TOWARDS SPATIAL SUPERSENSING IN VIDEO

Anonymous authors
Paper under double-blind review

ABSTRACT

We frame spatial *supersensing* in video as an overarching goal for multimodal intelligence and argue that progress requires a shift from long-context brute force to *predictive sensing*. Using a four-level taxonomy: *semantic perception*, *streaming event cognition*, *implicit 3D spatial cognition*, and *predictive world modeling*, we audit existing benchmarks and show they focus heavily on the first tier, with only partial coverage of streaming and spatial cognition, and almost never test true world modeling. To ground these gaps, we introduce VSI-SUPER, a two-part benchmark for continual spatial sensing: VSO (long-horizon spatial observation and recall) and VSC (continual counting under changing viewpoints and scenes). These tasks admit arbitrarily long video inputs and are specifically built so that simply scaling tokens or context length isn't enough. Within the current paradigm, we push spatial cognition by curating *VSI-590K* and training a new family of video MLLMs that deliver $+30\%$ absolute on VSI-BENCH without sacrificing general semantic perception. Yet these models still underperform on VSI-SUPER, exposing a paradigm gap. We then prototype *predictive sensing*: a self-supervised next latent-frame predictor whose *surprise* (prediction error) drives long-horizon memory and event segmentation. On VSI-SUPER, this approach substantially outperforms leading video MLLMs, evidencing that advancing spatial supersensing requires models that not only see but also anticipate, select, and organize experience.

1 INTRODUCTION

Video is a continuous sensory signal that projects a hidden, evolving 3D world onto pixels (Gibson, 2014; Marr, 2010). While multimodal LLMs (MLLMs) have advanced rapidly by pairing strong visual encoders with language models (Achiam et al., 2023; Team et al., 2024; Liu et al., 2023; Tong et al., 2024), most video extensions (Wang et al., 2024d; Li et al., 2024a; Bai et al., 2025a) still treat streams as sparse frames, underrepresent embodied spatial information (Yang et al., 2024e), and lean on knowledge recall. This undercuts the very capabilities that make video distinct, and leaves the central challenge of *world-level* reasoning underexplored. We propose *spatial supersensing* as the north star of multimodal intelligence, structuring requirements into four stages of capability (Fig. 1):

- **Semantic perception:** parsing pixels into objects, attributes, and relations. This corresponds to the strong multimodal understanding capabilities present in MLLMs.
- **Streaming event cognition:** operating on unbounded live streams with proactive support, aligning with efforts to make MLLMs real-time "watch-along" assistants.
- **Implicit 3D spatial cognition:** treating frames as 2D 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.
- **Predictive world modeling:** an internal model anticipates future states and uses expectation and surprise to organize perception for memory and decision-making, mirroring human "unconscious inference" (Von Helmholtz, 1867). Such predictive sensing is largely absent in current systems.

Our paper unfolds in three parts. **First**, we critically examine existing benchmarks through the lens of our supersensing hierarchy. We find that most benchmarks map to the first two levels, while a few such as VSI-Bench (Yang et al., 2024e) begin to probe Spatial Cognition. However, none sufficiently addresses the final, crucial level of Predictive World Modeling. To make this gap concrete and motivate a shift in approach, we introduce VSI-SUPER, a two-part benchmark for continual spatial sensing: VSO targets long-horizon spatial observation and recall, while VSC tests continual counting across changing viewpoints and scenes. Built from arbitrarily long spatial videos, these tasks are deliberately resistant to the current multimodal recipe; they require perception to be selectively filtered and structured rather than naively accumulated. We show that even state-of-the-art commercial long-context models struggle on them.

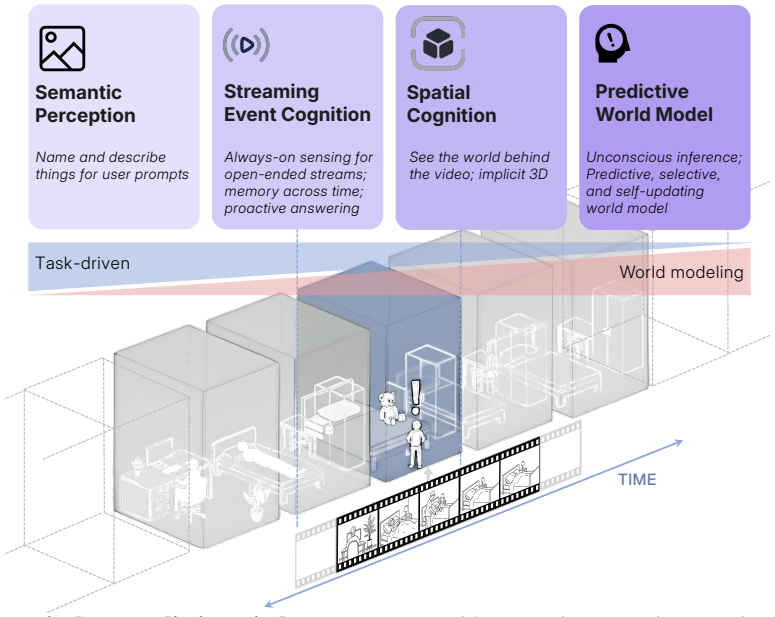


Figure 1: **From pixels to predictive minds:** systems start with semantic perception, naming and describing what they see. Streaming event cognition goes further, with always-on sensing across continuous streams, memory, and proactive answering. Spatial cognition captures the implicit 3D structure of video, enabling reasoning about objects, configurations, and metrics. Ultimately, a predictive world model emerges. One that learns passively from experience, updates through prediction and surprise, and retains information for future use. **Lower illustration:** video is the ideal testbed. Models must advance from frame-level Q&A to constructing implicit world models that enable deeper spatial reasoning, scale to unbounded horizons, and achieve supersensing rivaling (and ultimately surpassing) human visual intelligence.

Second, we ask whether spatial supersensing is simply a data problem. We curate *VSI-590K*, a spatially focused instruction-tuning corpus over images and videos, and introduce *Cambrian*-S, a family of video MLLMs. Under the current paradigm, careful data design and training push Cambrian-*S* to state-of-the-art spatial cognition on VSI-BENCH (>30% absolute gain) without sacrificing general semantic perception. Nevertheless, Cambrian-*S* still falls short on VSI-SUPER, indicating that while scale lays crucial groundwork, it alone is not sufficient for spatial supersensing.

This motivates the **third** and final part of our paper, where we propose *predictive sensing* as a first step toward a new paradigm. We present a proof-of-concept solution built on a self-supervised next-latent-frame prediction task. Here, we leverage the model's prediction error, or "surprise", for two key functions: 1) as a mechanism to manage memory, allocating more resources to unexpected events, and 2) as a signal for event segmentation, breaking an unbounded continuous stream into meaningful chunks. We demonstrate that this approach, though simple, significantly outperforms a strong long-context baseline on our two new tasks. While not a final solution, this result provides compelling evidence that the path to true supersensing requires models that don't just see, but actively predict and learn from the world.

To summarize, our contributions are: (**1**) We define a hierarchy for spatial supersensing and introduce two novel benchmarks that reveal the limitations of the current paradigm. (**2**) We develop Cambrian-*S*, a state-of-the-art model that pushes the limits of spatial cognition. This effort provides a powerful new baseline and, by revealing the precise boundaries of current methods on our new benchmarks, illuminates the path forward to a new paradigm. (**3**) We propose predictive sensing as a promising new direction, showing that leveraging model surprise is a more effective strategy for long-horizon spatial reasoning than passive context expansion.

2 BENCHMARKING SPATIAL SUPERSENSING

To ground our pursuit of spatial supersensing, we must first establish how to measure it. This section undertakes a two-part investigation into benchmarking this capability. We begin by auditing a suite of popular video MLLM benchmarks, where our analysis (Fig. 16) reveals that they overwhelmingly focus on semantic perception while neglecting the more advanced spatial and temporal reasoning required for supersensing (§2.1). To address this critical gap, we then introduce VSI-SUPER, a new

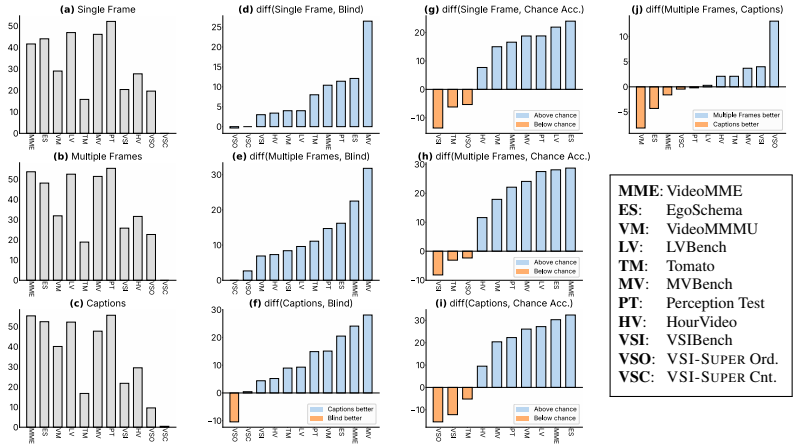


Figure 2: We evaluate performance under distinct input conditions: (**a**) a single (middle) frame, (**b**) multiple (32) uniformly sampled frames, and (**c**) frame captions. We compare these against chance-level and blind test results (visuals ignored). We first present the absolute accuracies achieved on each benchmark for input conditions (**a**–**c**). Next, we detail a series of performance differences (**d**–**j**) that arise from comparing these varied inputs and baselines (*e.g.*, single-frame *vs.*blind, frame captions *vs.*multi-frame). This comparative analysis indicates that visual inputs are substantially more critical for performance on benchmarks such as VSI (Yang et al., 2024e), Tomato (Shangguan et al., 2024), and HourVideo (Chandrasegaran et al., 2024), while their impact is less pronounced for benchmarks like VideoMME (Fu et al., 2024), MVBench (Li et al., 2024d), and VideoMMMU (Hu et al., 2025). VSO and VSC are new supersensing benchmarks we will introduce in Sec. 2.2.

benchmark specifically designed to probe these harder, continual aspects of spatial intelligence (§2.2). We use this benchmark to test the limits of the current paradigm throughout the rest of the paper.

2.1 Deconstructing Existing Video Benchmarks

To assess if existing benchmarks evaluate *true visual sensing* or simply rely on language priors, we conduct a series of diagnostic tests. We use our base Cambrian-1 model to probe a suite of representative video benchmarks under varied input conditions, allowing us to disentangle the underlying task demands from the capabilities of more complex video-specific architectures.

Diagnostic Setup. We establish five experimental conditions to isolate the contributions of different information sources. We provide the model with either a Single Frame (the middle frame), Multiple Frames (32 uniformly sampled frames), or textual Frame Captions generated from those 32 frames. We compare these against two baselines: a Blind Test, where the model only receives the question, and Chance Acc, which represents random guessing. By analyzing performance differences between these conditions—such as `diff(Multiple, Single)` to assess temporal cues or `diff(Multiple, Captions)` to control for textual solvability—we can create a fine-grained profile of each benchmark's characteristics.

Analysis of Results. Our findings, presented in Fig. 2, reveal a clear divide among popular benchmarks. Many can be surprisingly well-addressed with minimal or even non-visual input. For example, using only textual captions surpasses chance accuracy on all but 3 benchmarks—and by over 20% on benchmarks like EgoSchema (Mangalam et al., 2023), VideoMME (Fu et al., 2024), and VideoMMMU (Hu et al., 2025). This suggests these tasks can often be solved with high-level textual summaries, probing *language inference* more than *direct visual perception*. The performance gap between using multiple frames versus just captions is also telling (Fig. 2-j); a small margin on benchmarks like VideoMMMU and EgoSchema indicates a more language-centric nature.

Conversely, a few benchmarks demonstrate a strong reliance on visual sensing. Our image-based model struggles on **VSI-Bench** and **Tomato**, often performing below chance level with single-frame inputs. These benchmarks show the largest performance gains when provided with rich, multi-frame visual information, confirming that they effectively test the nuanced, spatiotemporal reasoning that is the hallmark of true video understanding.

Remark. We emphasize the *inherent challenges* in benchmarking and the impracticality of creating a single, all-encompassing benchmark. We do not intend that a reliance on language priors is an inherent flaw; world knowledge is crucial for many tasks. Rather, our goal is to highlight that "video

understanding" is not monolithic. Benchmarks should be chosen to align with the specific capabilities under investigation. This audit demonstrates a clear need for tasks that specifically drive progress towards the advanced spatial and continual sensing we aim to measure.

2.2 VSI-SUPER: PROBING SPATIAL SUPERSENSING IN MULTIMODAL LLMS

Spatial supersensing requires four key capabilities (see Fig. 1): *semantic perception*, *streaming event cognition*, *implicit 3D spatial cognition*, and *predictive world modeling*. Most existing video benchmarks evaluate basic semantic perception (Fu et al., 2024; Mangalam et al., 2023). Recent work has begun exploring proactive and real-time video QA (Chen et al., 2024d) and long-video modeling (Song et al., 2024; Li et al., 2024e; Zhang et al., 2024a) for streaming event cognition, while VSI-Bench (Yang et al., 2024e) assesses spatial cognition. However, no existing testbed probes high-level capability of predictive world modeling or examines spatial supersensing holistically. To ground the gaps between current MLLMs and spatial supersensing, we design VSI-SUPER, a two-part benchmark for continual spatial sensing that requires MLLMs to *selectively filter and accumulate visual signals on unbounded spatial videos* to answer questions. Details in Sec. C.

VSO: Long-horizon Spatial Observation and Recall. The VSO benchmark requires MLLMs to observe long spatiotemporal videos, and recall the specific locations of an unusual object in the correct order of its appearance. To construct this benchmark, human annotators use an image-editing model (Comanici et al., 2025) to insert surprising or out-of-place objects (*e.g.*, a Teddy Bear, Hello Kitty) into four distinct frames of a space-scanning video (Dai et al., 2017; Yeshwanth et al., 2023; Baruch et al., 2021) (see Fig. 3). This edited video is then concatenated with other similar space scan videos to create an arbitrarily long and continuous visual stream.

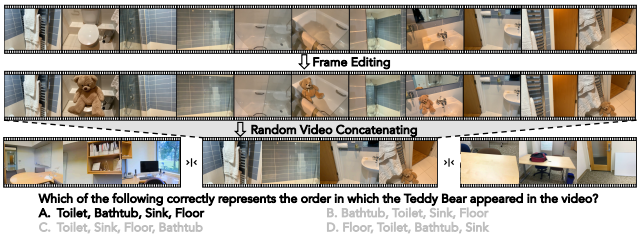


Figure 3: **VSO Task**. Recall the placement of objects in their correct appearance order from an arbitrarily long video.

VSC: Continual Counting under Changing Viewpoints and Scenes. While VSO mainly examines MLLMs to *recall* part of the observation from unbounded visual streams, VSC requires MLLMs to perform continuous, unique object counting in long-form spatial videos. The benchmark is constructed by concatenating multiple space-scanning clips from VSI-Bench (Yang et al., 2024e), and the task is to determine the total count of a specific object across the entire concatenated video (see Fig. 4). To evaluate the numerical answer question format, we adopt the mean relative accuracy ($\mathcal{MRA}$) metric following VSI-Bench (Yang et al., 2024e).

Figure 4: **VSC Task.** We concatenate several videos sampled from VSI-Bench and benchmark the model's counting ability.

Frontier Models Can't Crack VSI-SUPER. To see if VSI-SUPER can be readily solved by cutting-edge MLLMs, we put Gemini-2.5-Flash to the test. As shown in Tab. 1, despite its 1-million-token context length, the model still suffers from context overflow when processing 2-hour videos. Simply scaling up tokens or context length will never be enough, as VSI-SUPER can easily exceed any fixed context window by simply creating an arbitrarily long video. Even for 60-minute videos in VSI-SUPER that fall within its context window, performance remains limited, achieving only 34.7 on VSO and 10.9 on VSC. In contrast, Gemini-2.5-Flash excels at *semantic perception* and *knowledge retrieval* video benchmarks like VideoMME and VideoMMMU with around 80% accuracy.

Table 1: Gemini-2.5-Flash results on video benchmarks.

Model	VideoMME	VideoMMMU	VSI-Bench	VSO		VSC	
				60 Mins.	120 Mins.	60 Mins.	120 Mins.
Gemini-2.5-Flash	81.5	79.2	45.7	34.7	Out of Ctx.	10.9	Out of Ctx.

Challenging the Current Paradigm. The difficulty of VSI-SUPER extends beyond mere spatial reasoning, exposing fundamental limitations of the current MLLM paradigm. *First*, these tasks challenge the assumption that progress can be achieved by simply scaling resources. By admitting arbitrarily long video inputs, VSI-SUPER is designed to exceed any fixed context window, making brute-force approaches that process every frame computationally infeasible. Humans solve this by selectively attending to and retaining only a fraction of sensory input (**??**), a capability absent in current models. *Second*, the tasks demand advanced cognitive capabilities beyond simple perception.

For example, VSC requires not only the generalization of counting behavior to out-of-distribution scales but also the ability to segment a continuous stream into meaningful events—knowing when to start, continue, or reset a count across changing scenes. *This suite of challenges—spanning resource constraints, generalization, and cognitive functions like aggregation and segmentation—necessitates a paradigm shift* from purely data-driven approaches towards models that can form and leverage an internal world model to intelligently organize and reason about an unbounded visual world.

3 Pushing the Limits of Spatial Sensing in Current MLLMs

Is supersensing simply a data problem? We investigate this question by pushing the current data-centric MLLM paradigm to its limits. We begin by developing a strong base MLLM (§3.1); then after curating a large-scale, spatial instruction-tuning dataset, VSI-590K (§3.2), we produce a spatially-grounded Cambrian-*S* model family (§3.3). Our subsequent evaluation reveals a crucial split: this approach yields state-of-the-art results on existing spatial tasks but fails on the continual sensing challenges of VSI-Super, demonstrating the limitations of a purely data-driven approach (§3.4).

3.1 A Strong Foundation: Upgrading Cambrian-1

We begin by developing a powerful general MLLM as the starting point for our experiments, by upgrading Cambrian-1 with two modern components: SigLIP2-SO400m visual encoder (Tschannen et al., 2025) and the Qwen2.5-7B (Yang et al., 2024a) instruction-tuned LLM. Full implementation details are available in Sec. E.

