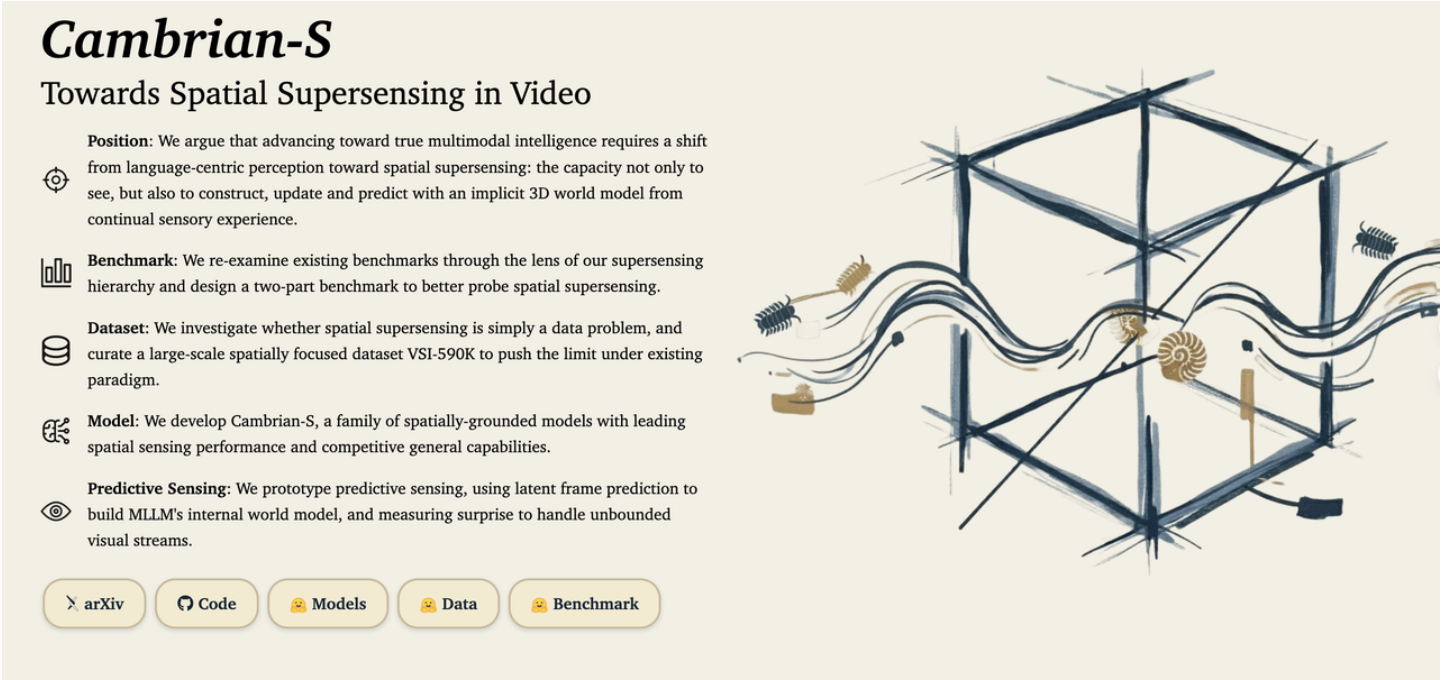
Cambrian-S

Cambrian-*S*: Towards Spatial Supersensing in Video

Shusheng Yang[1*] Jihan Yang[1*] Pinzhi Huang[1†] Ellis Brown[1†] Zihao Yang[1]
Yue Yu[1] Shengbang Tong[1] Zihan Zheng[1] Yifan Xu[1] Muhan Wang[1] Daohan Lu[1]
Rob Fergus[1] Yann LeCun[1] Li Fei-Fei[2] Saining Xie[1]

[1] New York University [2] Stanford University



研究背景：

尽管MLLMs在图像理解方面取得了显著进展，但它们在视频处理上的扩展仍然受限 。现有的视频MLLM大多将视频视为稀疏的帧序列，忽视了视频作为“一个隐藏的、不断演变的3D世界在像素上的连续高带宽投影”这一独特性 。因此，这些模型严重依赖文本召回能力，而对空间结构和动态的表征不足 。

