLaMA-2.7B	-	-	-	-	42.2	-	-	42.9	-	49.1	-	-	-
Qwen2.5-VL-3B	Qwen2.5-3B	26.8	22.7	-	-	**61.5**	-	-	**54.2**	-	**66.9**	39.3	-	-
Cambrian-S-3B	Qwen2.5-3B	**57.3**	**49.7**	**25.4**	**36.8**	60.2	**73.5**	**25.2**	52.3	**60.2**	65.9	**50.0**	**50.9**	**75.2**
SmolVLM2-2.2B	SmolLM2-1.7B	27.0	22.3	-	-	-	34.1	-	-	48.7	51.1	-	-	-
InternVL2.5-2B	InternLM2.5-1.8B	25.8	20.7	-	-	51.9	47.4	-	52.0	**68.8**	-	**45.3**	-	-
InternVL3.5-2B	Qwen3-1.7B	51.5	46.1	-	-	**58.4**	50.8	-	**57.4**	65.9	-	44.0	-	-
Cambrian-S-1.5B	Qwen2.5-1.5B	**54.8**	**47.5**	**22.5**	**31.4**	55.6	**68.8**	24.9	50.0	58.1	**63.2**	42.7	**51.9**	**69.6**
SmolVLM2-0.5B	SmolLM2-360M	26.1	23.1	-	-	-	20.3	-	-	43.7	44.8	-	-	-
LLaVA-One-Vision-0.5B	Qwen2-0.5B	28.5	20.6	-	-	44.0	26.8	-	45.8	45.5	49.2	28.7	-	55.5
InternVL2.5-1B	Qwen2.5-0.5B	22.5	17.5	-	-	50.3	39.8	-	47.9	**64.3**	-	**33.3**	-	-
InternVL3.5-1B	Qwen3-0.6B	49.9	41.8	-	-	**51.0**	41.5	**33.0**	**53.0**	61.0	-	32.0	-	-
Cambrian-S-0.5B	Qwen2.5-0.5B	**50.6**	**42.2**	**23.4**	**27.9**	44.0	**62.4**	15.7	44.0	51.8	**56.0**	26.0	**48.5**	**59.8**

Comparison of Cambrian-S with other leading MLLMs

Core Question:

- Performance improvements simply due to **broader data coverage**?
- Or has the model actually **developed stronger spatial cognition**?

Robust spatial reasoning on VSI-Bench-Debiased:

- A benchmark specifically designed to eliminate language shortcuts through debiasing
- Cambrian-S-7B still outperform proprietary counterparts

Spatial Sensing Under the Current Paradigm

- Empirical Results: Improved Spatial Cognition

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

VSI-Bench sub-task breakdown

1. Generalization in spatial reasoning: Cambrian-S-7B still performs strongly,
2. The training approach proves effective even with smaller model sizes
   - The smallest 0.5B model achieves performance comparable to Gemini-1.5 Pro on VSI-Bench

Spatial Sensing Under the Current Paradigm

- Empirical Results: Results on VSI-Super

	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

Cambrian-S-7B results on VSI-SUPER

Limitation 1

- Scores drop steadily from 38.3% to 6.0% as video length increases from 10 to 60 minutes

Limitation 2

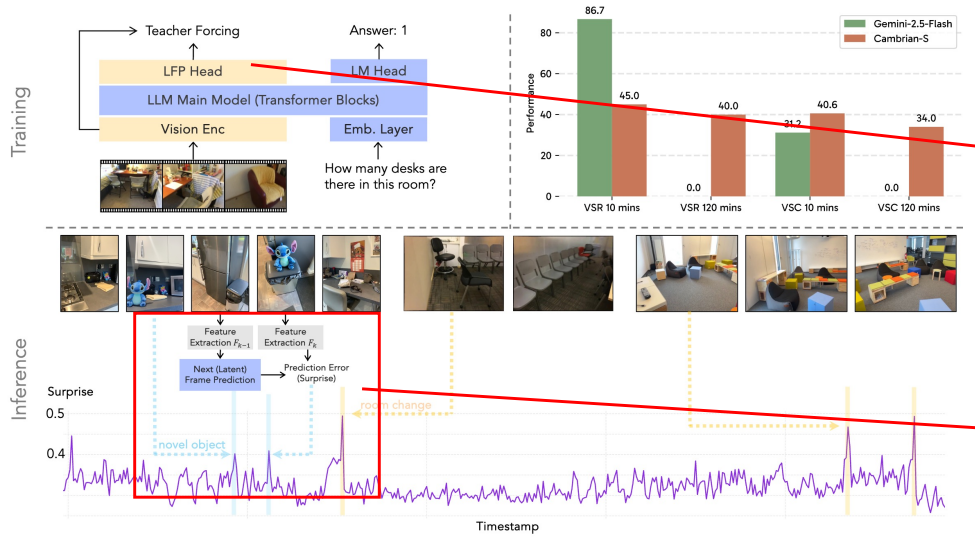
- The model has difficulty generalizing to new test scenarios
- Although trained on multi-room house tour videos, it fails to handle unseen examples with just a few additional rooms.