3.2 VSI-590K: Is Spatial Sensing Simply a Data Problem?

To understand the world, robust spatial sensing capabilities are essential. However, recent analysis in Thinking in Space (Yang et al., 2024e) reveals that even frontier MLLMs face significant challenges in visual spatial intelligence tasks. We believe this is due to a lack of high-quality spatially-grounded data in current instruction-tuning datasets (Zhang et al., 2024c; Cui et al., 2024; Ray et al., 2025). These observations motivate our curation of VSI-590K: a large-scale instruction-tuning dataset designed to impart visuospatial understanding.

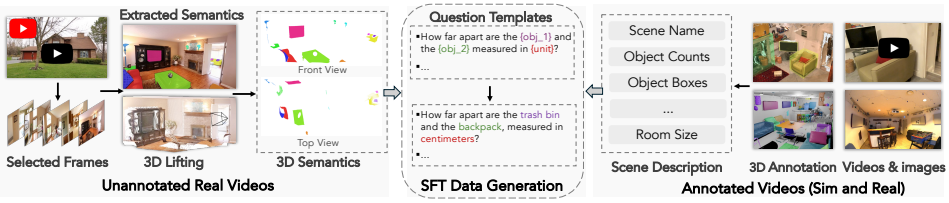


Figure 5: **VSI-590K Data Collection and Curation Pipeline.** Data is derived from 3D-annotated real and simulated video sources and from pseudo-annotated images of unannotated real videos. Question-Answer pairs are then automatically generated via question templates augmented for variety.

To construct a dataset that is both large-scale and high-quality, we combine data from three diverse source types, as illustrated in Fig. 5. First, for high-fidelity geometric grounding, we source annotated real videos from existing indoor scan and ego-vision datasets. Second, to increase scale and diversity beyond the scarcity of 3D-annotated data, we leverage embodied simulators to programmatically generate simulated data with rich spatial annotations. Finally, to capture the visual diversity of the web, we develop a pipeline to produce pseudo-annotated images from unannotated real videos sourced from YouTube and robotics datasets. The full details of our data curation and processing pipeline for each source are available in Sec. D.2.

The instruction-tuning data is generated via a comprehensive taxonomy of 12 spatiotemporal question types, augmented with varied phrasing and perspectives to ensure diversity (see Sec. D.1 for details). A detailed ablation study, presented in Sec. D.3, confirms the effectiveness of our data mixture. The study shows that training on the full VSI-590K dataset is critical for performance and that annotated real videos provide the most significant benefit, highlighting the value of high-quality video data for developing robust spatial understanding.

3.3 Cambrian-*S*: A Spatially-Grounded MLLM

To conduct our experiment, we develop **Cambrian-*S***, a family of spatially-grounded models with varying LLM scales: 0.5B, 1B, 3B, and 7B parameters. These models are the final artifacts of a carefully designed 4-stage training pipeline aimed at progressively building general and then specialized spatial capabilities, as illustrated in Fig. 6. The first two stages follow the Cambrian-1

process to establish strong image understanding. In Stage 3, we lift the models to video by performing general video instruction tuning on CamS-3M, a curated 3M-sample dataset mixture. This stage establishes a robust foundation for video understanding before introducing specialized skills.

The final and critical step is Stage 4, where we teach spatial sensing. In this stage, models are finetuned on a mixed corpus of our specialized VSI-590K and a proportional sample of the general video data from Stage 3. This data mixture is a deliberate choice; a detailed ablation study in Sec. G shows that while training on VSI-590K alone yields the highest scores on VSI-Bench, it degrades performance on general video benchmarks. Our mixed approach preserves broad video understanding while imparting strong spatial intelligence. Detailed finetuning setups such as data recipes and hyperparameters can be found in Sec. E.

3.4 Empirical Results: Success in Cognition, Failure in Continuity

We now evaluate the Cambrian-*S* models to test the efficacy and limitations of our data-centric approach. Our evaluation reveals a critical split: while the models achieve state-of-the-art performance on established spatial reasoning benchmarks, their architectural paradigm fundamentally falls short on the continual sensing tasks introduced in VSI-Super.

Figure 6: **Four-Stage Cambrian-*S* Training Pipeline.**

Success on Spatial Cognition Tasks. As shown in Tab. 3, our method yields a new state-of-the-art in spatial reasoning. Cambrian-*S*-7B achieves a score of 67.5% on VSI-Bench, significantly outperforming all open-source models and even surpassing the proprietary Gemini-2.5 Pro by over 16 absolute points. This strong performance includes remarkable generalization; on the complex "route planning" subtask, which was not present in our VSI-590K training, Cambrian-*S*-7B outperforms Gemini-1.5 Pro (see Tab. 17). Furthermore, our training recipe proves highly effective even at smaller scales, with our *0.5B model* rivaling Gemini-1.5 Pro on VSI-Bench. This focus on spatial skills does not compromise general capabilities, as Cambrian-*S* maintains competitive performance on standard video benchmarks (see Sec. F for full results).

Table 2: CambrianS-7B results on VSO and VSC.

	VSO					VSC			
Duration (in Mins.)	10	30	60	120	240	10	30	60	120
Cambrian-*S*-7B	38.3	35.0	6.0	0.0	0.0	0.6	0.0	0.0	0.0

Failure on Continual Sensing Tasks. Despite this success on tasks involving short, pre-segmented clips, the fixed-context architecture of Cambrian-*S* is ill-suited for the demands of continual sensing. When evaluated on the long-horizon tasks in VSI-Super, the limitations of the current paradigm become clear (see Tab. 2). On VSO, which tests long-term recall, the model's performance degrades significantly as the video length increases, dropping to 35.0% accuracy on videos longer than 30 minutes before eventually running out of memory. On VSC, which requires continual counting over an extended period, the model is unable to process the entire stream and fails to maintain an accurate count, achieving a final score of only 0.6% on 10-minutes videos, and 0.0% on videos longer than 30 minutes. These results demonstrate that a purely data-centric approach within a fixed-context architecture, no matter how well-tuned, hits a fundamental wall. Overcoming these challenges requires a paradigm shift towards models that can intelligently manage memory and process unbounded streams, which we explore in the following section.

4 A New Paradigm: Predictive Sensing for Unbounded Streams

The failure of the *fixed-context* Cambrian-*S* and Gemini-2.5 models on VSI-Super reveals a fundamental paradigm gap: simply scaling data and context is insufficient for the demands of unbounded, continuous streams. To bridge this gap, we propose an approach inspired by predictive coding in the human brain: **predictive sensing** (Friston, 2010; Von Helmholtz, 1867; Stahl & Feigenson, 2015; Kennedy et al., 2024). Instead of indiscriminately processing all sensory input, this paradigm uses an internal world model to continuously predict what comes next. The resulting prediction error, or "surprise," serves as an efficient, self-supervised signal for downstream cognitive tasks like selective memory and event segmentation. In this section, we detail our proof-of-concept implementation of this mechanism (§4.1) and then demonstrate its effectiveness on the very VSI-Super tasks where the previous approach failed (§4.2).

Table 3: **Comparison of Cambrian-S with other leading MLLMs.** Cambrian-*S* leads against proprietary and open-sourced models on various image and video visual-spatial benchmarks.

		Video									Image		
Model	Base LM	VSI-Bench	Tomato	HourVideo	Video^{MME}	EgoSchema	Video^{MMMU}	LVBench	MVBench	Percept. Test	RWQA	3DSR	CV-Bench
Proprietary Models													
Claude-3.5-sonnet	UNK.	-	27.8	-	62.9	-	65.8	-	-	-	51.9	48.2	-
GPT-4o	UNK.	34.0	37.7	37.2	71.9	-	61.2	66.7	-	-	-	44.2	-
Gemini-1.5-Pro	UNK.	48.8	36.1	37.3	75.0	72.2	53.9	64.0	-	-	67.5	-	-
Gemini-2.5 Pro	UNK.	51.5	-	-	-	-	83.6	67.4	-	-	-	-	-
Open-Source Models													
LLaVA-Video-7B	Qwen2-7B	35.6	22.5	28.6	63.3	57.3	36.1	58.2	58.6	67.9	66.4	-	75.7
LLaVA-One-Vision-7B	Qwen2-7B	32.4	25.5	28.3	58.2	60.1	33.9	56.4	56.7	57.1	66.3	-	74.3
Qwen-VL-2.5-7B	Qwen2.5-7B	33.5	-	-	65.1	65.0	47.4	56.0	69.6	-	-	48.4	-
InternVL2.5-8B	InternLM2.5-7B	34.6	-	-	64.2	50.6	-	60.0	72.0	-	**68.4**	50.9	-
InternVL3.5-8B	Qwen3-8B	56.3	-	-	**66.0**	61.2	**49.0**	**62.1**	**72.1**	-	67.5	-	-
Cambrian-S-7B	Qwen2.5-7B	**67.5**	**31.1**	**36.0**	63.3	**76.8**	38.6	59.4	64.5	**69.9**	65.9	**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	-	-	**61.5**	-	-	**54.2**	-	**66.9**	-	-	-
Cambrian-S-3B	Qwen2.5-3B	**57.3**	**25.4**	**36.8**	60.2	**73.5**	**25.2**	52.3	**60.2**	65.9	**60.1**	**50.9**	**75.2**
SmolVLM2-2.2B	SmolLM2-1.7B	27.0	-	-	-	34.1	-	-	48.7	51.1	-	-	-
InternVL2.5-2B	InternLM2.5-1.8B	25.8	-	-	51.9	47.4	-	52.0	**68.8**	-	60.1	-	-
InternVL3.5-2B	Qwen3-1.7B	51.5	-	-	**58.4**	50.8	-	**57.4**	65.9	-	**62.0**	-	-
Cambrian-S-1.5B	Qwen2.5-1.5B	**54.8**	**22.5**	**31.4**	55.6	**68.8**	24.9	50.0	58.1	**63.2**	54.5	**51.9**	**69.6**
SmolVLM2-0.5B	SmolLM2-360M	26.1	-	-	-	20.3	-	-	43.7	44.8	-	-	-
LLaVA-One-Vision-0.5B	Qwen2-0.5B	28.5	-	-	44.0	26.8	-	45.8	45.5	49.2	55.6	-	55.5
InternVL2.5-1B	Qwen2.5-0.5B	22.5	-	-	50.3	39.8	-	47.9	**64.3**	-	**58.1**	-	-
InternVL3.5-1B	Qwen3-0.6B	49.9	-	-	**51.0**	41.5	**33.0**	**53.0**	61.0	-	57.6	-	-
Cambrian-S-0.5B	Qwen2.5-0.5B	**50.6**	**23.4**	**27.9**	44.0	**62.4**	15.7	44.0	51.8	**56.0**	52.0	**48.5**	**59.8**

4.1 The Predictive World Model Primitive

We implement our predictive sensing paradigm through a lightweight, self-supervised module called the Latent Frame Prediction (LFP) head, which is trained jointly with the primary instruction-tuning objective. This is achieved by modifying the Stage 4 training recipe as follows:

- **Latent Frame Prediction Head.** We introduce an LFP Head, a two-layer MLP that operates in parallel with the language head, to predict the latent representation of the subsequent video frame. This architecture is illustrated in the top left of Fig. 7.
- **Learning Objectives.** To optimize the LFP head, we introduce two auxiliary losses, MSE and cosine similarity, which measure the discrepancy between the predicted latent and the ground-truth feature of the next frame. A coefficient balances the combined LFP loss against the primary instruction-tuning objective.
- **Dedicated Prediction Data.** We augment the Stage 4 data with a 290K-video subset from VSI-590K used exclusively for the LFP objective. Critically, while instruction-tuning videos are sampled uniformly to retain rich context for question answering, these LFP videos are sampled at a consistent 1 FPS to provide a fixed temporal interval for the prediction task.

During this modified Stage 4 finetuning, we train the connectors, language model, and both the language and LFP heads end-to-end, while the SigLIP vision encoder remains frozen. All other training dynamics are kept identical to the original stage-4.

Inference: Estimating Surprise via Prediction Error. At inference time, we use this trained LFP head to generate a "surprise" signal via a *Violation-of-Expectation* (VoE) process (Garrido et al., 2025). As the model receives new video frames, it continuously predicts the latent features of the next frame. We then compute the patch-averaged cosine distance between the model's prediction and the actual ground-truth feature of that next frame. This distance serves as a quantitative measure of surprise, with larger values indicating a greater violation of the model's learned expectations. This surprise score acts as a powerful, self-supervised guidance signal for the downstream tasks explored next.

4.2 Case Studies: How Predictive Sensing Helps VSI-Super

Case Study I: Surprise-driven Memory Management System for VSI-Super-Order. Our memory management system dynamically compresses and consolidates visual streams based on content surprise. As shown in Fig. 8 (a), we encode incoming frames using sliding window attention with window size W_s. Latent frame prediction module then measures a "surprise level" and assigns to each frame's KV caches. Frames with a surprise level below a predefined threshold T_s are compressed by half before being added to long-term memory. To maintain a stable GPU memory footprint, this

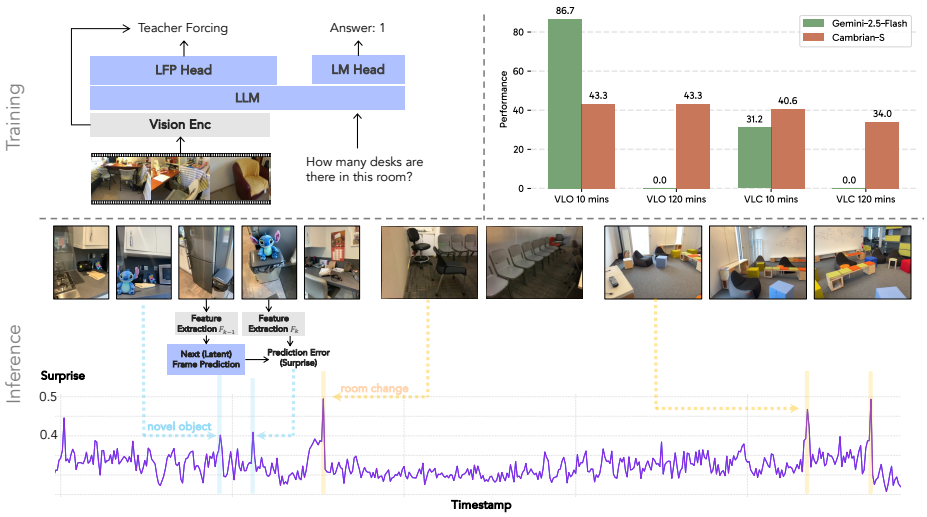


Figure 7: Training and inference pipeline for our latent frame prediction.

long-term memory is constrained to a fixed size $\mathcal{B}_{\text{long}}$. When this limit is reached, a consolidation module merges less surprising frames with adjacent ones (see Fig. 8 (b)). Finally, upon receiving a user query, the system retrieves the top-K most relevant frames from the long-term memory by calculating the cosine similarity between the query and the stored frame features (see Fig. 8 (c)). For more design detail, see Sec. H.3. While there exist related works on designing memory systems for long videos (Song et al., 2024; Zhang et al., 2024a), our focus differs. Rather than developing improved memory architectures, we care more about the potential of using predictive sensing errors (*i.e.*, surprise) as informative indicators.

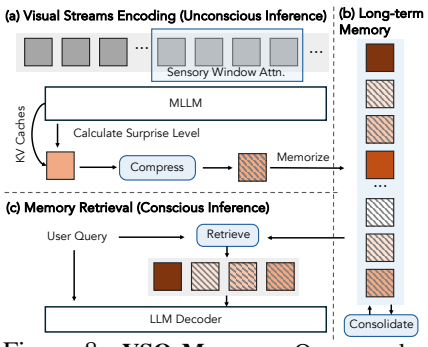


Figure 8: **VSO Memory.** Orange color intensity signifies surprise level. Hatched and solid boxes denote compressed and raw frames.

Results. We compare Cambrian-S with and without the surprise-based memory system, against two advanced proprietary models Gemini-1.5-Flash (Team et al., 2024) and Gemini-2.5-Flash (Comanici et al., 2025), on the VSO benchmark. As shown in Fig. 9, Cambrian-S (w/ Mem.) outperforms Gemini-1.5-Flash and Cambrian-S (w/o Mem.) at all video lengths, demonstrating consistent and remarkable spatial sensing across video lengths. Although Gemini-2.5-Flash yields promising results for videos within an hour, it fails to process longer inputs. On par with Cambrian-S (w/ Mem.) remarkable performance, as shown in Fig. 10, it maintains a stable GPU memory usage across different video lengths. This demonstrates that the unconscious surprise level inference effectively compresses the redundant data without losing critical information.

Ablation on Surprise Measurement. Central to our surprise-based memory system is the method for measuring surprise, as it determines which frames are compressed without foreknowledge of future queries. Here we compare our design, *i.e.*, predictive error as surprise, to another straightforward method: adjacent frame vision feature similarity as surprise. Specifically, we use SigLIP-2 as the vision encoder here. The experiment is conducted on VSO (10 mins version) and sweeps over a shared hyperparameter space for both methods to ensure a fair comparison. As shown in Fig. 11, using predictive error as surprise measurement demonstrates not only superior performance but also greater robustness across different surprise thresholds.

Case Study II: Surprise-driven continual video segment for VSI-SUPER-Count. In the VSI-SUPER-Count benchmark, we segment videos using the prediction error to detect "surprise". This is inspired by the "doorway effect" (Radvansky et al., 2011), a psychological phenomenon where people are more likely to forget items or tasks immediately after walking through a doorway into a different room. Motivated by this effect, our model treats frames with high surprise as space

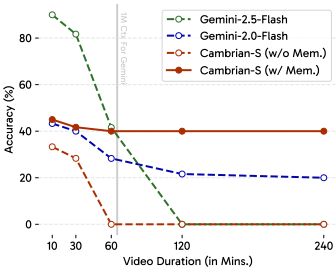


Figure 9: **Accuracy on VSO.**

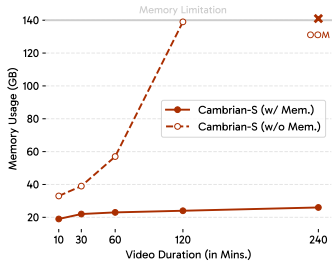


Figure 10: **GPU Memory Usage.**

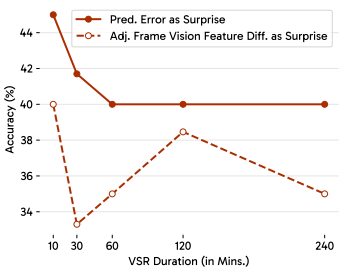


Figure 11: **Different Surprise.**

boundaries, allowing it to partition a long video into shorter but meaningful segments that are answered individually. As Fig. 12 shown, the model continuously buffers low-surprise frames in short-term memory. Upon detecting a high-surprise frame, the buffer is summarized to create a segment answer and then cleared. This process repeats until the end of the video. Finally, the final answer is aggregated by all segment answers.