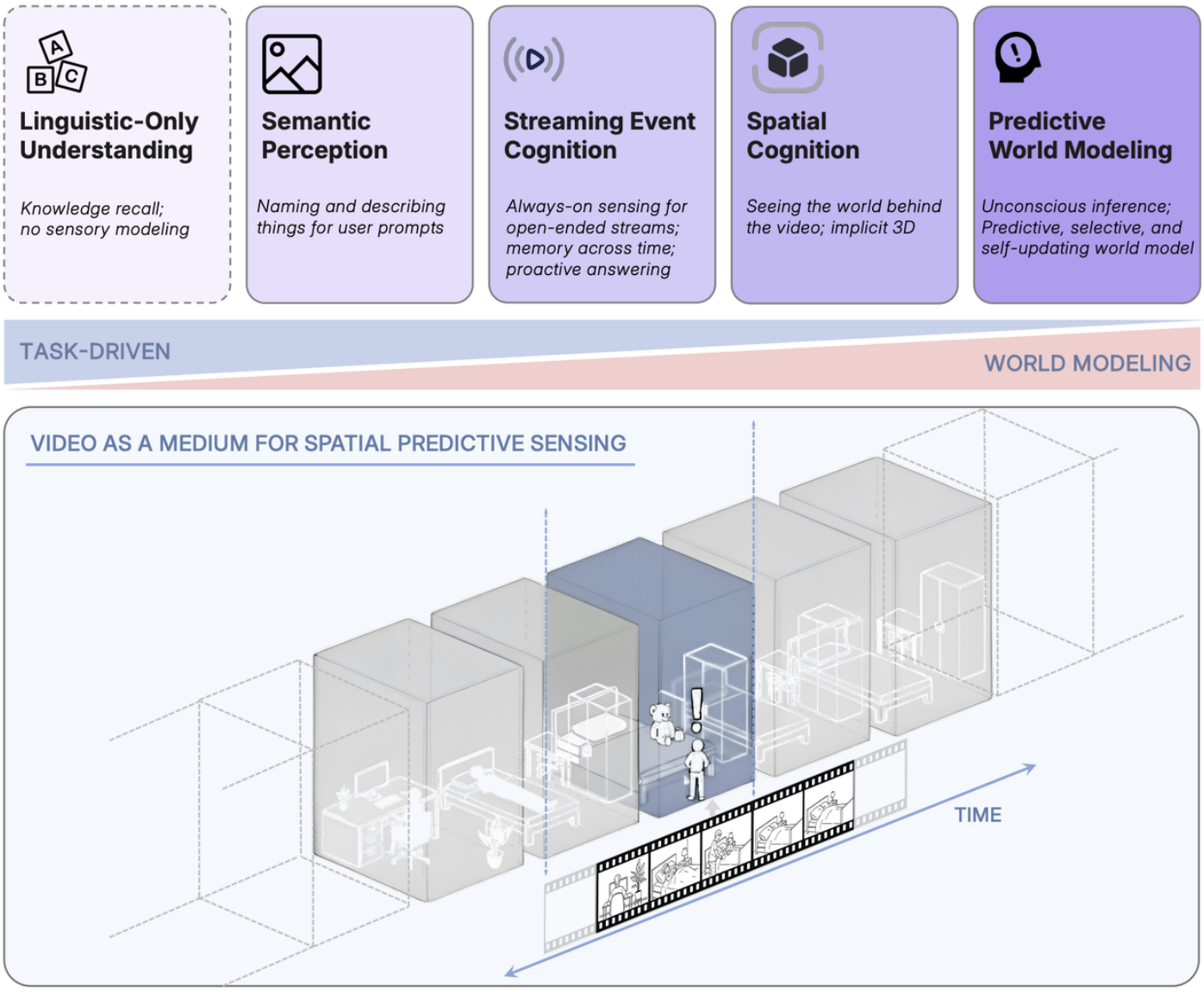
Long Video Und:

1. visual token compression 派：试图压缩减少visual token的数量来支持更长序列：(AuroraCap)，FastV, Llava-prumerge
2. LLM 改架构派，linear model / sparse attn model: (AuroraLong, VideoNSA), 还有一堆Mamba的大差不差, sparse除了videonsa没见人做，但是video gen那很多如Sparse VideoGen （很棒的方向，但是需要会底层infra）