A Purely data-driven approach within the current MLLM framework, no matter how much data or engineering effort is invested, faces fundamental limits
→ Need For *Predictive Sensing*

Predictive Sensing as a New Paradigm

- Predictive Sensing via Latent Frame Prediction

Predictive Sensing: An active sensory processing method in which AI predicts the future and utilizes the difference between prediction and reality ("surprise") to selectively recognize and organize important information



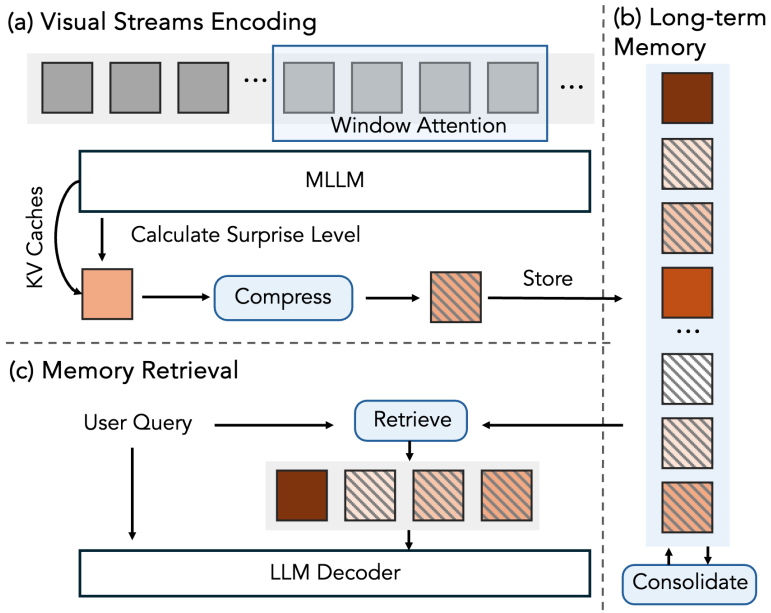
Training and inference pipeline for the latent frame prediction (LFP) approach

LFP Head: a two-layer MLP, that operates in parallel with the LFP head, to predict latent representation of the subsequent video frame

- Auxiliary losses: MSE(Mean Square Error) & Cosine Distance
  - Measure the discrepancy between the **predicted latent feature** and the **ground truth feature** of the next time

Predictive Sensing as a New Paradigm

- Case Study I: Surprise-driven Memory Management System for VSI-SUPER Recall

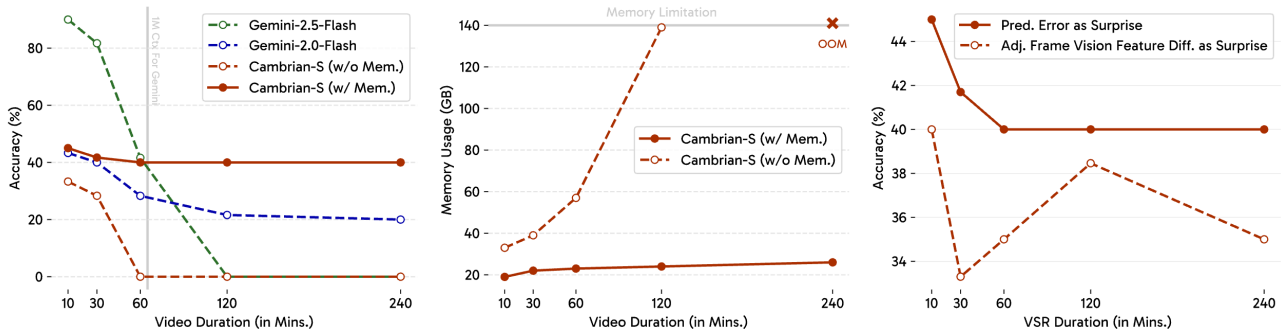


Surprise-driven memory management framework design

- **Goal:** Dynamic memory management for efficient processing of long video streams.
- **Mechanism:**
  - **Latent Frame Prediction:** Predicts the latent features of the subsequent frame.
  - **Surprise Calculation:** Quantifies a 'Surprise' score based on the difference between the actual and predicted frames.
- **Process:**
  - **Compression:** Frames with low surprise scores undergo 2× compression before being stored
  - **Consolidation:** Maintains a fixed memory budget by dropping or merging frames with the lowest surprise scores when full.
- **Retrieval:** Retrieves the Top-K relevant frames based on cosine similarity between the user query and stored memory features.

Predictive Sensing as a New Paradigm

- Case Study I: Surprise-driven Memory Management System for VSI-SUPER Recall



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

Performance analysis of surprise-driven memory on VSR

- **Surprise-based memory** effectively compresses redundant data without losing critical information.

Predictive Sensing as a New Paradigm

- Case Study II: Surprise-driven continual video segment for VSC

- **Definition:**
  - **Event:** A spatiotemporally coherent segment of experience (e.g., staying in one room) .
  - **Segmentation:** Parsing continuous sensory input into discrete, meaningful units based on changes in coherence.