Results. As shown in Fig. 13, Gemini-1.5-Flash achieves nearly zero performance in VSC, demonstrating the difficulty of this task. Although Gemini-2.5-Flash yields much better results on 10-minute videos, its performance declines rapidly on longer content. In contrast, the segment-and-conquer approach used by Cambrian-*S* (w/ LFP Seg) achieves superior and more stable performance across all video lengths. Segmenting the video using GT scene transitions (*i.e.*, Cambrian-*S* w/ GT Seg) improves performance even further. A deeper analysis in Fig. 15 reveals that Gemini-2.5-Flash's predictions are confined to a limited range. They do not grow as more objects appear in the video, while counts from Cambrian-*S* (w/ LFP Seg) scale correctly with the number of objects.

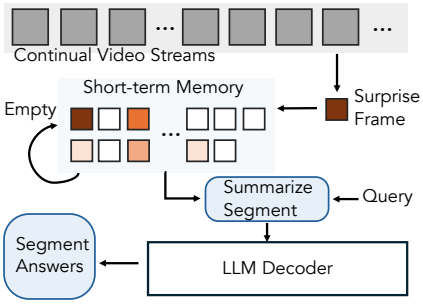


Figure 12: **Framework of VSC.**

Ablation on Surprise Measurement. We compare our surprise measurement against the baseline method which uses adjacent frame similarity to measure the surprise in Fig. 14. For both methods, we report their best results under a set of hyperparameters. As shown in the results, using predictive error as surprise consistently outperforms using appearance similarity as surprise in all cases and by a notable margin.

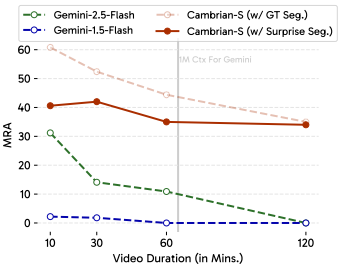


Figure 13: **VSC Results.**

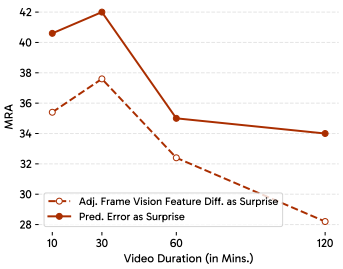


Figure 14: **Surprise measurement.**

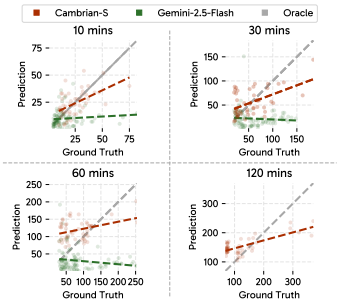


Figure 15: **GT *vs.* Pred. Distribution**

5 Conclusion

We propose *spatial supersensing* as the north star for MLLMs along with a probing benchmark VSI-Super. To examine if the challenge can be addressed by simply scaling data and compute, we train a spatially-grounded MLLM, Cambrian-*S*, on our curated VSI-590K. While Cambrian-*S* excels on standard spatial tasks, it still fails on the VSI-Super benchmark. This failure reveals a fundamental paradigm gap, which we explore by proposing predictive sensing via latent frame prediction. We validate this design through two case studies, which suggest that spatial supersensing requires models to go beyond mere perception, but actively predict the future and update their internal world models from visual streams.

Ethics Statement

This research on video spatial supersensing utilizes publicly available datasets, ensuring that all data complies with privacy regulations. We acknowledge the potential biases that can arise in automatic answer generation, particularly concerning gender, race, or other characteristics. We have taken measures to evaluate and minimize such biases, while remaining committed to further improvements. Additionally, we recognize the potential risks of misuse, such as generating misleading answers, and have checked the training dataset with safeguards against such applications.

Reproducibility Statement

To ensure the reproducibility of our research, we will make all necessary assets publicly available. Our complete source code, including scripts for data processing, model training, and evaluation, will be released. We will also release the weights for all Cambrian-*S* model variants. The curated VSI-590K dataset and the new VSI-Super benchmarks (VSI-Super Order and VSI-Super Count) will be made public to allow for verification and extension of our work. Comprehensive implementation details are provided throughout the paper; specifically, our model architecture, data mixtures, and training hyperparameters are described in Sec. 3 and further detailed in Sec. E. The data curation pipeline is outlined in Sec. 3.2 and Sec. D.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 22

Triantafyllos Afouras, Effrosyni Mavroudi, Tushar Nagarajan, Huiyu Wang, and Lorenzo Torresani. Ht-step: Aligning instructional articles with how-to videos. *Advances in Neural Information Processing Systems*, 36:50310–50326, 2023. 26

Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1534–1543, 2016. 22, 23

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, 2023. 19

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. 19

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023a. 18

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023b. 18

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025a. 1, 18

Yutong Bai, Danny Tran, Amir Bar, Yann LeCun, Trevor Darrell, and Jitendra Malik. Whole-body conditioned egocentric video prediction. *arXiv preprint arXiv:2506.21552*, 2025b. 19

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *CVPR*, 2025. 19

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *NeurIPS*, 2021. URL https://openreview.net/forum?id=tjZjv_qh_CE. 4, 22, 23

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 18

Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Xindong He, Xu Huang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. In *IROS*. IEEE, 2025. 22, 23

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 961–970, 2015. 18

Wenxiao Cai, Yaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. *arXiv preprint arXiv:2406.13642*, 2024. 19

Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 18

Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 18

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Fei-Fei Li. Hourvideo: 1-hour video-language understanding. *Advances in Neural Information Processing Systems*, 37:53168–53197, 2024. 3, 18

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *CVPR*, 2024a. 19

Chang Chen, Fei Deng, Kenji Kawaguchi, Caglar Gulcehre, and Sungjin Ahn. Simple hierarchical planning with diffusion. In *ICLR*, 2024b. 19

Dongping Chen, Yue Huang, Siyuan Wu, Jingyu Tang, Liuyi Chen, Yilin Bai, Zhigang He, Chenlong Wang, Huichi Zhou, Yiqiang Li, et al. Gui-world: A video benchmark and dataset for multimodal gui-oriented understanding. *arXiv preprint arXiv:2406.10819*, 2024c. 26

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18407–18418, 2024d. 4

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*, 2024e. 19

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024f. 18

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. In *NeurIPS*, 2024. 19

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 4, 8

Kenneth James Williams Craik. *The nature of explanation*. CUP Archive, 1967. 19

Erfei Cui, Yinan He, Zheng Ma, Zhe Chen, Hao Tian, Weiyun Wang, Kunchang Li, Yi Wang, Wenhai Wang, Xizhou Zhu, Lewei Lu, Tong Lu, Yali Wang, Limin Wang, Yu Qiao, and Jifeng Dai. Sharegpt-4o: Comprehensive multimodal annotations with gpt-4o, 2024. URL https://sharegpt4o.github.io/. 5, 26

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 4, 22, 23

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pp. 933–941. PMLR, 2017. 25

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation. *Advances in Neural Information Processing Systems*, 35:5982–5994, 2022. 22, 23

David Fan, Shengbang Tong, Jiachen Zhu, Koustuv Sinha, Zhuang Liu, Xinlei Chen, Michael Rabbat, Nicolas Ballas, Yann LeCun, Amir Bar, et al. Scaling language-free visual representation learning. *arXiv preprint arXiv:2504.01017*, 2025. 18

Karl Friston. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2): 127–138, 2010. 6, 19

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 3, 4, 18, 20

Carlos E Garcia, David M Prett, and Manfred Morari. Model predictive control: Theory and practice—a survey. *Automatica*, 25(3):335–348, 1989. 19

Quentin Garrido, Nicolas Ballas, Mahmoud Assran, Adrien Bardes, Laurent Najman, Michael Rabbat, Emmanuel Dupoux, and Yann LeCun. Intuitive physics understanding emerges from self-supervised pretraining on natural videos. *arXiv preprint arXiv:2502.11831*, 2025. 7, 19

James J Gibson. *The ecological approach to visual perception: classic edition*. Psychology press, 2014. 1

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022. 26

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 19

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. 19

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022. 18

Jakob Hohwy. *The predictive mind*. OUP Oxford, 2013. 19

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-MMMU: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025. 3, 18

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 18

Jindong Jiang, Xiuyu Li, Zhijian Liu, Muyang Li, Guo Chen, Zhiqi Li, De-An Huang, Guilin Liu, Zhiding Yu, Kurt Keutzer, et al. Token-efficient long video understanding for multimodal llms. *arXiv preprint arXiv:2503.04130*, 2025. 19

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*, 2024. 19

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pp. 5156–5165. PMLR, 2020. 19

Nicholas GW Kennedy, Jessica C Lee, Simon Killcross, R Fred Westbrook, and Nathan M Holmes. Prediction error determines how memories are organized in the brain. *Elife*, 13:RP95849, 2024. 6

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. 18

Bruno Korbar, Yongqin Xian, Alessio Tonioni, Andrew Zisserman, and Federico Tombari. Text-conditioned resampler for long form video understanding. In *European Conference on Computer Vision*, pp. 271–288. Springer, 2024. 19

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a. 1, 18, 25

Chengzu Li, Caiqi Zhang, Han Zhou, Nigel Collier, Anna Korhonen, and Ivan Vulić. Topviewrs: Vision-language models as top-view spatial reasoners. *arXiv preprint arXiv:2406.02537*, 2024b. 19

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023a. 18

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023b. 18

Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *European Conference on Computer Vision*, pp. 237–255. Springer, 2024c. 19

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. MVbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024d. 3, 26

Wei Li, Bing Hu, Rui Shao, Leyang Shen, and Liqiang Nie. Lion-fs: Fast & slow video-language thinker as online video assistant. *arXiv preprint arXiv:2503.03663*, 2025. 19

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024e. 4, 19

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pp. 323–340. Springer, 2024f. 18

Benlin Liu, Yuhao Dong, Yiqin Wang, Yongming Rao, Yansong Tang, Wei-Chiu Ma, and Ranjay Krishna. Coarse correspondence elicit 3d spacetime understanding in multimodal language model. *arXiv preprint arXiv:2408.00754*, 2024a. 19

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 1, 18, 25

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024b. 25

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pp. 38–55. Springer, 2024c. 22

Yiren Lu, Yunlai Zhou, Disheng Liu, Tuo Liang, and Yu Yin. Bard-gs: Blur-aware reconstruction of dynamic scenes via gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16532–16542, 2025. 19

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Videogpt+: Integrating image and video encoders for enhanced video understanding. *arXiv preprint arXiv:2406.09418*, 2024. 26

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023. 3, 4, 18

Kevis-Kokitsi Maninis, Kaifeng Chen, Soham Ghosh, Arjun Karpur, Koert Chen, Ye Xia, Bingyi Cao, Daniel Salz, Guangxing Han, Jan Dlabal, et al. Tips: Text-image pretraining with spatial awareness. *arXiv preprint arXiv:2410.16512*, 2024. 25

David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010. 1

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2630–2640, 2019. 26

Muhammad Ferjad Naeem, Yongqin Xian, Xiaohua Zhai, Lukas Hoyer, Luc Van Gool, and Federico Tombari. Silc: Improving vision language pretraining with self-distillation. In *European Conference on Computer Vision*, pp. 38–55. Springer, 2024. 25

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903. IEEE, 2024. 22, 23

Junwen Pan, Rui Zhang, Xin Wan, Yuan Zhang, Ming Lu, and Qi She. Timesearch: Hierarchical video search with spotlight and reflection for human-like long video understanding. *arXiv preprint arXiv:2504.01407*, 2025. 19

Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20133–20143, 2023. 22, 23

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36: 42748–42761, 2023. 18, 26

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021. 18, 25

Gabriel A Radvansky, Sabine A Krawietz, and Andrea K Tamplin. Walking through doorways causes forgetting: Further explorations. *Quarterly journal of experimental psychology*, 64(8):1632–1645, 2011. 8

Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 1999. 19

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 22

Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, and Kate Saenko. SAT: Spatial Aptitude Training for Multimodal Language Models. In *COLM*, 2025. 5

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14313–14323, 2024. 26

Weiming Ren, Wentao Ma, Huan Yang, Cong Wei, Ge Zhang, and Wenhu Chen. Vamba: Understanding hour-long videos with hybrid mamba-transformers. *arXiv preprint arXiv:2503.11579*, 2025. 19

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10912–10922, 2021. 22, 23

Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation models. *arXiv preprint arXiv:2410.23266*, 2024. 3

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024. 18, 19

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15768–15780, 2025. 19

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18221–18232, 2024. 4, 8, 18, 26

Aimee E Stahl and Lisa Feigenson. Observing the unexpected enhances infants' learning and exploration. *Science*, 2015. 6

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 18

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 1, 8

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri Iyer, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs. In *NeurIPS*, 2024. 1, 18, 25

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a. 18

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b. 18

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 5, 18, 25

Ujjwal Upadhyay, Mukul Ranjan, Zhiqiang Shen, and Mohamed Elhoseiny. Time blindness: Why video-language models can't see what humans can? *arXiv preprint arXiv:2505.24867*, 2025. 19

Paul Voigtlaender, Soravit Changpinyo, Jordi Pont-Tuset, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with video localized narratives. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2461–2471, 2023. 26

Hermann Von Helmholtz. *Handbuch der physiologischen Optik*, volume 9. L. Voss, 1867. 1, 6, 19

Bo Wan, Michael Tschannen, Yongqin Xian, Filip Pavetic, Ibrahim M Alabdulmohsin, Xiao Wang, André Susano Pinto, Andreas Steiner, Lucas Beyer, and Xiaohua Zhai. Locca: Visual pretraining with location-aware captioners. *Advances in Neural Information Processing Systems*, 37:116355–116387, 2024. 25

Alex Jinpeng Wang, Linjie Li, Kevin Qinghong Lin, Jianfeng Wang, Kevin Lin, Zhengyuan Yang, Lijuan Wang, and Mike Zheng Shou. Cosmo: Contrastive streamlined multimodal model with interleaved pre-training. *arXiv preprint arXiv:2401.00849*, 2024a. 26

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. *arXiv preprint arXiv:2503.11651*, 2025. 22

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b. 18

Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020. 19

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. LVBench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024c. 18

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pp. 396–416. Springer, 2024d. 1

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5288–5296, 2016. 18

Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5036–5045, 2022. 26

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024a. 5

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024b. 25

Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. Vript: A video is worth thousands of words. *Advances in Neural Information Processing Systems*, 37:57240–57261, 2024c. 26

Jihan Yang, Runyu Ding, Ellis Brown, Xiaojuan Qi, and Saining Xie. V-irl: Grounding virtual intelligence in real life. In *ECCV*, 2024d. 19

Jihan Yang, Shusheng Yang, Anjali Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in Space: How Multimodal Large Language Models See, Remember and Recall Spaces. *CVPR*, 2024e. 1, 3, 4, 5, 19, 20, 21, 22

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, et al. Egolife: Towards egocentric life assistant. *arXiv preprint arXiv:2503.03803*, 2025. 26

Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *ICLR*, 2024f. 19

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *ICCV*, 2023. 4, 22, 23

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023. 18, 25

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 18

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-vstream: Memory-based real-time understanding for long video streams. *arXiv preprint arXiv:2406.08085*, 2024a. 4, 8

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024b. 19

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024c. URL https://arxiv.org/abs/2410.02713. 5, 18, 22, 26

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. In *ICML*, 2025. 19

Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 18

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv preprint arXiv:2409.18125*, 2024. 19

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 18

Language Model Use Statement

Large language models (LLMs) were used only for light editorial purposes, such as minor grammar checking and language polishing. They were not used for generating scientific content, research ideation, experiment design, or analysis. The authors take full responsibility for the entirety of the paper, and LLMs are not considered contributors or eligible for authorship.

A Appendix Outline

This supplementary material provides additional details for our work and is organized as follows:

- **Related Work.** A summary of key related works, including recent advancements in video MLLMs, visual spatial intelligence, and memory mechanisms within LLM and MLLM fields.
- **VSI-Super Details.** Further details about our new VSI-Super benchmarks.
- **VSI-590K Dataset: Additional Details.** Further details on the VSI-590K dataset, covering its construction process, task taxonomy (or taskonomy, if specific to your field), and supplementary examples.
- **Model Implementation Details: Our upgraded Cambrian-1 and Cambrian-S.** A detailed account of the implementation of our upgraded Cambrian-1 and Cambrian-S, including architecture, data processing, and training procedures.
- **Additional Benchmark Results and Analysis of Cambrian-1.** Supplementary benchmark results and comparative analysis for Cambrian-1.
- **Cambrian-*S* Training Recipe Ablation Study.** Ablation study on the influence of base video model and data recipe.
- **Predictive Sensing.** Further specifics on the training and inference implementation of our latent frame prediction and ablation study, with addditional details about our memory framework.
- **Discussion, Limitation, and Future work.** We discuss the limitations of our current work and outline potential directions for further improvement and exploration.

B Related Work

B.1 Video Multimodal Large Language Model

The unprecedented language and reasoning capabilities of large-scale pretrained LLMs (Brown et al., 2020; Touvron et al., 2023a; Bai et al., 2023a; Touvron et al., 2023b), coupled with well-developed visual feature extractors (Radford et al., 2021; Zhai et al., 2023; Tschannen et al., 2025; He et al., 2022; Fan et al., 2025), have driven significant advancements in empowering LLMs to understand visual content like still images (Hurst et al., 2024; Liu et al., 2023; Li et al., 2024a; Bai et al., 2023b; Tong et al., 2024; Team et al., 2023; Chen et al., 2024f; Wang et al., 2024b; Li et al., 2023a), and spurred a growing interest in building video MLLMs (Li et al., 2024f;a; Zhang et al., 2024c; Song et al., 2024; Bai et al., 2025a; Zhu et al., 2025; Zhang et al., 2023; Li et al., 2023b; Shen et al., 2024) which is considered as a key step toward grounding MLLMs to real world applications like embodied agents (Kim et al., 2024).