3. memory派：试图构建某些memory机制来做streaming video understanding：（MovieChat），Cambrain-S

研究动机：

要实现真正的多模态智能，需要从当前“任务驱动”的系统转向一种更广泛的“超感”（supersensing）范式。作者的核心动机是推动模型发展出“空间超感”（spatial supersensing），即不仅能“看”，还能从连续的感官体验中“构建、更新和预测”一个隐式的3D世界模型。



0. **(Linguistic-only understanding):** no sensory capabilities; reasoning confined to text and symbols. Current MLLMs have progressed beyond this stage, yet still retain traces of its bias.
1. **Semantic perception:** parsing pixels into objects, attributes, and relations. This corresponds to the strong multimodal *"show and tell"* capabilities present in MLLMs.
2. **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.
3. **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.
4. **Predictive world modeling:** the brain makes *unconscious inferences* [130] by predicting latent world states based on prior expectations. When these predictions are violated, surprise guides attention, memory, and learning [41, 120, 60]. However, current multimodal systems lack an internal model that anticipates future states and uses surprise to organize perception for memory and decision making.

核心问题：

1. 如何设计新的基准来评估更高级的“空间超感”能力（特别是流式认知和预测性建模）？
2. 当前“数据驱动、暴力扩展上下文”的MLLM范式是否足以实现“空间超感”？
3. 如果当前范式不足，什么样的新范式（如“预测感知”）可能引领前进的道路？