- **Cognitive Inspiration: The "Doorway Effect"**
  - Humans naturally reset their working memory when entering a new environment (boundaries of experience)
  - This allows efficient processing without being overwhelmed by the past.

Predictive Sensing as a New Paradigm

- Case Study II: Surprise-driven continual video segment for VSC

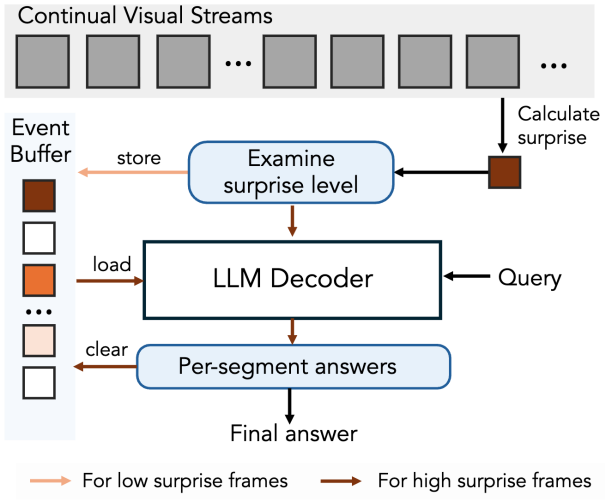
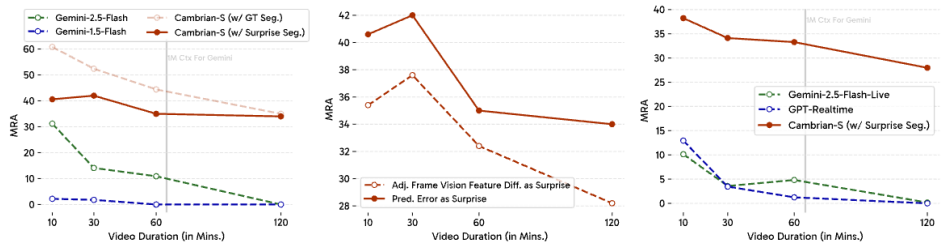


Illustration of our surprise-driven event segmentation framework for VSC.

- **Solution:**
  - **Buffer:** Accumulate frames in a temporary **Event Buffer** .
  - **Detect Boundary:** A sudden **Spike in Surprise** indicates a scene change (e.g., entering a new room).
  - **Action & Reset:**
    - Trigger the LLM to answer for the current segment (e.g., "Count chairs here").
    - **Clear the buffer (Reset)** to free up working memory .
- **Result:** Effectively mimics the human **"Doorway Effect"** to handle unbounded video lengths .

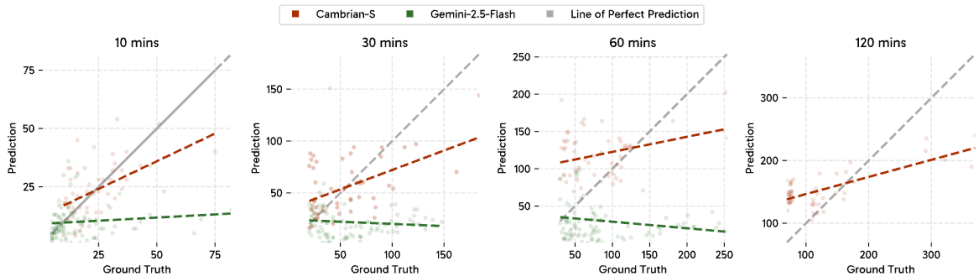
Predictive Sensing as a New Paradigm

- Case Study II: Surprise-driven continual video segment for VSI-SUPER Count



(a) VSC results (b) Different surprise measurement (c) Streaming evaluation

Performance analysis on VSC



Cambrian-S scales to higher ground truth object counts whereas Gemini saturates

- **Streaming Robustness:** Outperforms commercial "Live" models (GPT-4o, Gemini) which fail in long streams .
- **True Generalization:** Unlike Gemini which saturates, Cambrian-S shows **linear correlation** with ground truth counts, indicating true counting behavior rather than pattern matching .

Conclusion

1. Exposing MLLM Limitations:
   - New **VSI-SUPER** Benchmark
   - **Cambrian-S**
2. Proposing **Predictive Sensing**
   - Introduced a new paradigm: using **"surprise"** (prediction error).
   - Proof-of-Concept: Demonstrated significant performance gains on VSI-SUPER

Limitations

- Proof-of-Concept Prototype
- Current Limitations: These designs are limited in:
  - Quality: Room for further refinement and robustness.
  - Scale: Insufficient for achieving full spatial supersensing.
  - Generalizability: Limited in application to diverse real-world scenarios.
- Future Work Directions:
  - Future research should focus on:
    - Diverse & Embodied Scenarios: Exploring more varied and interactive environments
    - Stronger Integration: Connecting with advances in vision, language, and world modeling