Despite great progress has been witnessed on building more competitive video MLLMs and better benchmarks (Fu et al., 2024; Hu et al., 2025; Chandrasegaran et al., 2024; Mangalam et al., 2023; Wang et al., 2024c; Patraucean et al., 2023) to properly evaluate their capabilities, in this paper, we argue that current video MLLMs and benchmarks majorly focus on the *recognition* capabilities (Caba Heilbron et al., 2015; Zhou et al., 2018; Carreira et al., 2018; 2019; Xu et al., 2016) while overlook other critical properties inherently in video modality. Instead, our work concentrates on other two missing pieces which are of critical importance for building the real *spatial supersensing* intelligence: the ability to *understand the space* and *perceive in a continuous manner*.

B.2 Spatial Understanding

Spatial understanding, the ability to perceive and comprehend spatial relationships within an environment, is crucial for embodied agents to effectively interact with the real world. Unlike recognition abilities, which primarily align with language semantics, spatial understanding is more physically grounded, presenting a significant challenge for current Multimodal Large Language Models (MLLMs). While recent efforts have aimed to enhance MLLMs' spatial understanding (Yang et al., 2024d; Chen et al., 2024a; Cheng et al., 2024; Cai et al., 2024; Liu et al., 2024a; Li et al., 2024b; Zhu et al., 2024; Song et al., 2025; Lu et al., 2025; Upadhyay et al., 2025), most focus on static images, which poorly reflect real-world embodied scenarios. The most relevant work to ours is Thinking in Space (Yang et al., 2024e), which proposes a video-based benchmark with several primitives to assess MLLMs' spatial intelligence. Building on this, our work introduces VSI-590K dataset to advance the visual spatial capabilities of video MLLMs.

B.3 Always-on Video Understanding

Humans effortlessly perceive and process a continuous—potentially infinite—stream of visual signals from their surroundings, both intentionally and subconsciously. Equipping embodied agents or lifelong assistants with similar capabilities is essential for enabling continuous learning and adaptation through real-world interaction. However, the unbounded length of video streams poses a major challenge for current Video Multi-Modal Large Language Models (MLLMs), primarily due to escalating computational and storage demands. Recent works have attempted to address this challenge from several perspectives:

- *Efficient architectural design*. The quadratic complexity of self-attention becomes a bottleneck for long video sequences. Inspired by advances in language modeling, some approaches (Li et al., 2024c; Ren et al., 2025) adopt more efficient architectures (e.g., linear or sub-quadratic attention (Wang et al., 2020; Gu & Dao, 2023; Katharopoulos et al., 2020)) to reduce computational overhead and accommodate longer inputs.
- *Context window expansion*. The fixed-length context window of pre-trained LLMs inherently limits their ability to comprehend extended temporal content. Expanding this window (Chen et al., 2024e; Zhang et al., 2024b) allows models to process and reason over longer video segments.
- *Retrieval-augmented video understanding*. To handle long video streams, some methods retrieve relevant segments from a larger corpus (Korbar et al., 2024; Pan et al., 2025), using them as context for downstream understanding tasks.
- *Visual token reduction or compression*. Reducing the number of visual tokens (either per frame or across frames) (Shen et al., 2024; Li et al., 2024e; Jiang et al., 2025; Li et al., 2025) helps manage long video sequences by shortening the effective input length.

B.4 Predictive Modeling

A learned internal predictive model (Craik, 1967; Ha & Schmidhuber, 2018) allows an intelligent agent to represent and simulate aspects of its environment, enabling more effective planning and decision-making. Model predictive control (MPC) (Garcia et al., 1989) applies similar principles in control theory, leveraging internal forward models to anticipate future trajectories and select optimal actions in real time. This concept draws inspiration from how humans form mental models of the world (Rao & Ballard, 1999; Hohwy, 2013; Friston, 2010) and how these internal representations influence behavior (*e.g.*, *unconscious inference* (Von Helmholtz, 1867)), serving as simplified abstractions of reality that enable prediction and efficient action. A growing body of work has explored the idea of predictive modeling through self-supervised representation learning (Assran et al., 2023; 2025), and text- or action-conditioned video generation (Zhou et al., 2025; Yang et al., 2024f; Bar et al., 2025; Chen et al., 2024b; Bai et al., 2025b; Garrido et al., 2025; Kang et al., 2024). In this paper, motivated by how humans leverage internal world models to process unbounded sensory input efficiently and effectively, we investigate how to equip MLLMs with a similar predictive sensing capability.

Figure 16: Illustrations of how spatial sensing is conceptualized in current video benchmarks. The left panel features examples from the "spatial reasoning" subcategory of VideoMME (Fu et al., 2024), including a visual-effects clip of "`What if the Moon Crashed into the Earth?`" from *Shutter Authority*, and the ground-truth answer refers to the gravitational pull of the Moon—an explanation that is physically impossible.) 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 (Yang et al., 2024e), which highlight visual-spatial reasoning tasks such as object counting, identifying relative directions, route planning, and related scenarios.

C VSI-SUPER DETAILS

C.1 VSO BENCHMARK

Figure 19 shows more edited frame examples of our VSI-SUPER-Order benchmark. By using image editing models, the generated edited frames can be really realistic.

C.2 VSC BENCHMARK

VSC is constructed by concatenating multiple space-scanning clips from VSI-Bench, and the task is to determine the total count of a specific object across the entire concatenated video (see Fig. 4). We construct the benchmark with 4 different video durations, from 10 minutes to 120 minutes, to throughly reflect the genralizability of MLLMs' spatial counting ability. For metric, following VSI-Bench, we choose to use $\mathcal{MRA}$ starting from 0.5 to 0.95 as the major metric.

C.3 HUMAN PERFORMANCE ON VSI-SUPER

To evaluate human performance on VSI-SUPER, we recruited 10 volunteers to complete two tasks: VSC (10 mins) and VSO (60 mins), which includes 50 and 60 questions, respectively. As shown in Tab. 4, humans achieved near-perfect results on VSO with 95.2% accuracy, and significantly outperformed MLLMs on VSC (76.5% *vs.* 32.1%).

Table 4: We analyzed human performance on VSI-SUPER and found it to be significantly superior to that of Gemini-2.5-Flash.

	Metric	Human Performance	Gemini-2.5-Flash
VSC (10mins)	MRA	76.5	32.1
VSO (60mins)	Acc.	95.2	34.7

D ADDITIONAL DETAILS OF THE VSI-590K DATASET

In this section, we provide more details for the dataset, including the question type definition, question-answer pair construction pipeline, and some examples for each data source.

D.1 QUESTION TYPE DEFINITION

We define 12 question types across a spatiotemporal taxonomy to create a comprehensive and diverse set of questions for instruction-tuning. We define five main question types—`size`,

`direction`, `count`, `distance`, and `appearance order`—broadly categorized as measuring `configuration`, `measurement`, or `spatiotemporal` capabilities following VSI-Bench. For all question types except `appearance order`, we also define *relative* and *absolute* versions of each question type, as both relative and absolute judgments are crucial to visual-spatial understanding (Yang et al., 2024e). For example, for `size`, we ask for both size comparison between two objects and the metric dimensions of an object. To increase diversity, we vary the perspective from which `direction` and `distance` questions are formulated. For example, for `distance`, we may ask which of two objects is closer to the camera or which of two objects is closer to a third, different object. Finally, we further augment diversity by varying both *phrasing* and *measurement units* for each question.

Taxonomy. When curating visual-spatial intelligence supervised fine-tuning datasets, an important perspective is how to define the question type. Inspired by (Yang et al., 2024e), we expand its task definition in a more systematic manner. As shown in Tab. 5, we distinguish these question types in four perspectives:

- `Spatial-temporal attributes:` we categorize questions into five distinct spatial-temporal attribute types: size (comparing or measuring object/space dimensions), direction (orientation in space), count (enumeration of objects), distance (proximity between objects), and appearance order (temporal sequence of objects appearing in videos).
- `Relative versus absolute:` questions are classified as relative when they involve comparison between multiple objects (e.g., "which is larger?"), or absolute when they require specific measurements or quantities (e.g., "what is the height in meters?"). This distinction applies across most attribute types.
- `Perspective taking:` this dimension captures the viewpoint from which spatial relationships are evaluated. Questions may be posed from the camera's perspective (e.g., "from the camera's perspective, is the object on the left or right?") or from the perspective of specific objects in the scene (e.g., "facing the object$_1$ from object$_2$...")
- `Modality:` questions are categorized based on whether they can be answered using static images only, or require dynamic video information. Some attribute types like appearance order are only applicable to videos, while others like size can be addressed in either modality.

Table 5: **Taxonomy of spatiotemporal question types in VSI-590K**. Questions in VSI-590K are stratified along five axes: attribute type, relative versus absolute, perspective, modality, and question template.

Types	Rel./Abs.	Perspective	Modality	Example template
Size	Relative	—	Video / Image	"Between {*object*$_1$} and {*object*$_2$}, which is larger?"
	Absolute	—	Video / Image	"What is the height of the {*object*} in {*unit*}?"
	Absolute	—	Video / Image	"What is the room's size in {*unit*}?"
Direction	Relative	Camera	Image	"From the camera's perspective, is the {*object*} on the left or the right?"
	Relative	Object	Video / Image	"Facing the {*object*$_1$} from the {*object*$_2$}, would the {*object*$_3$} be placed left, right, or back?"
	Absolute	Object	Video / Image	"Standing at {*object*$_1$}, facing toward {*object*$_2$}, how far clockwise do I rotate (in degrees) to see the {*object*$_3$}?"
Count	Relative	—	Video / Image	"Are there fewer {*object*$_1$} than {*object*$_2$} ?"
	Absolute	—	Video / Image	"How many {*object*} are present?"
Distance	Relative	Camera	Image	"Which object is closer to the camera, the {*object_1*} or the {*object_2*}?"
	Relative	Object	Video / Image	"Which is nearer to the {*object_3*}, the {*object_1*} or the {*object_2*}?"
	Absolute	Object	Video / Image	"What is the distance between the {*object_1*} and the {*object_2*} in {*unit*}?"
Appearance Order	—	—	Video	"Determine how {*object_1*}, {*object_2*}, {*object_3*}, and {*object_4*} are ordered by their initial appearances in the video"

Question templates augmentation question types. Besides, as shown in Tab. 5, for each question type, we provide adequate templates to prevent MLLMs from overfitting to specific formats or

measurement units. These diverse templates were initially created by humans, then augmented using GPT-4o (Achiam et al., 2023), and finally validated and corrected by human reviewers. We provide concrete question templates in Tabs. 21 to 33.

D.2 Detailed QA-Pair Construction Pipeline

We construct VSI-590K from a diverse span of data sources and types (*i.e.*, simulated and real). This creates a dataset significantly stronger than a highly homogeneous dataset of a similar size. See Tab. 6 for the data sources and for dataset statistics on the number of videos, images, and QA pairs from each dataset. Below we describe how our main data source types are processed to generate question-answer pairs.

- *Annotated Real Videos.* As proposed in VSI-Bench (Yang et al., 2024e), multimodal visual-spatial reasoning requires 3D geometric and spatial understanding. In this regard, we first follow VSI-Bench to re-purpose the training split of existing indoor scans and ego-vision datasets containing 3D instance-level annotations, including S3DIS (Armeni et al., 2016), ScanNet (Dai et al., 2017), ScanNet++ V2 (Yeshwanth et al., 2023), ARKitScenes (Baruch et al., 2021), and ADT (Pan et al., 2023). For each dataset, the annotations are first organized into a meta-information file containing the attributes that describes each scene: object counts by category, object bounding boxes, room size, and more. Question templates are then automatically propagated to generate a plethora of questions.
- *Simulated Data.* Given the scarce nature of 3D-annotated data, it is impossible to collect a very large-scale and diverse 3D-annotated SFT dataset solely by relying on annotated real videos. We leverage embodied simulators to programmatically generate spatially grounded video trajectories and QA pairs. We render 625 videos traversals through ProcTHOR (Deitke et al., 2022) scenes with diverse layouts, object placements, and appearances. We adapt the same methodology to Hypersim (Roberts et al., 2021), sampling 5,113 images from 461 indoor scenes; given instance-level bounding boxes, we construct supervision consistent with our annotated real-video setup.
- *Unannotated Real Videos.* Web-sourced videos, despite unannotated, provide rich diversity in room types, regions, and layouts. We web-crawled around 19K room tour videos from YouTube, and also source videos from the robotic learning datasets Open-X-Embodiment (O'Neill et al., 2024) and AgiBot-World (Bu et al., 2025). Since these videos lack the necessary 3D annotations for curating spatial instruction-tuning data, we build a pseudo-annotation pipeline. As shown in Algorithm 1, we implement a multi-stage processing pipeline. We begin by sampling frames at regular intervals and filtering out blurry images. For each valid frame, we employ the open-vocabulary object detector Grounding-DINO (Liu et al., 2024c) with predefined categories of interest. When a frame contains sufficient valid objects, we use SAM2 (Ravi et al., 2024) to extract instance-wise semantic masks. Besides, to transform 2D image content into 3D representations, we employ VGGT (Wang et al., 2025) to extract 3D point sets for each image and integrate them with the previously generated instance masks. Notably, we apply an erosion algorithm to refine the instance masks, which mitigates inaccurate point cloud estimations at object boundaries. This pipeline has enabled us to create pseudo-annotations from approximately 19,000 room tour videos from YouTube and robotic learning datasets, yielding diverse spatial question-answer pairs across various room types and layouts without manual 3D annotations. By processing individual frames rather than complete videos, our pipeline ensures higher quality semantic extraction and more reliable reconstruction results, avoiding the noise and inconsistent issues typically encountered when applying reconstruction and semantic extraction techniques to entire video sequences.

D.3 VSI-590K Data Source Ablation

To evaluate the effectiveness of our proposed VSI-590K dataset, we perform an ablation study by finetuning the improved Cambrian-1 image MLLM described in Section 3.1 with part of the video instruction tuning samples from LLaVA-Video (Zhang et al., 2024c). This model serves as the *baseline* in Tab. 7. We measure the contribution of each dataset source by conducting individual and combined fine-tuning with this model. Notably, fine-tuning on the full VSI-590K mixture yields

Table 6: **Statistics for VSI-590K.** The curated data draws from 10 sources to improve diversity.

Dataset	# Videos	# Images	# QA Pairs
Annotated Real Videos			
S3DIS (Armeni et al., 2016)	199	-	5,187
Aria Digital Twin (Pan et al., 2023)	183	-	60,207
ScanNet (Dai et al., 2017)	1,201	-	92,145
ScanNet++ (Yeshwanth et al., 2023)	856	-	138,701
ARKitScenes (Baruch et al., 2021)	2,899	-	57,816
Simulated Data			
ProcTHOR (Deitke et al., 2022)	625	-	20,092
Hypersim (Roberts et al., 2021)	-	5,113	176,774
Unannotated Real Videos			
YouTube Room Tour	-	20,100	20,100
Open X-E (O'Neill et al., 2024)	-	14,801	14,801
AgiBot-World (Bu et al., 2025)	-	4,844	4,844
Total	5,963	44,858	590,667

Algorithm 1: QA Generation Pipeline for Unannotated Web-crawled Video

Input: Video sequence V, valid category list $\mathcal{C}_{\text{valid}}$, invalid category list $\mathcal{C}_{\text{invalid}}$, sampling interval Δt, blur threshold τ_{blur}, minimum object count θ_{min}, minimum 3D point count θ_{3D}, erosion kernel K_{erosion}
Output: Selected frame set $\mathcal{F}$, Question-answer pairs $\mathcal{Q}$

Initialize $\mathcal{F} \leftarrow \emptyset$, $\mathcal{Q} \leftarrow \emptyset$;
$\mathcal{S} \leftarrow \text{SampleFrames}(V, \Delta t)$; `// Sample frames at interval` Δt
foreach *frame* $f \in \mathcal{S}$ **do**
  if *BlurDetection*$(f) > \tau_{blur}$ **then**
    continue;
  $\mathcal{O} \leftarrow \text{GroundingDINO}(f, \mathcal{C}_{\text{valid}} \cup \mathcal{C}_{\text{invalid}})$; `// Detect objects from both category lists`
  if $\exists o \in \mathcal{O} : \textit{category}(o) \in \mathcal{C}_{\textit{invalid}}$ **then**
    continue;
  $\mathcal{O}_{\text{valid}} \leftarrow \{o \in \mathcal{O} : \text{category}(o) \in \mathcal{C}_{\text{valid}}\}$;
  if $|\mathcal{O}_{\textit{valid}}| < \theta_{\textit{min}}$ **then**
    continue;
  $\mathcal{M} \leftarrow \emptyset$; `// Initialize mask set`
  foreach *object* $o \in \mathcal{O}_{\textit{valid}}$ **do**
    $b \leftarrow \text{GetBoundingBox}(o)$;
    $m \leftarrow \text{SAM2}(f, b)$; `// Generate mask using SAM2`
    $m' \leftarrow \text{Erode}(m, K_{\text{erosion}})$; `// Apply erosion on the masks`
    $\mathcal{M} \leftarrow \mathcal{M} \cup \{m'\}$;
  $\mathcal{P}_{\text{map}} \leftarrow \text{VGGT}(f)$; `// Generate 3D point map using VGGT`
  $\mathcal{P} \leftarrow \emptyset$; `// Initialize 3D point set`
  foreach *mask* $m \in \mathcal{M}$ **do**
    $P \leftarrow \text{ExtractMaskedPoints}(m, \mathcal{P}_{\text{map}})$; `// Extract 3D points covered by mask`
    if $|P_{\textit{valid}}| \geq \theta_{\textit{3D}}$ **then**
      $\mathcal{P} \leftarrow \mathcal{P} \cup \{P\}$;
  if $|\mathcal{P}| > 0$ **then**
    $q \leftarrow \text{QAGenerator}(\mathcal{P})$; `// Generate QA pairs from 3D geometry`
    $\mathcal{Q} \leftarrow \mathcal{Q} \cup \{q\}$;
    $\mathcal{F} \leftarrow \mathcal{F} \cup \{f\}$;
Return $\mathcal{F}$, $\mathcal{Q}$;

the highest performance on most video spatial reasoning tasks, clearly surpassing the baseline and single-source variants. Furthermore, we observe a clear hierarchy in the utility of different data source groups for improving visual-spatial understanding: annotated real videos provide the most significant improvements, followed by simulated data, and lastly pseudo-annotated images. This suggests that videos are more valuable than images for spatial reasoning, as training exclusively on video data (as opposed to single images) produces superior performance on both video and image spatial reasoning

Table 7: **Cumulative benefits of the data mixture in VSI-590K.** Our proposed dataset, VSI-590K (All-in-One), yields by far the strongest performance on VSI-Bench. Annotated real video sources provide the most benefit, followed by simulated data, then pseudo-annotated images. [1]RealWorldQA

	Image			VSI-Bench (Video)								
VSI Data Mixture	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	65.4	54.9	75.3	41.6	63.8	21.0	44.9	37.0	43.8	47.4	34.0	41.1
+ ADT	65.9	56.5	77.5	41.0	51.0	29.8	52.5	40.2	42.3	38.8	34.0	39.8
+ 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	56.3	70.9	37.9	67.5	59.3	57.0	46.7	35.1	76.1
+ ScanNet++ V2	66.1	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	62.2	55.7	74.9	36.4	21.0	29.7	49.3	3.8	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**

Table 8: **VSI-590K task group ablations on VSI-Bench.** We report the performance on VSI-Bench by deducting different single sub-tasks or a certain group of sub-tasks.