当前bench能反映问题吗？

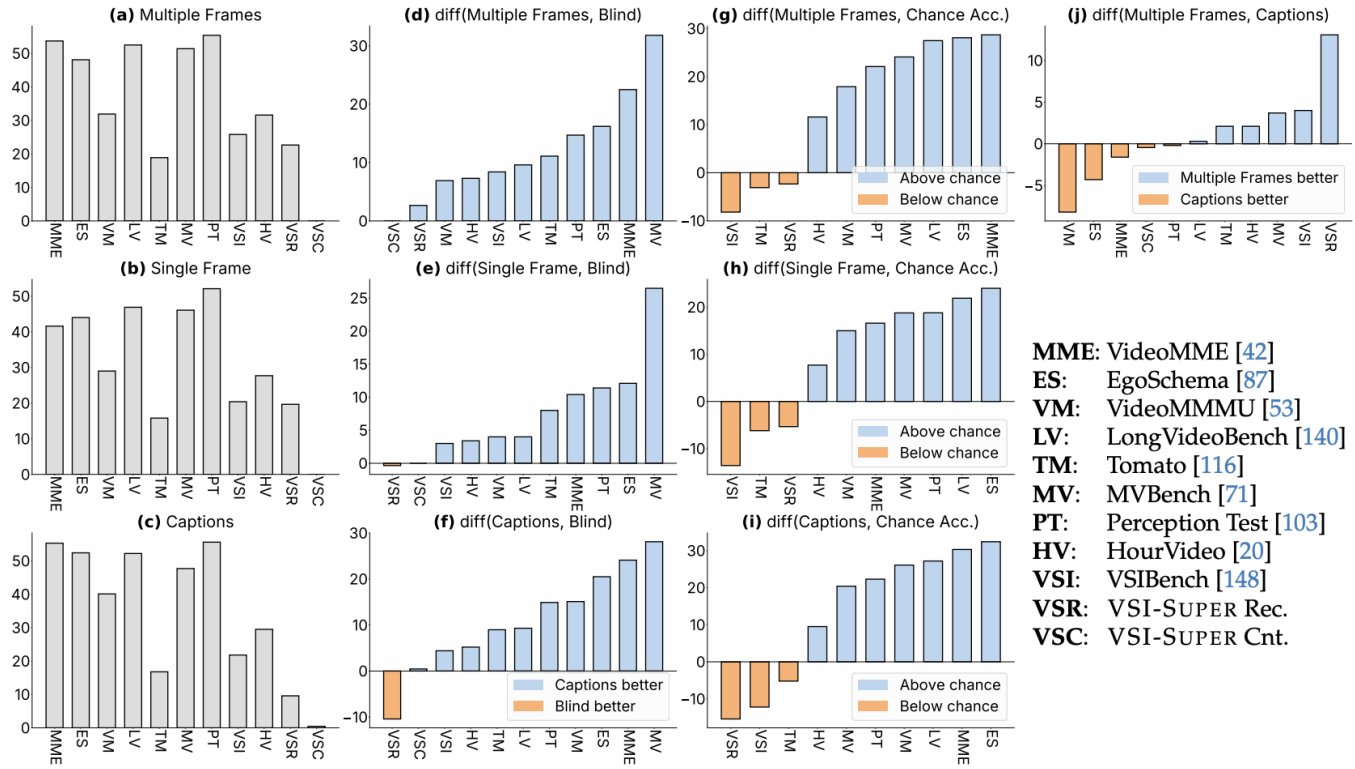


Figure 2 | **Benchmark diagnostic results reveal varying dependence on visual input.** We evaluate model under distinct input conditions: (**a**) multiple (32) uniformly sampled frames, (**b**) a single (middle) frame, and (**c**) frame captions, benchmarked against chance-level and blind test results (visual input ignored). Panels (**a–c**) show absolute accuracies; panels (**d–j**) show performance differences between conditions. Visual inputs are substantially more critical for VSI-Bench [148], Tomato [116], and HourVideo [20], while their impact is less pronounced for VideoMME [42], MVBench [71], and VideoMMMU [53]. VSR and VSC are new supersensing benchmarks introduced in Sec. 2.2.

现有基准（如VideoMME, MVBench等）大多测试语义感知，甚至可以通过“帧字幕”（Frame Captions）而非真实视觉输入来解答，显示出它们更偏向于测试语言先验知识

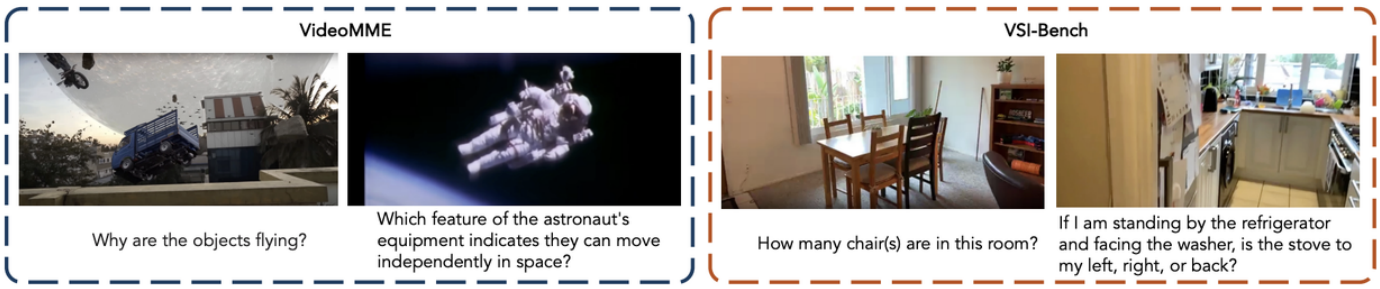


Figure 3 | **Illustrations of how spatial sensing is conceptualized in current video benchmarks**. The left panel features examples from the “spatial reasoning” subcategory of VideoMME [42], including a question regarding gravity from Shutter Authority’s “`What if the Moon Crashed into the Earth?`” and 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.