Types	Avg	Obj Ct	Abs Dst	Obj Sz	Rm Sz	Rel Dst	Rel Dir	Rte Pln	Ap Ord
Baseline	63.2	73.5	49.4	71.4	70.1	66.9	61.5	36.6	76.6
Task Groups									
- Configuration	47.5	48.7	44.2	70.4	60.5	46.2	39.1	27.3	43.9
- Measurement	43.1	73.1	13.9	32.3	27.7	66.5	55.1	33.5	43.0
- Spatiotemporal	58.1	73.7	47.7	70.9	65.2	68.3	58.9	32.5	47.6

benchmarks. This aligns with the intuition that the temporal and multi-view nature in of videos aids in developing robust spatial understanding.

In Tab. 8, we ablate three task groups of our VSI-590K. We notice that, configuration data contribute the most to route planning, while measurement contribute the least.

D.4 Examples of VSI-590K

To better illustrate VSI-590K, we provide qualitative visualization results in Figs. 20 to 26. These visualizations demonstrate that VSI-590K delivers great diversity and quality for spatial question-answering supervised fine-tuning.

D.5 Error Analysis of Pseudo Annotation Pipeline

To analyze the systematic errors in our pseudo-annotation pipeline, we randomly sampled 500 question-answer pairs from the pseudo-annotated data and manually verified their correctness. As shown in Tab. 9, we find that while some systematic errors are present, the overall data quality is satisfactory. This finding is consistent with our observation in Tab. 7, which indicates that pseudo-annotated data contributes the least to performance on VSI-Bench.

E Model Implementation Details

This section elaborates the details of both our upgraded Cambrian-1 and Cambrian-S, from the architecture design, to data and training recipes.

Table 9: **Error analysis of pseudo-annotation pipeline.**

Question Type	Metrics	Human Performance
Abs. Count	MRA	64.2
Rel. Count	ACC	65.0
Rel. Dir. (Camera)	ACC	83.8
Rel. Dir. (Object)	ACC	64.0
Rel. Dist. (Camera)	ACC	86.3
Rel. Dist. (Camera)	ACC	78.6

E.1 Base Architecture

Following the original Cambrian-1 (Tong et al., 2024) and common practices in most MLLMs (Liu et al., 2023; Li et al., 2024a), our model (both Cambrian-1† and Cambrian-S) integrates a pre-trained vision encoder, a pre-trained language model as the decoder, and a vision-language adapter to bridge these two modalities. Specifically, we employ SigLIP2-So400M (Tschannen et al., 2025) as the vision encoder. This encoder was trained using a combination of losses: text next-token-prediction (LocCa (Wan et al., 2024)), image-text contrastive (Sigmoid (Radford et al., 2021; Zhai et al., 2023)), and masked self-prediction (SILC (Naeem et al., 2024)/TIPS (Maninis et al., 2024)). For the language model, we utilize the instruction-tuned Qwen2.5-7B model (Yang et al., 2024b). Unlike Cambrian-1, which used SVA for a deeper vision-language fusion, Cambrian-1† and Cambrian-S employ a simpler GELU-activated (Dauphin et al., 2017) two-layer MLP as the vision-language adapter to maintain a balance between performance and efficiency.

Table 10: Training Configuration for Stage 1 and Stage 2

	Stage 1 (Alignment)	Stage 2 (Instruction Tuning)
Model		
Vision Encoder	SigLIP2-So400M	
Language Decoder	Qwen2.5-7B-Instruct	
VL-Adapter	2×MLP-GELU	
Data Recipe		
Data	Cambrian-Alignment-2.5M	Cambrian-7M
Image Resolution	Pad (384×384)	AnyRes (At most 9 sub-images)
# of Tokens / Image	729	At most 7290
Training Recipe		
Max Sequence Length	2048	8192
Trainable Module	VL-Adapter	VL-Adapter & LLM
Learning Rate	1×10^{-3}	1×10^{-5}
Batch Size	512	256
Warmup Ratio	0.06	0.03

E.2 Stage 1 & 2: Cambrian-1 Training

Our upgraded base image MLLM Cambrian-1 is trained using a 2-stage training recipe, similar to Cambrian-1.

Stage 1: Vision-Language Alignment. We freeze most of the model's parameters, training only the vision-language adapter on the Cambrian-Alignment-2.5M dataset. Input images are padded to a fixed resolution of 384×384, and the maximum sequence length is set to 2048.

Stage 2: Instruction Tuning. We unfreeze both the vision-language adapter and the LLM decoder, keeping the vision encoder frozen. The model is then fine-tuned on the Cambrian-7M image-text dataset. Slightly different from Cambrian-1, we adopt an any-resolution strategy (Liu et al., 2024b) during this stage. More specifically, input images are resized to a certain resolution while maintaining their aspect ratio. These resized images are then divided into several 384×384 sub-images. This approach enables the model to handle input images at higher and more dynamic resolutions.

Table 10 details the first 2 training stages' setups. Noteworthy, stage 1 and stage 2 costs ∼1000 and ∼9700 TPU-v4-core hours for a 7B model, respectively.

E.3 Stage 3 & 4: Cambrian-*S* Training

Cambrian-S is created by fine-tuning the Cambrian-1 base model through two additional video-centric training stages. While the architecture remains the same, these stages adapt the model from a static image understanding expert to a dynamic, spatially-aware video reasoner.

Stage 3: General Video Instruction Tuning. The primary goal of this stage is to lift the model's capabilities from static images to dynamic video, establishing a robust foundation for general video understanding. To achieve this, we finetune the model on CamS-3M, a curated 3M-sample dataset mixture of existing public video instruction-tuning datasets. During this stage, we keep the vision encoder frozen while training the LLM decoder and the vision-language adapter, allowing the model to learn temporal relationships and general video-language concepts.

Stage 4: Spatial Sensing Tuning. This final stage hones the model's specialized spatial intelligence. We finetune the model from Stage 3 on a mixed corpus combining our specialized VSI-590K with a proportional sample of the general video data from CamS-3M. As demonstrated in our ablation study (Sec. G), this data mixture is crucial. It allows us to maximize spatial performance on challenging benchmarks like VSI-Bench while preserving the broad video understanding capabilities developed in Stage 3.

The specific training configurations for these final two stages are detailed in Table 11.

Table 11: Video Instruction Post-Training Configuration for **Cambrian-S-7B** (Stage 3 and Stage 4).

	Stage 3 (General Video Instruction Tuning)	**Stage 4** (Spatial Instruction Tuning)
Model		
Vision Encoder	SigLIP2-So400M	
Language Decoder	Qwen2.5-7B-Instruct	
VL-Adapter	2×MLP-GELU	
Data Recipe		
Data Source	CamS-3M General Video Mixture (Tab. 12)	VSI-590K + a proportional sample of CamS-3M
Video Frame Resolution	Pad (384×384)	Pad (384×384)
Frame Sampling Strategy	Uniform	Uniform
# Frames per Video	64	128
# Tokens per Video Frame	64	64
Training Recipe		
Max Sequence Length	8192	16384
Trainable Modules	VL-Adapter and MLLM	
Learning Rate	1×10^{-5}	
Global Batch Size	256	
Warmup Ratio	0.03	

Table 12: Data sources for the **CamS-3M** general video instruction tuning mixture used in Stage 3 & 4.

Source Datasets	
LLaVA-Video (Zhang et al., 2024c)	VideoChatGPT-Plus (Maaz et al., 2024)
ShareGPT4o (Cui et al., 2024)	Ego4D (Grauman et al., 2022)
VideoChat2 (Li et al., 2024d)	HowTo100M (Miech et al., 2019)
MovieChat (Song et al., 2024)	HD-VILA (Xue et al., 2022)
EgoIT (Yang et al., 2025)	HTStep (Afouras et al., 2023)
Perception Test (Patraucean et al., 2023)	TimeIT (Ren et al., 2024)
Vript (Yang et al., 2024c)	HowToInterlink7M (Wang et al., 2024a)
GUI-World (Chen et al., 2024c)	Video-Localized-Narratives (Voigtlaender et al., 2023)

F Additional Benchmark Results and Analysis of Cambrian-1 and Cambrian-*S*

F.1 Image MLLM Benchmarks

Table 13 details the performance of our improved Cambrian-1-7B and Cambrian-S-7B on image-based MLLM benchmarks.

Table 13: Evaluate our upgraded Cambrian-1-7B and Cambrian-S-7B on image-based MLLM benchmarks.

	General					Knowledge					OCR & Chart					Vision-Centric				
Method	Avg	MME^P	MMB	$SEED^I$	GQA	Avg	SQA^I	$MMMU^V$	$MathVista^M$	AI2D	Avg	ChartQA	OCRBench	TextVQA	DocVQA	Avg	MMVP	RealworldQA	$CV\text{-}Bench^{2D}$	$CV\text{-}Bench^{3D}$
Open-source Models																				
Mini-Gemini-HD-8B	72.7	1606.0	72.7	73.2	64.5	55.7	75.1	37.3	37.0	73.5	62.9	59.1	47.7	70.2	74.6	51.5	18.7	62.1	62.2	63.0
LLaVA-NeXT-8B	72.5	1603.7	72.1	72.7	65.2	55.6	72.8	41.7	36.3	71.6	63.9	69.5	49.0	64.6	72.6	56.6	38.7	60.1	62.2	65.3
Cambrian-1-8B	73.1	1,547.1	75.9	74.7	64.6	61.3	80.4	42.7	49.0	73.0	71.3	73.3	62.4	71.7	77.8	65.0	51.3	64.2	72.3	72.0
LLaVA-OneVision-7B	-	1,580.0	80.8	75.4	-	-	96.0	48.8	63.2	81.4	-	80.0	-	-	87.5	-	-	66.3	-	-
Cambrian-1-7B (Ours)	75.0	1,611.9	78.9	76.3	64.3	64.1	84.2	48.7	45.5	78.0	80.5	78.9	73.3	79.1	90.6	66.3	53.3	67.7	70.0	74.0
Cambrian-S-7B	74.7	1,578.8	79.7	77.0	63.0	64.9	83.7	48.3	49.6	77.9	77.5	77.2	68.7	75.9	88.1	70.4	60.0	67.5	73.4	80.6

F.2 Video MLLM Benchmarks

In §2, we analyze the behaviors when video-based benchmarks meets kinds of different evaluation setups, using our upgraded Cambrian-1-7B as a prober. Detailed results of Fig. 2 are listed in Tab. 14.

Table 14: Our upgraded Cambrian-1's performance on video benchmarks, under different evaluation setups.

Model	VSI-B	Tomato	HourVideo	$Video^{MME}$	EgoSchema	$Video^{MMMU}$	LVBench	MVBench	Percept. Test
Chance-Level	34.0	22.0	20.0	25.00	20.00	14.00	25.0	27.3	33.3
Cambrian-1-7B									
Blind Test	17.4	7.8	24.3	31.2	31.9	25.0	42.9	19.6	40.7
Single Frame	20.4	15.8	27.7	41.6	44.0	29.0	46.9	46.1	52.1
Multiple (32) Frames	25.8	18.9	31.6	53.7	48.1	31.9	52.5	51.4	55.4
(32) Frame Captions	21.8	16.8	29.5	55.3	52.4	40.1	52.2	47.7	55.6
Cambrian-S-7B	58.7	27.2	37.2	61.3	75.7	36.6	54.7	59.3	68.3

F.3 Contributions from Image-based and Video-based Instruction Tuning

To elucidate the respective contributions of image-based and video-based instruction tuning to a model's final video understanding capabilities, we conducted a series of experiments. These experiments employed varying proportions of image and video data during the finetuning stages, and we observed the resulting performance trends across diverse video benchmarks.

More specifically, for the initial image MLLM training, we randomly sampled 1M, 4M, and 7M image question-answering (QA) pairs from Cambrian-7M to train distinct models. Subsequently, for video-specific finetuning, we randomly sampled 25%, 50%, 75%, and 100% of video QA pairs from LLaVA-Video-178K ($\sim$1.6M data samples in total) to perform video-only finetuning on each of these pretrained image MLLMs. The hyperparameters for image instruction tuning and video finetuning were maintained as detailed in Table 10 and Table 11, respectively. The experimental results, presented in Table 15, yield the following observations:

- *Models trained with more image data do not inherently outperform those trained with less when evaluated on video benchmarks without finetuning.* As indicated in the table, direct evaluation on video benchmarks reveals comparable performance across all three models, which were initially trained on 1M, 4M, and 7M image datasets, respectively.

- *Finetuning on video data can be generally beneficial for models pretrained with larger image datasets, though not universally.* When all models were finetuned on 100% video data, the model initially trained on 7M images outperformed the other two on 5 out of 9 video benchmarks (specifically, HourVideo, VideoMME, EgoSchema, LVBench, and Perception Test).
- *Incorporating video data into the training process consistently benefits performance across all video benchmarks.* We observed that finetuning an image-based Multimodal Large Language Model (MLLM) with video data, even a small portion such as 25%, improved its performance on all evaluated video benchmarks.
- *Increasing the amount of video data used for finetuning does not guarantee consistent performance improvements across all benchmarks.* While video finetuning is generally advantageous, some benchmarks (e.g., VideoMME, VSI-Bench, Tomato) do not show further gains with more video data. For instance, models finetuned with 100% video data exhibited performance on par with those finetuned with only 25% video data on the VideoMME benchmark. Only EgoSchema, MVBench, and Perception Test demonstrated consistent benefits from increased video data, a phenomenon we hypothesize is related to the underlying video distribution of the training videos.

Table 15: Video MLLM performance trained with different proportions of image and video data.

Image data	Video data	VSI-B	Tomato	HourVideo	Video^{MME}	EgoSchema	Video^{MMMU}	LVBench	MVBench	Percept. Test
Chance-Level	-	34.0	22.0	20.0	25.0	20.0	14.0	25.0	27.3	33.3
1M	0%	26.0	20.2	32.5	52.1	46.9	32.0	51.4	50.5	54.2
	25%	32.4	25.4	36.2	60.4	47.0	40.1	53.5	57.0	61.9
	50%	33.3	27.2	36.2	61.7	47.1	40.1	53.2	59.2	64.3
	75%	32.7	28.8	34.4	60.7	48.7	37.7	53.3	59.5	66.3
	100%	34.4	28.4	35.1	61.3	48.9	39.6	53.0	60.1	67.5
4M	0%	26.7	20.5	31.8	53.1	44.8	32.0	52.1	51.5	54.9
	25%	32.3	26.7	37.0	61.3	45.0	38.6	53.1	57.6	61.9
	50%	31.9	27.4	37.2	61.9	45.7	38.1	54.2	59.5	65.2
	75%	33.8	27.9	36.2	61.1	47.3	40.9	53.1	60.1	67.0
	100%	33.8	28.0	35.5	60.5	50.2	40.2	52.2	60.5	67.7
7M	0%	25.8	18.9	31.6	53.7	48.1	31.9	52.5	51.4	55.4
	25%	31.5	24.6	36.7	61.3	48.8	37.7	54.7	58.3	62.3
	50%	31.4	27.6	36.6	61.0	49.0	37.9	53.6	59.7	65.6
	75%	31.8	27.0	35.7	61.8	50.7	38.0	53.0	60.2	67.9
	100%	32.6	27.7	37.3	62.1	52.4	39.4	54.3	60.6	68.8

G CAMBRIAN-*S* TRAINING RECIPE ABLATION STUDY

To determine the optimal training recipe for Cambrian-*S*, we studied two key factors: the importance of the pretrained base video model and the composition of the instruction-tuning dataset. As shown in Tab. 16, we start from four base models with diverse video capabilities:

- **A1**, which is only trained with image-text alignment on Cambrian-1 alignment data. No image instruction tuning is conducted.
- **A2**, our improved Cambrian-1 model, which is finetuned with image instruction tuning on top of A1.
- **A3**, which is initialized from A2 and finetuned on 429K general video instruction tuning samples.
- **A4**, which is initialized from A2 and finetuned on a larger 3M set of general video instruction tuning samples.

We finetune each of these checkpoints using two different setups: (i) finetuning only on VSI-590K, and (ii) finetuning on a mixture of VSI-590K and additional general video instruction tuning data. The results in Tab. 16 lead to two key observations that informed our final recipe:

- A stronger base model (indicated by better performance on general benchmarks like VideoMME and EgoSchema) consistently leads to better spatial understanding after finetuning. This implies that strong spatial sensing is built upon a foundation of capable general video understanding.
- Compared to mixed-data finetuning, VSI-590K-only finetuning yields the highest performance on VSI-Bench. However, this specialization comes at the cost of a significant performance drop on other general video benchmarks.

Table 16: **Spatial Sensing Tuning Recipe Investigation**. We take four base models with various general video capability and study their different trends when conducting spatial sensing tuning with two different data recipe. A1: only the connector is trained during image-language alignment, A2: A1 *w/.* Cambrian-7M instruction tuning data, A3: A2 finetuned on 429K video instruction tuning data, A4: A2 finetuned on 3M video instruction tuning data. From A1 to A4, video understanding ability improves monotonically. I-IT and V-IT denotes instruction finetuning on image and video data respectively.