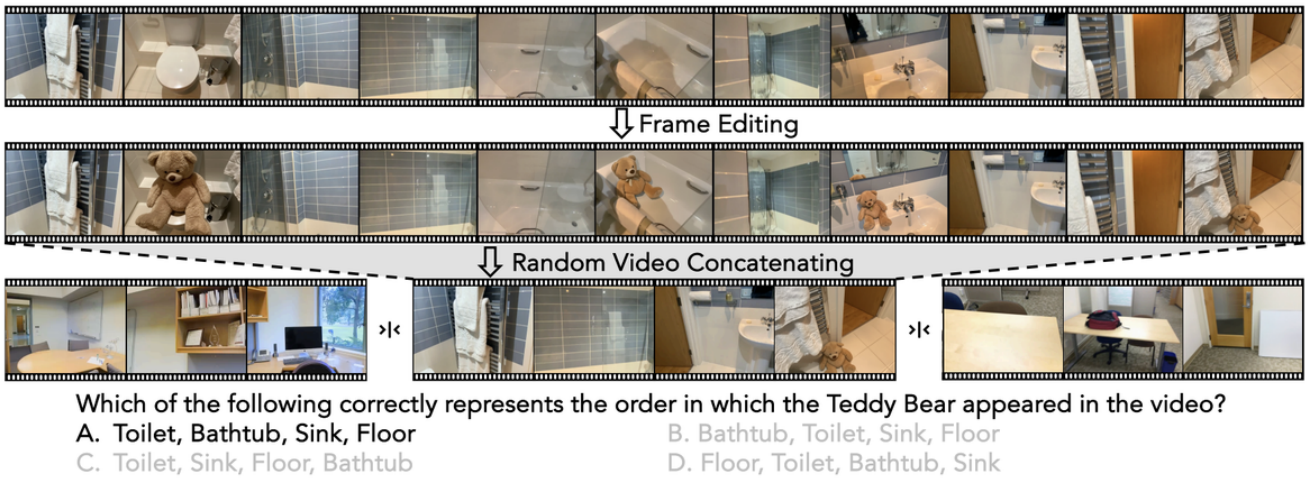


Figure 4 | **Illustration of the VSR benchmark’s construction process and format**. We use generative models to edit videos by inserting *surprising* or out-of-place objects into the space. The core task then challenges models to recall the spatial placements of these objects in the correct order of their appearance across arbitrarily long videos.

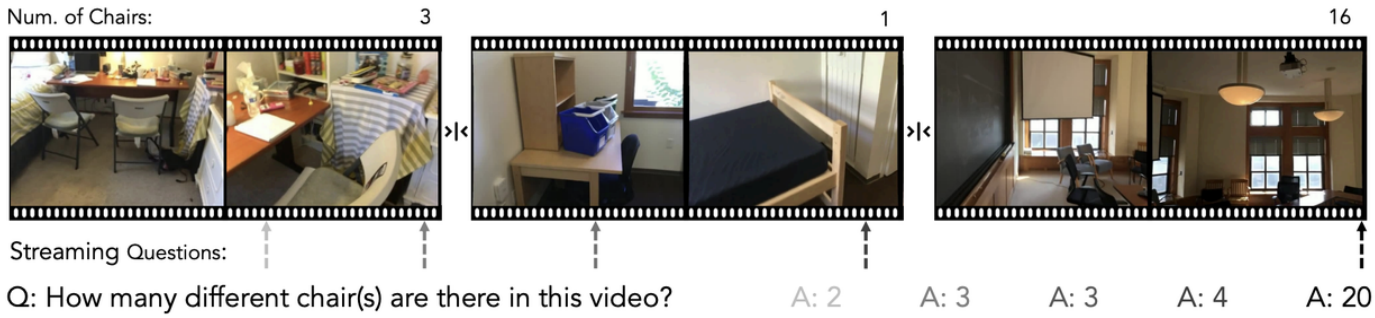


Figure 5 | **Overview of the VSC benchmark.** The benchmark evaluates counting capabilities on long-horizon, multi-room videos composed of concatenated scenes. Queries are posed at various time points to simulate a streaming question-answering setting.

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.

暴力扩展当前范式可以解决问题吗?

3.2. Spatial Video Data Curation: VSI-590K

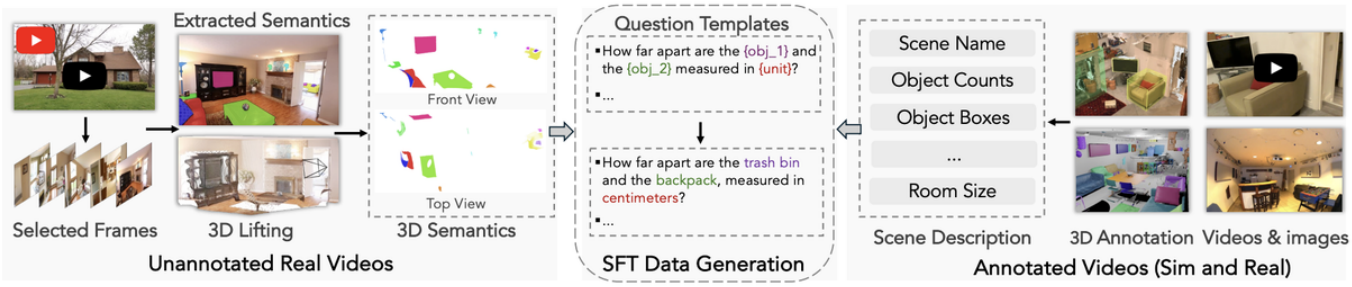


Figure 7 | **VSI-590K data curation pipeline.** We collect data from 3D-annotated real and simulated video sources, as well as from pseudo-annotated frames extracted from web videos. We then use diverse templates to automatically generate question–answer pairs for instruction tuning.

Table 2 | **Data statistics for VSI-590K.** We collect data from 10 sources with different video types and annotations to improve diversity.

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

Table 5 | **Comparison of Cambrian-*S* with other leading MLLMs.** Cambrian-*S* outperforms both proprietary and open-source models across a range of image and video visual–spatial benchmarks and model sizes. For video evaluation, we uniformly sample 128 frames as input. Detailed evaluation settings are provided in Sec. E.

		Video										Image		
Model	Base LM	VSI-Bench	VSI-Bench$^{\text{Debiased}}$	Tomato	HourVideo	Video$^{\text{MME}}$	EgoSchema	Video$^{\text{MMMU}}$	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**

在现有benchmark上成功干到sota

However:

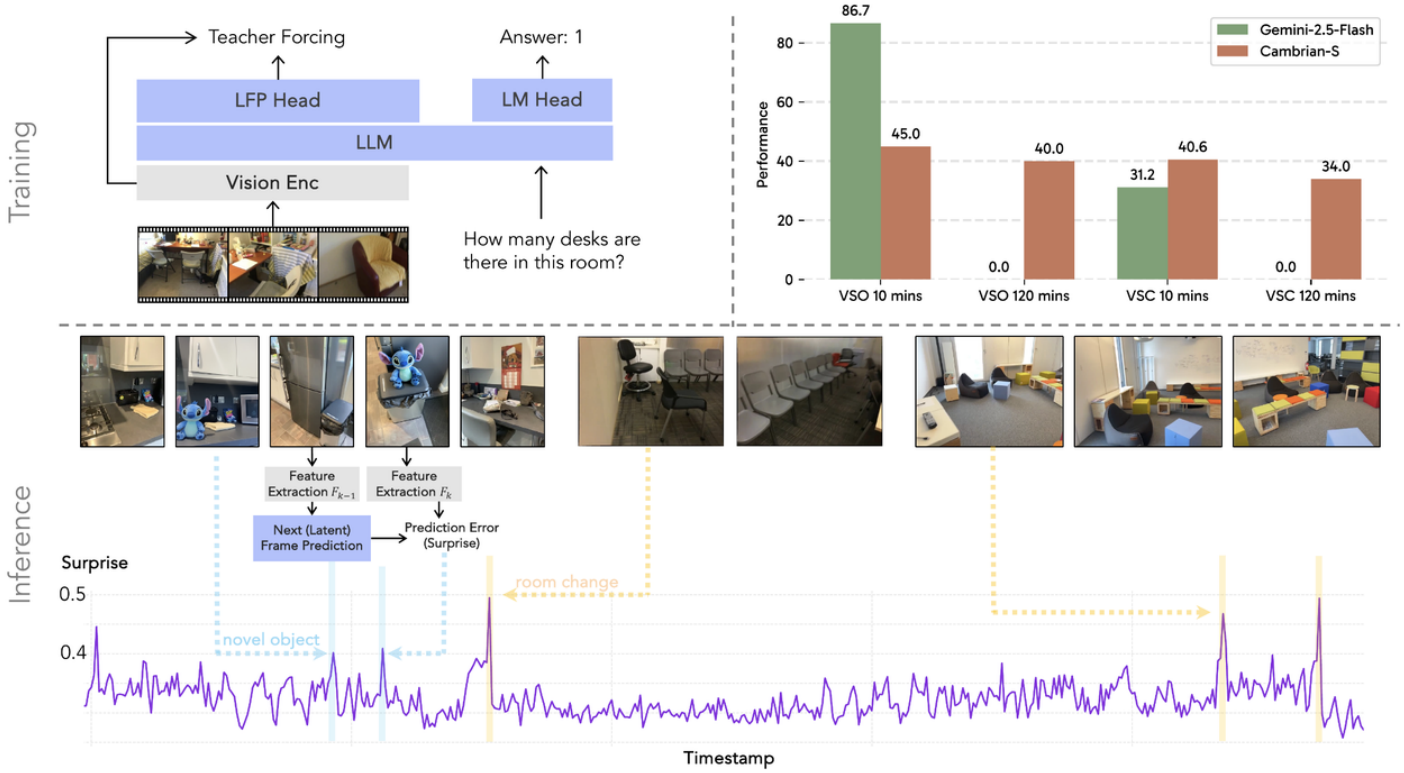
Eval Setup	VSR					VSC			
	10 min	30 min	60 min	120 min	240 min	10 mins	30 min	60 min	120 min
Uni. Sampling, 128F	26.7	21.7	23.3	30.0	28.2	16.0	0.0	0.0	0.0
FPS Sampling, 1FPS	38.3	35.0	6.0	0.0	0.0	0.6	0.0	0.0	0.0

Table: 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.