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 + more 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**

Table 17: **VSI-Bench full results.** Best results are **highlited**. It's notable that without any route planning data, Cambrian-*S*-7B can outperform Gemini-1.5-Pro on route planning subtask, which does not only requires spatial sensing but also reasoning.

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

H Predictive Sensing Details

H.1 Next Frame Prediction Head Architecture

As shown in Algorithm 2, our next frame prediction head is a simple two-layer MLP with GELU activation function. The output dimension is set to 1152, which is the same as the dimension of our vision encoder's(*i.e.*, SigLIP2-So400M) output.

Algorithm 2: Next Frame Prediction Head Architecutre (in PyTorch Style).

```
NFPHead(
  Sequential(
    (0): Linear(in_features=3584, out_features=3584, bias=True)
    (1): GELU(approximate=none)
    (2): Linear(in_features=3584, out_features=1152, bias=True)
    )
)
```

H.2 Additional Ablations and Results

In Table 18, we study the impact of NFP loss on model's general video understanding capability and visual spatial intelligence. Specifically, we observe that with a proper tuned loss weight (*i.e.*, 0.1), the NFP loss can have less to none negative impact on the model's video understanding ability.

Table 18: Impact of NFP loss on video understanding.

Loss Weight		Accuracy		
Cosine	MSE	VideoMME	EgoSchema	VSI-Bench
0.0	0.0	63.4	76.3	67.5
0.1	0.1	63.9	76.8	66.1
0.5	0.5	63.6	77.2	60.8
1.0	1.0	60.9	73.1	60.2

H.3 Memory Framework Design for VSO

As introduced in main paper (and shown in Algorithm 3), our predictive memory mechanism comprises three distinct memory levels (M_s, M_l, M_w) and four key transition functions governing their interaction: *Sensory Streaming*, *Memory Compression*, *Memory Consolidation*, and *Retrieval*. This section details the implementation of these functions.

Basic memory units. For our implementation, we utilize the *encoded key-value pairs* from each Large Language Model (LLM) layer as the basic memory units. This choice, rather than using output latent features from a vision encoder or vision-language connector, allows us to fully leverage the LLM's internal capabilities for memory construction without requiring external modules. This design decision will be elaborated upon in subsequent sections.

Streaming sensing. Each incoming frame is initially processed independently by the vision encoder and the vision-language connector with a window size of W_s. Subsequently, it is further encoded by the LLM, referencing selected previous frames. The key-value pairs from these preceding frames, cached in the *Sensory memory buffer* (M_s), provide the necessary context for this encoding step.

Surprise-based memory compression. In the meantime of encoding a single frame, we assess its "surprise" level. This is achieved by calculating the difference between the model's prediction for the current frame and the actual ground truth observation (both in the latent feature space). When a frame of timestamp t is moved from the sensory memory buffer M_s to the long-term memory M_l, if it is deemed non-surprising (*i.e.*, its surprise score is below a predefined threshold T_s), we will downsample its' key-value pairs by a factor of 2 along the spatial ($H \times W$) dimension. This surprise-based compression mitigates redundancy in the information stored within M_l.

Surprise-based memory consolidation. Long-term memory M_l is initialized with a predefined budget size B_{long} (*e.g.*, 32,768 tokens). When the volume of memory tokens surpasses this budget, we apply a *surprise-based* consolidation function to M_l to ensure it remains within the allocated limit. Our consolidation function is straightforward yet effective: we identify the surprise score associated with each frame in M_l. Then, the frame with the lowest surprise score is removed (or "forgotten"). Then, we merge or drop some of these frames according to their surprise scores (we tried three different strategies here: 1. forget the oldest memory, 2. forget the least surprise memory, and 3. forget the least surprise memory while merging adjacent surprise memories if any adjacent surprise memories exist). This process is iterated until the total size of M_l falls below the budget.

Retrieval. Upon receiving a user query q, we first retrieve the most relevant frames from the long-term memory (M_l) to construct the working memory (M_w). This M_w then serves as the context for answering the user's query. To perform this retrieval efficiently without resorting to external modules, we utilize the inherent similarity measurement capabilities of the LLM's attention mechanism. Specifically, for each transformer layer, the user query q is transformed into the attention mechanism's query feature space. We then compute the similarity between this query feature and the key features of each frame stored in M_l. Similarity is measured using cosine distance, and for simplicity, multi-head features are treated as a single feature. The k frames with the highest similarity scores have their key-value pairs selected and utilized by the attention mechanism to further encode the user query.

Algorithm 3: Memory Framework Demonstration

Input: Frames $\{f_1, \ldots, f_T\}$, User query q
Input: Encoder $\mathcal{E}$, Decoder $\mathcal{D}$, Surprise Estimator $\mathcal{S}$, Surprise threshold τ
Input: Compression function $\mathcal{C}$, Consolidation function $\mathcal{G}$, Retrieval function $\mathcal{R}$
Input: Sensory memory $\mathcal{M}_s \leftarrow \emptyset$ with budget B_s, Long-term memory $\mathcal{M}_l \leftarrow \emptyset$ with budget B_l, Working memory $\mathcal{M}_w \leftarrow \emptyset$

for $t \leftarrow 1$ **to** T **do**
  $z_t \leftarrow \mathcal{E}(f_t, \mathcal{M}_s)$;
  $\mathcal{M}_s \leftarrow \mathcal{M}_s \cup \{z_t\}$; // Streaming sensing
  $s_t \leftarrow \mathcal{S}(f_t, \mathcal{M}_s)$; // Surprise estimation
  while $|\mathcal{M}_s| > B_s$ **do**
    Dequeue z_{old} from $\mathcal{M}_s$;
    $m \leftarrow \mathbf{1}[s_t \geq \tau] \cdot z_{\text{old}} + \mathbf{1}[s_t < \tau] \cdot \mathcal{C}(z_{\text{old}})$; // Selective compression
    $\mathcal{M}_l \leftarrow \mathcal{M}_l \cup \{m\}$;
    if $|\mathcal{M}_l| > B_l$ **then**
      $\mathcal{M}_l \leftarrow \mathcal{G}(\mathcal{M}_l)$; // Memory consolidation
$\mathcal{M}_w \leftarrow \mathcal{R}(q, \mathcal{M}_l)$; // Retrieve working memory
$\hat{a} \leftarrow \mathcal{D}(q, \mathcal{M}_w)$; // Answering query with $\mathcal{M}_w$
return $\hat{a}$

H.4 Memory Framework Design for VSC

Algorithm 4 presents our agentic framework for the VSI-SUPER Count task. Similar to the memory design in Algorithm 3, we encode sensory frames using a sliding window approach with a window size of W_s. The latent frame prediction module continuously estimates the expected next frame and computes the prediction error to quantify how "surprise" the actual next frame is. As new frame arrivs, the oldest frames that exceed the sensory memory window are dequeued and stored in the long-term memory. If a dequeued frame is deemed "surprising" (*i.e.*, its prediction error exceeds a predefined threshold τ), which may indicate a scene or spatial boundary, we trigger a query response using the accumulated long-term memory and reset it afterward. The generated response is then stored in the answer memory bank. The final answer is computed as the aggregation of all intermediate answers stored in this bank.

Algorithm 4: Agentic framework design for VSI-SUPER Count task.

Input: Frames $\{f_1, \ldots, f_T\}$, user query q
Input: Encoder $\mathcal{E}$, Decoder $\mathcal{D}$, Surprise Estimator $\mathcal{S}$, threshold τ
Input: Sensory memory $\mathcal{M}_s \leftarrow \emptyset$ with budget B_s
Input: Long-term memory $\mathcal{M}_l \leftarrow \emptyset$, Answer memory bank $\mathcal{M}_{\text{Ans}} \leftarrow \emptyset$
for $t \leftarrow 1$ **to** T **do**
  $z_t \leftarrow \mathcal{E}(f_t, \mathcal{M}_s)$;
  $\mathcal{M}_s \leftarrow \mathcal{M}_s \cup \{z_t\}$; // Streaming sensing
  $s_t \leftarrow \mathcal{S}(f_t, \mathcal{M}_s)$; // Surprise estimation
  if $|\mathcal{M}_s| > B_s$ **then**
    Remove oldest z_{old} from $\mathcal{M}_s$;
    $\mathcal{M}_l \leftarrow \mathcal{M}_l \cup \{z_{\text{old}}\}$; // Store to long-term memory
  if $s_t \geq \tau$ **then**
    $\hat{a} \leftarrow \mathcal{D}(q, \mathcal{M}_l)$; // Answer query using long-term memory
    $\mathcal{M}_{\text{Ans}} \leftarrow \mathcal{M}_{\text{Ans}} \cup \{\hat{a}\}$;
    $\mathcal{M}_l \leftarrow \emptyset$; // Reset long-term memory
return Sum($\mathcal{M}_{\text{Ans}}$)

H.5 COMPARISONS WITH EXISTING MEMORY METHOD

In Tab. 19, we compare our memory design with MovieChat and Flash-VStream, which are both designed for general long-video understanding. Our memory yield consistently better results than MovieChat and Flash-VStream.

We also compare our memory design with MovieChat and Flash-VStream on existing video benchmarks (*i.e.*, VSI-Bench, Video-MME, EgoSchema). As shown in Tab. 20, our memory design outperforms both MovieChat and Flash-VStream across all benchmarks.

Table 19: We compare our memory design with MovieChat and Flash-VStream on VSO benchmark. Notebaly, our memory design outperforms both MovieChat and Flash-VStream across all setups, frequent by a large margin.

Model	10 Mins.	30 Mins.	60 Mins.	120 Mins.	240 Mins.
Our Memory	43.3	45.0	45.0	43.3	43.3
MovieChat	31.7	28.3	28.3	25.0	21.7
Flash-VStream	20.0	31.7	23.3	21.7	20.3

Table 20: **Memory compairson on video benchmarks.** For all methods, video inputs are sampled at 1 FPS. OOM indicates out-of-memory.

Method	VSI-Bench	Video-MME	EgoSchema
Naive Inference	65.3	OOM	76.8
With Memory and Predictive sensing	64.7	61.3	75.8
MovieChat	53.3	59.4	74.7
Flash-VStream	52.1	55.4	73.3

H.6 REPEAT SEQUENCE EXPERIMENT ON "PREDICTIVE ERROR AS SURPRISE"

Following the suggestion of Reviewer x9om, we conducted a repeated sequence experiment to study the difference between "predictive error as surprise" and "adjacent frame feature similarity as surprise". Specifically, we sampled the first two frames from the 288 videos used in VSI-Bench and repeated them 10 times to form 20-frame sequences (pattern: "ABABAB..."). We then fed these sequences into the model to measure surprise scores using both metrics.

We visualize these scores in Fig. 17. As shown, the "adjacent frame feature similarity" scores remain constant because the sequence consists of only two alternating frames. In contrast, a distinct pattern emerges for "predictive error as surprise": the scores initially decrease before gradually increasing.

Here, the surprise is majorly affected by two factors: prior observation, which helps reduce surprise, and the out-of-distribution (OOD) input (the model has never seen these repeated two frames during training), which results in surprise increase. Initially, the prior observation does help to decrease surprise. However, as the sequence gets longer, the increasingly severe OOD input lead to a larger overall increase in surprise.

We also visualize the temporal distribution of when minimum surprise score occur in each sequence. As shown in Fig. 18 (left), the minimum surprise score frequently occurs after the second or third repetition, supporting our claim that past observations enable the model to reduce surprise in subsequent frames. Since interleaved repetition causes significant video jitter, we conducted a similar experiment repeating the first two frames 10 times using an "ABBAABBA..." pattern. As shown in Fig. 18 (right), given this smoother video sequence, the minimum surprise score occurs later in the sequence.

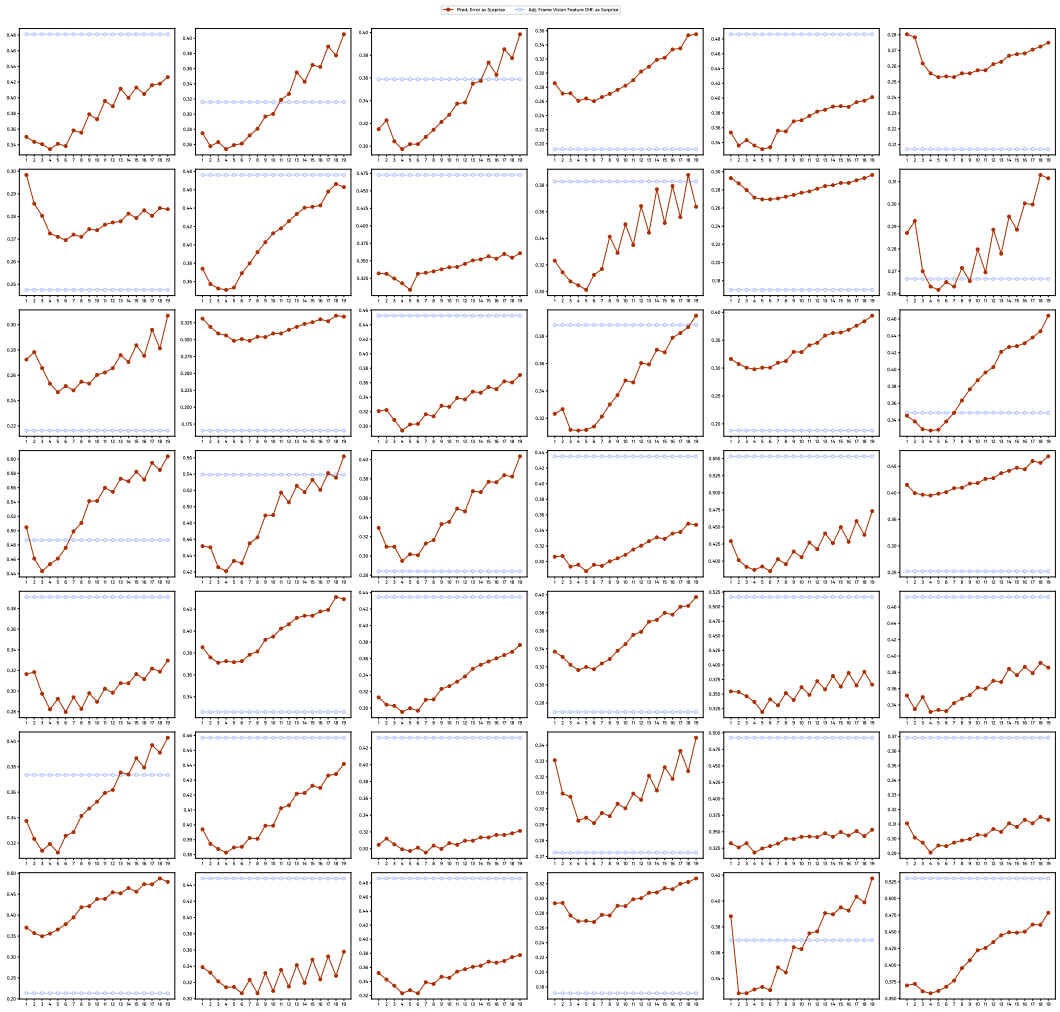


Figure 17: Visualization of surprise scores for both "predictive error as surprise" and "adjacent frame feature similarity as surprise" on 2-frame repeated sequences. Only 19 surprise scores are included because the first frame (with index 0) has no previous frame to compute the surprise score.

I Discussion, Limitation, and Future Work

Summarization. We highlight the importance of and propose a hierarchy for spatial *supersensing* capabilities in videos, arguing that achieving superintelligence requires AI systems to move beyond text-based knowledge and semantic perception, the current focus of most MLLMs, to also develop

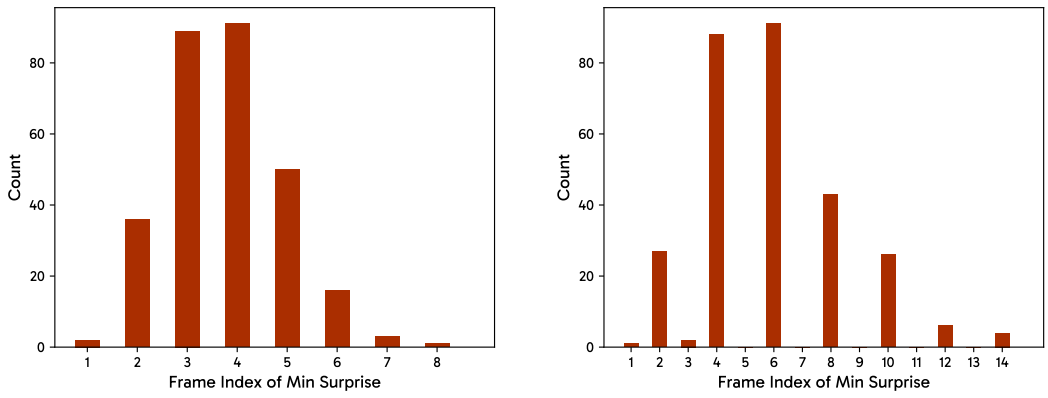


Figure 18: Visualization of the distribution of when the minimum surprise score occurs in each sequence. Left: Video sequence with "ABABAB..." pattern. Right: Video sequence with "ABBAABBA..." pattern.

spatial cognition and predictive world models. To measure progress, we introduce VSI-SUPER and find that current MLLMs struggle with it. To test whether current progress is limited by data, we curate VSI-590K and train our spatially grounded MLLM, Cambrian-*S*, on it. Although Cambrian-*S* performs well on standard benchmarks, its results on VSI-SUPER reveal the limitations of the current MLLM paradigm. We prototype predictive sensing, using latent frame prediction and surprise estimation to handle unbounded visual streams. It improves Cambrian-*S* performance on VSI-SUPER and marks an early step toward spatial supersensing.

Limitations and Future Work. Our goal is to present a conceptual framework that encourages the community to reconsider the importance of developing spatial supersensing. As a long-term research direction, our current benchmark, dataset, and model design remain limited in quality, scale, and generalizability. While a meaningful progress, our current progress is limited by the following factors:

- VSI-SUPER is built with concatenated videos, and covers a limited scope of spatial supersensing.
- Our current "predictive sensing" remains far from the way human do it. For example, humans do not only predict the next frame and measure the surprise, but also learn from the observations and the surprise quickly to update their internal world model. However, our current models are far from achieving this.
- As a compromise to training efficiency and resource constrain, our current model is trained on 1 FPS video, which is far from the real-world video sampling rate and will result in unnegligible information loss.
- Inspite of our models leading performance on VSI-Bench, they remain far away from the human level visual spatial intelligence.