Scaling data and models is essential, but alone it cannot unlock true spatial supersensing.

需要什么样的新范式?

人类认知理论：与当前能够分词和处理整条数据流的视频多模态模型不同，人类的感知（和记忆）具有高度选择性，只保留了一小部分感官输入。 大脑不断更新内部模型以预测输入刺激，压缩或舍弃那些没有新意信息的可预测输入。 相比之下，意外的感官信息违背预测会产生“惊讶”，并推动注意力和记忆编码的增加。



1. 训练：LFP 模块 (Latent Frame Prediction) 使模型具有“预知未来”的能力

- **架构：** 在阶段 4 训练中，加入一个轻量级的“潜在帧预测头”（LFP Head），它是一个与 LM Head 并行的 MLP。
- **目标：** LFP Head 被训练用于预测下一帧的潜在特征（Latent Representation）。训练目标函数 L_{total} 是标准指令微调损失 L_{IT} 和 LFP 损失 L_{LFP} 的加权和：

  $$L_{total} = L_{IT} + \lambda \cdot L_{LFP}$$

  其中 L_{LFP} 由均方误差（MSE）和余弦距离（Cosine Distance）组成，用于衡量 LFP 头的预测 z_{pred} 与下一帧的真实特征 z_{gt} 之间的差异。

2. 推理：利用“Surprise” 惊喜： 未来与我的预知有很大不同

在推理时，模型计算 LFP 的预测 z_{pred} 与真实下一帧 z_{gt} 之间的余弦距离，将其作为“Surprise”的量化指标：

$$Surprise = 1 - \frac{z_{pred} \cdot z_{gt}}{||z_{pred}|| \cdot ||z_{gt}||}$$

这个 $Surprise$ 信号被用于两个下游任务：

应用 1 (VSR 任务)：意外驱动的内存管理系统

- **压缩 (Compression):** $Surprise$ 值低的帧（即符合预期的、冗余的帧），其 KV Caches 在存入长期记忆 M_l 之前会被 2 倍下采样压缩。
- **整合 (Consolidation):** 当长期记忆 M_l 超过预设预算 B_{long} 时，系统会“忘记” $Surprise$ 得分最低（即最不重要）的帧，以维持恒定的显存占用。

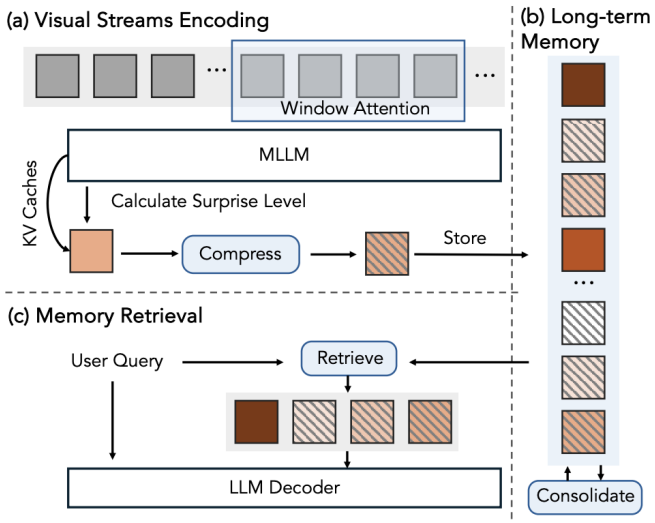


Figure 10 | **Surprise-driven memory management framework design.** The proposed memory system (a) encodes incoming visual streams, compressing frames with low surprise; (b) performs consolidation when memory is full by dropping or merging the least surprising frames; and (c) retrieves relevant frames during query answering. Color shading (dark→light) reflects the degree of surprise, with hatched boxes denoting compressed frames and solid boxes representing uncompressed ones.

应用 2 (VSC 任务)：意外驱动的事件分割

- **分割 (Segmentation):** $Surprise$ 值高的帧（ $s_t \geq \tau$ ）被视为“事件边界”（Event Boundary），例如从一个房间进入另一个房间。
- **处理 (Processing):** 当检测到边界时，模型将缓存区 M_l 中积累的帧视为一个完整的“事件”（如一个房间），对其进行一次计数（ $\hat{a}$ ）。
- **聚合 (Aggregation):** 处理完所有事件后，将所有分段答案（ $\hat{a}$ ）聚合（Sum）起来，得到最终的总数。

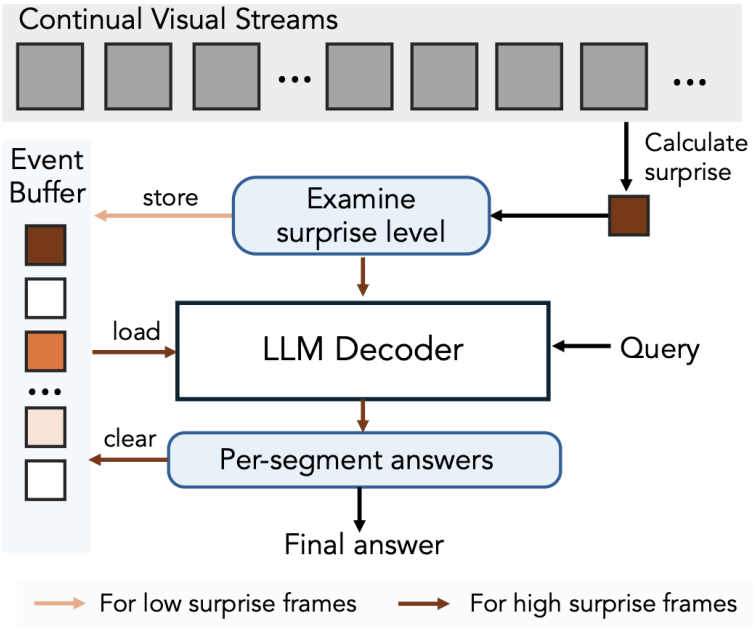


Figure 12 | **Illustration of our surprise-driven event segmentation framework for VSC.** The model continuously accumulates frame features in an event buffer. When a high-surprise frame is detected, the buffered features are summarized to produce a segment-level answer, and the buffer is cleared to start a new segment. This process repeats until the end of the video, after which all segment answers are aggregated to form the final output. Color shading (dark→light) reflects the degree of surprise.

However：Cambrain-S是篇完美的paper吗?

1. 论文得出的“当前堆数据范式无法解决空间超感问题”这一结论，在逻辑推导上不够严密 。

作者用于验证 Scaling 有效性的训练数据（VSI-590K），本质上仍主要由短时序、静态 3D 扫描或切片化的数据组成 。然而，测试基准 VSI-SUPER 却是一个要求处理“超长时序”和“流式连续性”的任务模型在 VSI-SUPER 上的失败，很可能仅仅是因为**训练与测试数据的分布严重不匹配（OOD）**，而非当前范式本身的绝对缺陷。要严格证明“Scaling 无效”，必须进行控制变量实验：尝试将符合 VSI-SUPER 模式的长时序流式数据加入训练。只有当模型在“见过”类似数据后依然无法学会感知与记忆，我们才能宣判当前范式的死刑。

2. “预测感知（LFP）”的成功可能源于对 Benchmark 人工痕迹的“捷径利用”

论文提出的新范式 LFP（Latent Frame Prediction）在 VSI-SUPER 上表现优异，但这种成功可能建立在基准测试的人工构造缺陷之上，而非真正的世界模型能力。

VSI-SUPER 的构建依赖于强烈的“反先验”操作：VSR（召回任务）是在正常帧中**人工插入**突兀的物体，VSC（计数任务）是将不同视频**人工拼接**在一起 。

- **批判视角：** LFP 的核心机制是检测“预测误差（Surprise）” 。在 VSI-SUPER 中，人工插入的物体和视频拼接的硬切口（Hard Cut），天然会产生巨大的、非自然的预测误差。

这并不是一篇benchmark-method paper，更像是一篇position paper，核心论调在于：

现在的MLLM范式是不行的，是暂时的妥协，但是是无法长期依赖的。

未来可能是什么：

1. Video Encoder
2. Native MLLM
3. RNN/Mamba-like memory system / TTT