- VSI-SUPER is constructed using concatenated videos and currently covers a limited scope of spatial supersensing scenarios.
- Our implementation of "predictive sensing" differs significantly from human cognition. Humans do not merely predict the next frame and measure surprise; they also rapidly update their internal world models based on these observations. Our current models lack this dynamic adaptation capability.
- To balance training efficiency with resource constraints, our model is trained on 1 FPS video. This sampling rate is much lower than real-world visual streams, resulting in non-negligible information loss.
- Despite our models achieving leading performance on VSI-Bench, they still lag significantly behind human-level visual spatial intelligence.

To address these limitations, we must design more realistic benchmarks and curate larger-scale datasets with diverse scenarios. Furthermore, we emphasize the need for advanced algorithms that extend beyond static training paradigms to enable test-time learning and rapid adaptation.

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

Which of the following correctly represents the order in which the Golden Retriever appeared in the video?
A. Bed, Table, Chest of drawers, Floor
B. Table, Chest of drawers, Bed, Floor
C. Chest of drawers, Floor, Table, Bed
D. Floor, Bed, Chest of drawers, Table

Which of the following correctly represents the order in which white Ragdoll cat appeared in the video?
A. Ground, Trash bin, Bench, Table
B. Table, Bench, Ground, Trash bin
C. Ground, Trash bin, Table, Bench
D. Trash bin, Bench, Table, Ground

Figure 19: Examples of our Sequential Order Recall benchmark. Only edited frames are visualized. Ground truth answers are **highlighted**.

Absolute Direction (Object)

Standing by the backpack, looking toward the table, how far counterclockwise in degrees must I turn to see the trash bin?
Answer: 334.09

Absolute Distance

Measuring from the closest points of each, how far apart are the chair and the door in meters?
Answer: 2.32

Absolute Distance

Considering the chair and the door, which object's longest edge is the shorter?
A. Door
B. Chair

Figure 20: Examples of VSI-590K (Annotated Real Video).

Object Appearance Order

Determine the initial appearance order of these categories in the video: door, chair, lamp, refrigerator.
A. refrigerator, door, lamp, chair
B. refrigerator, chair, door, lamp
C. refrigerator, chair, lamp, door
D. door, chair, lamp, refrigerator

Absolute Size

Provide the longest side's length for the door in inches.
Answer: 72.00

Room Size

Indicate the room's dimensions in square feet. If there's more than one room, estimate their total size.
Answer: 232.76

Figure 21: Examples of VSI-590K (Annotated Real Video).

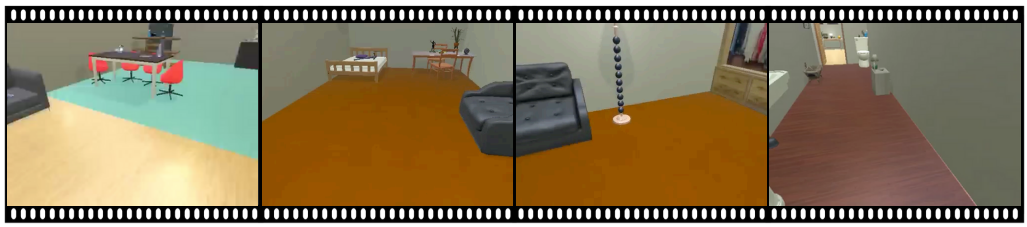
Relative Direction (Object Perspective)

Facing the door while standing near the window, in which of the following positions is the board relative to me: front-left, front-right, back-left, or back-right? Use Cartesian quadrants, with me at the origin looking toward positive y-axis
A. Back-right
B. Front-right
C. Front-left
D. Back-left

Relative Distance (Object Perspective)

Identify the object among (bookcase, chair, board, door) that is closest to the window based on the shortest distance between their closest points. Choose the nearest instance if several exist.
A. Bookcase
B. Chair
C. Board
D. Door

Figure 22: Examples of VSI-590K (Annotated Real Video).



Relative Distance (Object Perspective)

If I am standing by the dresser and facing the chair, is the closet to my left, right, or back? An object is to my back if I would have to turn at least 135 degrees in order to face it.
A. Left
B. Right
C. Back

Figure 23: Examples of VSI-590K (Annotated Simulated Video).

Relative Direction (Object Perspective)

With the toilet beside me and facing the cabinet, is the lamp positioned front-left, front-right, back-left, or back-right relative to me, based on Cartesian plane quadrants?
A. Back-right
B. Front-right
C. Front-left
D. Back-left

Relative Distance (Object Perspective)

Identify the object among (bookcase, chair, board, door) that is closest to the window based on the shortest distance between their closest points. Choose the nearest instance if several exist.
A. Bookcase
B. Chair
C. Board
D. Door

Figure 24: Examples of VSI-590K (Annotated Simulated Video (Frame)).

Object Counting (Relative)

If counted, would chairs be fewer than, more than, or equal in number to tables?
A. Fewer
B. More
C. Equal

Relative Direction (Camera Perspective)

Through the camera's lens, is the sink captured on the left or right part of the scene?
A. Right
B. Left

Figure 25: Examples of VSI-590K (Unannotated Real Video (Frame)).

Object Counting (Absolute)

What would be the count if you tallied all the chairs?
Answer: 6

Relative Distance (Camera Perspective)

In terms of proximity to the camera, which is closer: a table or a sofa?
A. Table
B. Sofa

Figure 26: Examples of VSI-590K (Unannotated Real Video (Frame)).

We manually reviewed all LLM-suggested edits to ensure factual correctness. The LLM is acknowledged here for editorial assistance only and was not involved as an author.

Table 21: Absolute Count Question Template

Absolute Count Question Template

1. What's the number of `{object_name}`(s) present in this room?
2. Can you count how many `{object_name}`(s) are in this room?
3. Could you tell me the total number of `{object_name}`(s) in this room?
4. Exactly how many `{object_name}`(s) are in here?
5. In this room, how many `{object_name}`(s) can be found?
6. What's the count of `{object_name}`(s) in this room?
7. Tell me how many `{object_name}`(s) are located here.
8. Do you know how many `{object_name}`(s) are inside this room?
9. What's the exact quantity of `{object_name}`(s) in this room?
10. Could you specify how many `{object_name}`(s) exist in this room?
11. I'd like to know the number of `{object_name}`(s) in this room.
12. Can you inform me how many `{object_name}`(s) there are here?
13. What's the total number of `{object_name}`(s) found in this room?
14. How many `{object_name}`(s) can we find in this area?
15. Please provide the count of `{object_name}`(s) in this room.
16. What's the exact count of `{object_name}`(s) present here?
17. Could you clarify how many `{object_name}`(s) there are in this room?
18. How many `{object_name}`(s) do we have in this room?
19. What's the quantity of `{object_name}`(s) seen in this room?
20. Could you indicate how many `{object_name}`(s) are present?
21. Can you verify the number of `{object_name}`(s) in this room?
22. I'd appreciate knowing how many `{object_name}`(s) are here.
23. Precisely how many `{object_name}`(s) does this room contain?
24. How many `{object_name}`(s) does this room have?
25. Could you give me the number of `{object_name}`(s) in this space?

Table 22: Absolute Direction Question Template (Object Perspective)

Absolute Direction Question Template (Object Perspective)

1. I'm at the {object_1}, looking towards the {object_2}. How many degrees {direction} should I rotate to look at the {object_3}?
2. Standing by the {object_1} and facing toward the {object_2}, how far in degrees do I turn {direction} to face the {object_3}?
3. From the {object_1}, oriented toward the {object_2}, what's the angle of rotation {direction} needed to look directly at the {object_3}?
4. If I'm positioned at the {object_1}, looking directly at the {object_2}, how many degrees must I rotate {direction} to align myself with the {object_3}?
5. Standing at the {object_1} and directed toward the {object_2}, what's the degree measurement I need to rotate {direction} to face the {object_3}?
6. At the {object_1}, when facing the {object_2}, how many degrees should I turn {direction} to face toward the {object_3}?
7. Standing at the {object_1}, facing toward the {object_2}, how far {direction} do I rotate (in degrees) to see the {object_3}?
8. Starting from the {object_1} and looking at the {object_2}, how many degrees of {direction} rotation are required to look at the {object_3}?
9. At the location of the {object_1}, facing the {object_2}, what angle (in degrees) {direction} do I rotate to directly see the {object_3}?
10. Standing at the {object_1}, facing the {object_2}, how many degrees do I have to rotate in {direction} to face the {object_3} exactly?
11. Positioned at the {object_1} and oriented toward the {object_2}, how many degrees {direction} should I turn to face the {object_3}?
12. If at the {object_1} and directly facing the {object_2}, what {direction} angle adjustment is needed to look at the {object_3}?
13. From my position at the {object_1}, looking toward the {object_2}, how many degrees should I rotate {direction} to view the {object_3}?
14. Standing by the {object_1}, looking toward the {object_2}, how far {direction} in degrees must I turn to see the {object_3}?
15. I'm at the {object_1} facing the {object_2}; how many degrees {direction} must I rotate to align my view with the {object_3}?
16. Located at the {object_1} and facing toward the {object_2}, how many degrees {direction} rotation will it take to directly face the {object_3}?
17. Standing at the {object_1} with eyes toward the {object_2}, how many degrees must I rotate {direction} to point toward the {object_3}?
18. When positioned at the {object_1} and viewing the {object_2}, what's the required angle of {direction} rotation to face the {object_3}?
19. From my standpoint at the {object_1}, facing toward the {object_2}, how many degrees {direction} must I rotate to directly face the {object_3}?
20. If I'm standing at the {object_1}, looking at the {object_2}, what's the {direction} degree measurement needed to see the {object_3}?
21. At the {object_1}, facing the {object_2}, what's the precise angle of {direction} rotation required to turn toward the {object_3}?
22. Standing at the {object_1} and oriented to the {object_2}, how much should I rotate {direction}, in degrees, to face the {object_3}?
23. If I'm located at the {object_1}, viewing the {object_2}, how many degrees of rotation {direction} are necessary to face the {object_3}?
24. Standing at the {object_1}, looking toward the {object_2}, what's the number of degrees needed to rotate {direction} to look directly at the {object_3}?
25. From the {object_1} and facing the {object_2}, how many degrees {direction} do I precisely rotate to face the {object_3}?

Table 23: Absolute Distance Question Template (Object Perspective)

Absolute Distance Question Template (Object Perspective)

1. Measuring from the closest points of each, how far apart are the {object_1} and the {object_2} in {unit}?
2. Using the nearest points, what's the spacing between the {object_1} and the {object_2} in {unit}?
3. Considering their closest points, what's the separation of the {object_1} from the {object_2} in {unit}?
4. From the closest points of each object, what's the gap between the {object_1} and the {object_2} expressed in {unit}?
5. Measured at their closest points, what's the length separating the {object_1} and the {object_2} in {unit}?
6. Using their nearest points, measure the distance from the {object_1} to the {object_2} in {unit}.
7. Counting from the closest points, how many {unit} lie between the {object_1} and the {object_2}?
8. What's the measure of space between the nearest points of the {object_1} and the {object_2} expressed in {unit}?
9. State the distance from the closest point on the {object_1} to the nearest point on the {object_2} in {unit}.
10. Determine the shortest distance between the {object_1} and the {object_2} in {unit}.
11. Provide the separation distance between the closest points of the {object_1} and the {object_2} in {unit}.
12. In {unit}, what's the smallest distance between the {object_1} and the {object_2} using their closest points?
13. Measured from their nearest points, what's the extent between the {object_1} and the {object_2} in {unit}?
14. Express in {unit} how far apart the closest points of the {object_1} and the {object_2} are.
15. Quantify the shortest space separating the {object_1} from the {object_2} in {unit}.
16. How many {unit} separate the closest points of the {object_1} and the {object_2}?
17. What's the linear distance between the nearest points of the {object_1} and the {object_2} in {unit}?
18. Calculate the minimal span from the {object_1} to the {object_2} using their closest points, in {unit}.
19. Report the shortest distance measurement between the {object_1} and the {object_2} in {unit}.
20. What's the measurable gap from the nearest point on the {object_1} to the closest point on the {object_2} in {unit}?
21. Specify precisely how far apart the {object_1} and the {object_2} are at their closest points, expressed in {unit}.
22. Can you provide the distance separating the nearest points of the {object_1} and the {object_2}, measured in {unit}?
23. Find the shortest distance between the {object_1} and the {object_2} using their closest points, in {unit}.
24. Measured at their nearest points, state how far apart the {object_1} is from the {object_2} in {unit}.
25. What is the precise shortest measurement between the {object_1} and the {object_2} in {unit}?

Table 24: Absolute Size Question Template

Absolute Size Question Template

1. How long is the longest side of the `{object_name}` measured in `{unit}`?
2. What's the measurement of the `{object_name}`'s longest side in `{unit}`?
3. Can you provide the length of the longest edge of the `{object_name}` in `{unit}`?
4. In `{unit}`, what's the longest dimension of the `{object_name}`?
5. What's the length of the longest side of the `{object_name}`, expressed in `{unit}`?
6. Tell me the measurement in `{unit}` of the `{object_name}`'s longest side.
7. What's the `{unit}` length of the longest edge of the `{object_name}`?
8. Provide the longest side's length for the `{object_name}` in `{unit}`.
9. How many `{unit}` is the longest side of the `{object_name}`?
10. Could you specify the longest edge length of the `{object_name}` using `{unit}`?
11. What's the maximum length of the `{object_name}` measured in `{unit}`?
12. Give the longest dimension of the `{object_name}` in `{unit}`.
13. What's the length in `{unit}` of the `{object_name}`'s longest side?
14. Expressed in `{unit}`, how long is the longest side of the `{object_name}`?
15. What's the size of the longest side of the `{object_name}` in terms of `{unit}`?
16. How lengthy is the longest side of the `{object_name}` in `{unit}`?
17. What's the measure of the `{object_name}`'s largest side in `{unit}`?
18. Report the longest side's length of the `{object_name}` in `{unit}`.
19. What's the longest side measurement of the `{object_name}`, using `{unit}`?
20. State the length of the `{object_name}`'s longest side in `{unit}`.
21. How many `{unit}` long is the longest edge of the `{object_name}`?
22. What's the longest side dimension of the `{object_name}`, stated in `{unit}`?
23. Identify the length of the longest side of the `{object_name}` in `{unit}`.
24. What's the longest side of the `{object_name}` measured as in `{unit}`?
25. Give the longest side of the `{object_name}` in `{unit}`.

Table 25: Relative Count Question Template

Relative Count Question Template

1. Does this room have more or fewer {category_1}(s) compared to {category_2}(s)?
2. In this room, are there more or fewer {category_1}(s) relative to {category_2}(s)?
3. Is the number of {category_1}(s) greater or smaller than the number of {category_2}(s) in this room?
4. Are there more {category_1}(s) or fewer {category_1}(s) than {category_2}(s) in this room?
5. In this room, do {category_1}(s) outnumber {category_2}(s), or are there fewer?
6. Do you find more or fewer {category_1}(s) compared with {category_2}(s) here?
7. Are there more {category_1}(s) or fewer of them compared to {category_2}(s) in the room?
8. Is the count of {category_1}(s) higher or lower than the count of {category_2}(s) in this room?
9. In terms of quantity, are there more or fewer {category_1}(s) than {category_2}(s) present here?
10. Does this room contain more or fewer {category_1}(s) than it does {category_2}(s)?
11. Are {category_1}(s) more numerous or less numerous than {category_2}(s) in this room?
12. Are there more or fewer {category_1}(s) than there are {category_2}(s) inside this room?
13. Within this room, is the quantity of {category_1}(s) greater or smaller compared to {category_2}(s)?
14. Do we have a greater or lesser number of {category_1}(s) than {category_2}(s) in this room?
15. Compared to {category_2}(s), are there more or fewer {category_1}(s) in this room?
16. In this room, are {category_1}(s) more or fewer plentiful than {category_2}(s)?
17. Are {category_1}(s) found in greater or smaller numbers than {category_2}(s) here?
18. Are there more {category_1}(s) present, or are there fewer, as compared to {category_2}(s) in this space?
19. Is there a higher or lower count of {category_1}(s) than {category_2}(s) in this room?
20. In comparison with {category_2}(s), are there more or fewer {category_1}(s) in this room?
21. Does this room have more or fewer quantities of {category_1}(s) than {category_2}(s)?
22. Are there more or fewer {category_1}(s) here than there are {category_2}(s)?
23. Can you tell if the number of {category_1}(s) is higher or lower than that of {category_2}(s) in this room?
24. Within this room, do we have more or fewer {category_1}(s) compared to {category_2}(s)?
25. Is the amount of {category_1}(s) in this room greater or lesser compared with {category_2}(s)?

Table 26: Relative Direction (Hard) Question Template (Object Perspective)

Relative Direction (Hard) Question Template (Object Perspective)

1. I'm at the {object_1}, looking toward the {object_2}. Is the {object_3} located at my front-left, front-right, back-left, or back-right?
2. From the position of the {object_1} facing the {object_2}, where is the {object_3} relative to me: front-left, front-right, back-left, or back-right?
3. Standing near the {object_1} and looking at the {object_2}, is the {object_3} positioned at my front-left, front-right, back-left, or back-right?
4. At the spot of the {object_1}, facing toward the {object_2}, is the {object_3} in my front-left, front-right, back-left, or back-right?
5. If I'm positioned at the {object_1} and facing the {object_2}, would the {object_3} be in my front-left, front-right, back-left, or back-right?
6. With the {object_1} as my location and looking at the {object_2}, in which direction is the {object_3}: front-left, front-right, back-left, or back-right?
7. From the viewpoint at the {object_1} looking toward the {object_2}, is the {object_3} at my front-left, front-right, back-left, or back-right?
8. When standing at the {object_1} and oriented toward the {object_2}, where does the {object_3} appear: front-left, front-right, back-left, or back-right?
9. At the location of the {object_1}, while facing the {object_2}, is the {object_3} situated front-left, front-right, back-left, or back-right of me?
10. Standing at the {object_1}, facing directly toward the {object_2}, would the {object_3} be located at my front-left, front-right, back-left, or back-right?
11. From the place of the {object_1} looking at the {object_2}, can you confirm if the {object_3} is toward my front-left, front-right, back-left, or back-right?
12. If I'm at the {object_1}, oriented toward the {object_2}, which quadrant is the {object_3} in: front-left, front-right, back-left, or back-right?
13. At the point of the {object_1}, facing the {object_2}, identify if the {object_3} is at my front-left, front-right, back-left, or back-right.
14. I'm located at the {object_1}, facing the {object_2}; is the {object_3} in my front-left, front-right, back-left, or back-right direction?
15. When positioned at the {object_1} and looking toward the {object_2}, in which direction would I find the {object_3}: front-left, front-right, back-left, or back-right?
16. At the {object_1}, looking straight at the {object_2}, is the {object_3} situated front-left, front-right, back-left, or back-right of me?
17. If standing near the {object_1} and facing the {object_2}, would the {object_3} be front-left, front-right, back-left, or back-right relative to my view?
18. With my position at the {object_1} looking toward the {object_2}, determine if the {object_3} is at my front-left, front-right, back-left, or back-right.
19. Standing by the {object_1}, directed toward the {object_2}, does the {object_3} lie front-left, front-right, back-left, or back-right from my viewpoint?
20. If I'm standing at the {object_1} facing the {object_2}, can you tell if the {object_3} is in my front-left, front-right, back-left, or back-right?
21. At the {object_1}, with my gaze fixed on the {object_2}, is the {object_3} positioned front-left, front-right, back-left, or back-right relative to me?
22. Standing at the {object_1}, oriented toward the {object_2}, would the {object_3} appear at my front-left, front-right, back-left, or back-right?
23. Positioned at the {object_1}, looking directly toward the {object_2}, where exactly is the {object_3}: front-left, front-right, back-left, or back-right?
24. From the standpoint of the {object_1} and facing the {object_2}, is the {object_3} found front-left, front-right, back-left, or back-right?
25. If located at the {object_1} and looking toward the {object_2}, in what direction is the {object_3}: front-left, front-right, back-left, or back-right?

Table 27: Relative Direction (Medium) Question Template (Object Perspective)

Relative Direction (Medium) Question Template (Object Perspective)

1. If I'm next to the {object_1}, looking towards the {object_2}, is the {object_3} on my left, right, or behind me? 'Behind' means turning at least 135 degrees to face it.
2. Standing near the {object_1} and facing the {object_2}, would the {object_3} be positioned to my left, right, or rear? 'Rear' implies needing at least a 135-degree rotation to face it.
3. If I stand by the {object_1}, oriented toward the {object_2}, where is the {object_3}: left, right, or behind? An object behind requires turning 135 degrees or more to face it directly.
4. Facing the {object_2} while positioned by the {object_1}, is the {object_3} located to my left, right, or back? 'Back' means I'd have to turn at least 135 degrees to face it.
5. From my position near the {object_1}, looking directly at the {object_2}, is the {object_3} on my left side, right side, or behind? 'Behind' involves turning at least 135 degrees to face it.
6. If my location is beside the {object_1} and I'm facing towards the {object_2}, is the {object_3} to my left, right, or rear? 'Rear' means I'd need to rotate at least 135 degrees.
7. When standing next to the {object_1} and gazing at the {object_2}, would the {object_3} be on my left, right, or behind? Behind indicates needing to turn 135 degrees or more to face it.
8. I'm positioned at the {object_1}, looking toward the {object_2}; is the {object_3} placed to my left, right, or behind me? 'Behind' suggests turning at least 135 degrees to see it.
9. Standing alongside the {object_1} and facing the {object_2}, does the {object_3} lie to my left, right, or behind? Behind means a minimum 135-degree turn is needed.
10. If I am near the {object_1}, turned toward the {object_2}, is the {object_3} found on my left, right, or rear? To my rear means rotating at least 135 degrees.
11. Looking at the {object_2} from my spot by the {object_1}, is the {object_3} situated left, right, or behind? Behind means I must rotate at least 135 degrees.
12. If I stand beside the {object_1}, directed towards the {object_2}, would I find the {object_3} to my left, right, or behind? 'Behind' implies turning 135 degrees or more.
13. From my stance at the {object_1}, looking straight towards the {object_2}, is the {object_3} positioned on my left, right, or at my back? 'Back' means a turn of at least 135 degrees.
14. Standing close to the {object_1}, facing the {object_2}, where is the {object_3}: to my left, right, or behind? 'Behind' means turning at least 135 degrees around.
15. If I'm at the {object_1}, looking at the {object_2}, is the {object_3} located to my left, right, or behind me? Behind means rotating 135 degrees or more to see it clearly.
16. I'm near the {object_1}, oriented toward the {object_2}. Is the {object_3} found to my left, right, or behind? 'Behind' implies I need at least a 135-degree turn.
17. Standing at the {object_1} and facing directly towards the {object_2}, is the {object_3} on my left, right, or to my back? 'To my back' means I'd need a 135-degree or greater rotation.
18. When next to the {object_1}, viewing the {object_2}, is the {object_3} situated on my left, right, or behind? Behind involves turning at least 135 degrees.
19. Facing the {object_2} from the {object_1}, would the {object_3} be placed left, right, or behind me? 'Behind' means I'd have to turn at least 135 degrees.
20. If I position myself at the {object_1}, aiming toward the {object_2}, is the {object_3} to my left, right, or behind me? Behind signifies at least a 135-degree rotation is required.
21. Standing near the {object_1}, directing my view towards the {object_2}, is the {object_3} located left, right, or behind? Behind means I would turn 135 degrees or more to face it.
22. From my place by the {object_1}, facing the {object_2}, is the {object_3} on my left side, right side, or behind? 'Behind' indicates needing to rotate at least 135 degrees.
23. Standing by the {object_1}, if I'm looking towards the {object_2}, is the {object_3} situated to my left, right, or behind? Behind requires turning at least 135 degrees to face it.
24. If I'm positioned near the {object_1} and looking at the {object_2}, would the {object_3} be found to my left, right, or back? 'Back' means rotating at least 135 degrees to face it.
25. When next to the {object_1} and directed towards the {object_2}, does the {object_3} lie to my left, right, or behind me? 'Behind' means turning at least 135 degrees around to face it.

Table 28: Relative Direction (Easy) Question Template (Object Perspective)

Relative Direction (Easy) Question Template (Object Perspective)

1. Standing next to the {object_1} and looking toward the {object_2}, is the {object_3} on the left or right of the {object_2}?
2. If I'm positioned at the {object_1} and oriented toward the {object_2}, would the {object_3} be to the left or right of it?
3. When I'm beside the {object_1} facing toward the {object_2}, is the {object_3} located to its left or right?
4. While standing at the {object_1} and facing the {object_2}, which side of the {object_2} is the {object_3} on—left or right?
5. If I'm at the {object_1}, looking directly at the {object_2}, does the {object_3} sit on its left or right side?
6. I'm standing near the {object_1} and facing toward the {object_2}; is the {object_3} situated to the left or right of the {object_2}?
7. With the {object_1} next to me and the {object_2} ahead, is the {object_3} positioned on the left or right of the {object_2}?
8. If I stand alongside the {object_1} and face the {object_2}, will I find the {object_3} to the left or the right of the {object_2}?
9. Facing the {object_2} from the {object_1}, can you confirm if the {object_3} is on its left side or its right side?
10. When positioned beside the {object_1} and looking at the {object_2}, is the {object_3} placed on the left or right?
11. If I'm standing by the {object_1}, looking toward the {object_2}, would the {object_3} appear on the left or right side of it?
12. Standing next to the {object_1} and looking toward the {object_2}, should I expect the {object_3} to my left or right of the {object_2}?
13. From my position at the {object_1}, facing toward the {object_2}, is the {object_3} to the left or right of the {object_2}?
14. Standing at the {object_1} facing the {object_2}, does the {object_3} lie on its left or right?
15. If I'm located by the {object_1} and oriented toward the {object_2}, would the {object_3} be positioned to its left or right side?
16. Standing beside the {object_1} and facing the {object_2}, which side—left or right—is the {object_3} located on?
17. When at the {object_1}, facing the {object_2}, is the {object_3} found to the left or to the right?
18. If I'm next to the {object_1} looking at the {object_2}, will the {object_3} be seen to its left or right?
19. Positioned by the {object_1} and facing toward the {object_2}, on which side—left or right—is the {object_3}?
20. If I stand near the {object_1}, looking toward the {object_2}, is the {object_3} on the {object_2}'s left or right?
21. Standing adjacent to the {object_1} and viewing the {object_2}, is the {object_3} on the left side or the right side?
22. From the perspective of standing at the {object_1} and facing toward the {object_2}, does the {object_3} lie to its left or right?
23. When I'm at the {object_1}, looking directly toward the {object_2}, is the {object_3} located to the left or right?
24. Standing next to the {object_1}, and facing the {object_2}, do you see the {object_3} positioned to the left or to the right?
25. If I'm standing near the {object_1} looking at the {object_2}, is the {object_3} to the left or to the right of it?

Table 29: Relative Distance Question Template (Object Perspective)

Relative Distance Question Template (Object Perspective)

1. When measuring from the nearest points, which object among ({choice_a}, {choice_b}, {choice_c}, {choice_d}) is nearest to the {category}? In case multiple instances exist, measure to the closest.
2. Considering the nearest point of each object, identify the object from ({choice_a}, {choice_b}, {choice_c}, {choice_d}) that's closest to the {category}. If multiple objects exist, choose the nearest instance.
3. Which one of these items ({choice_a}, {choice_b}, {choice_c}, {choice_d}) lies closest to the {category} when measured from their nearest points? Use the nearest instance if multiple exist.
4. By measuring from the closest points of these objects, which object ({choice_a}, {choice_b}, {choice_c}, {choice_d}) is closest to the {category}? If duplicates occur, measure to the nearest.
5. Using the closest points as reference, which of these ({choice_a}, {choice_b}, {choice_c}, {choice_d}) is closest to the {category}? When multiple instances exist, refer to the nearest one.
6. From the closest point of each object, determine which among ({choice_a}, {choice_b}, {choice_c}, {choice_d}) is nearest the {category}. If multiple exist, use the nearest instance.
7. Considering proximity at their nearest points, which object out of ({choice_a}, {choice_b}, {choice_c}, {choice_d}) is closest to the {category}? In case of multiples, measure the nearest instance.
8. Identify which of these items ({choice_a}, {choice_b}, {choice_c}, {choice_d}) is nearest to the {category}, measured from their closest points. Select the nearest if multiple instances are present.
9. Which object among these options ({choice_a}, {choice_b}, {choice_c}, {choice_d}) is closest to the {category} when measuring from the closest point? Measure the closest instance if multiple exist.
10. By using the closest points, identify which of these ({choice_a}, {choice_b}, {choice_c}, {choice_d}) is nearest to the {category}. If there are multiple objects, select the closest one.
11. Which of these ({choice_a}, {choice_b}, {choice_c}, {choice_d}) is closest to the {category}, measuring from their nearest points? If more than one exists, use the closest instance.
12. From their nearest points, which object ({choice_a}, {choice_b}, {choice_c}, {choice_d}) is closest to the {category}? If multiple instances appear, pick the nearest one.
13. Measure from the closest point: among these options ({choice_a}, {choice_b}, {choice_c}, {choice_d}), which is nearest to the {category}? Use the closest instance if multiples occur.
14. Considering each object's nearest point, which of these objects ({choice_a}, {choice_b}, {choice_c}, {choice_d}) is nearest the {category}? In case of duplicates, measure to the closest.
15. When measuring from their nearest points, which of these objects ({choice_a}, {choice_b}, {choice_c}, {choice_d}) is closest to the {category}? If multiple instances exist, select the closest.
16. Identify the object among ({choice_a}, {choice_b}, {choice_c}, {choice_d}) that is closest to the {category} based on the shortest distance between their closest points. Choose the nearest instance if several exist.
17. Using proximity from their closest points, determine the closest object from ({choice_a}, {choice_b}, {choice_c}, {choice_d}) to the {category}. For multiples, measure the nearest one.
18. From the nearest points, which one of ({choice_a}, {choice_b}, {choice_c}, {choice_d}) is closest to the {category}? When multiple objects are present, measure to the closest.
19. Measuring distance from the nearest points, select the closest object ({choice_a}, {choice_b}, {choice_c}, {choice_d}) to the {category}. If multiple exist, use the nearest instance.
20. Based on measuring from their closest points, which among ({choice_a}, {choice_b}, {choice_c}, {choice_d}) lies nearest to the {category}? If several exist, measure to the closest one.
21. Considering the nearest points, which of these ({choice_a}, {choice_b}, {choice_c}, {choice_d}) is closest to the {category}? If there are multiple, identify the closest instance.
22. Determine which object ({choice_a}, {choice_b}, {choice_c}, {choice_d}) is closest to the {category} from their nearest points. If multiple instances appear, pick the nearest.
23. Measuring the closest points, which item among ({choice_a}, {choice_b}, {choice_c}, {choice_d}) is nearest to the {category}? For multiple occurrences, measure the nearest instance.
24. Identify the closest object from these options ({choice_a}, {choice_b}, {choice_c}, {choice_d}) to the {category}, using the closest point as reference. Use the closest instance if multiple exist.
25. Considering distances from each object's nearest point, which object ({choice_a}, {choice_b}, {choice_c}, {choice_d}) is nearest to the {category}? If multiple instances exist, select the closest one.

Table 30: Relative Size Question Template

Relative Size Question Template

1. Comparing the {object_1} and the {object_2}, which one has the {adjective} longest edge?
2. Which of these two, the {object_1} or the {object_2}, has the {adjective} longest side?
3. Of the {object_1} and the {object_2}, whose longest edge is the {adjective}?
4. Between the {object_1} and the {object_2}, whose edge is the {adjective} in length?
5. Among the {object_1} and the {object_2}, which possesses the {adjective} longest edge?
6. Considering the {object_1} and the {object_2}, which object's longest edge is the {adjective}?
7. Which object's longest dimension, the {object_1} or the {object_2}, is the {adjective}?
8. Between the {object_1} and the {object_2}, which features the {adjective} longest edge?
9. When looking at the {object_1} and the {object_2}, which has the {adjective} longest side?
10. From the {object_1} and the {object_2}, whose longest side is the {adjective}?
11. Between the {object_1} and the {object_2}, which contains the {adjective} longest edge?
12. Which one, the {object_1} or the {object_2}, has a longest edge that is the {adjective}?
13. Comparing longest edges of the {object_1} and the {object_2}, which is the {adjective}?
14. Which has the {adjective} longest dimension: the {object_1} or the {object_2}?
15. Of these two objects, the {object_1} and the {object_2}, which edge is the {adjective} longest?
16. Between the {object_1} and the {object_2}, whose longest edge measures the {adjective}?
17. Between the {object_1} and the {object_2}, which object's longest edge is the {adjective}?
18. Is the longest edge of the {object_1} or the {object_2} the {adjective}?
19. Between the {object_1} and the {object_2}, whose longest side is the {adjective}?
20. Among the longest edges of the {object_1} and the {object_2}, which is the {adjective}?
21. Between the {object_1} and the {object_2}, whose longest edge length is the {adjective}?
22. Looking at the {object_1} and the {object_2}, whose longest edge comes out as the {adjective}?
23. Between the {object_1} and the {object_2}, which object's longest edge appears the {adjective}?
24. Which has the {adjective} maximum length: the {object_1}'s longest edge or the {object_2}'s?
25. Comparing the {object_1} and the {object_2}, whose longest edge length is the {adjective}?

Table 31: Room Size Question Template

Room Size Question Template

1. Could you provide the room dimensions in {unit}? If multiple rooms are displayed, please estimate their total size.
2. What's the area of the room measured in {unit}? For several rooms, estimate the combined size.
3. How large is the room in terms of {unit}? If more than one room is shown, estimate their total size.
4. Please indicate the size of the room using {unit}. If there are multiple rooms, estimate the combined area.
5. What's the room size in {unit}? If multiple rooms appear, calculate the combined area.
6. Could you estimate the room size in {unit}? When multiple rooms are present, provide the total area.
7. In {unit}, what's the measurement of the room? If showing several rooms, estimate the combined space.
8. Provide the dimensions of the room in {unit}. For multiple rooms, estimate the total size.
9. How much space does the room cover in {unit}? If multiple rooms, estimate the combined measurement.
10. What is the total size of the room expressed in {unit}? Estimate combined dimensions if multiple rooms are visible.
11. Indicate the room's dimensions in {unit}. If there's more than one room, estimate their total size.
12. What's the area measurement of the room in {unit}? Estimate the total size if multiple rooms are shown.
13. Please state the size of the room in {unit}. Estimate the combined space if several rooms are provided.
14. Can you specify the room size in {unit}? If several rooms are presented, estimate the combined area.
15. In terms of {unit}, what's the room's size? If multiple rooms appear, estimate their total area.
16. Could you clarify the size of the room using {unit}? Estimate the total if multiple rooms are involved.
17. Please give the room size measured in {unit}. When multiple rooms are shown, estimate their combined size.
18. What's the room dimension in {unit}? For multiple rooms, provide an estimate of their total area.
19. State the room size in terms of {unit}. If multiple rooms are shown, estimate their combined dimensions.
20. Can you tell me the room size in {unit}? For several rooms, estimate the overall size.
21. What's the measurement of the room in {unit}? Estimate combined space if multiple rooms are visible.
22. Could you provide the dimensions of the room using {unit}? If multiple rooms appear, estimate their total space.
23. What is the room's size in {unit}? Provide the combined size if more than one room is depicted.
24. How large is the room measured in {unit}? If multiple rooms, estimate the combined area.
25. Please specify the room's dimensions in {unit}. Estimate the total size for multiple rooms shown.

Table 32: Relative Direction Question Template (Camera Perspective)

Relative Direction Question Template (Camera Perspective)

1. From the camera's perspective, is the `{object_1}` positioned on the left or right?
2. Looking through the camera, does the `{object_1}` appear on the left side or the right side?
3. In the camera frame, which side is the `{object_1}` located on – left or right?
4. If viewing through the camera lens, is the `{object_1}` situated to the left or to the right?
5. From the camera's viewpoint, is the `{object_1}` positioned on the left-hand side or right-hand side?
6. As seen by the camera, is the `{object_1}` on the left or on the right portion of the image?
7. When looking at the camera view, does the `{object_1}` fall on the left or right section?
8. From the perspective of someone behind the camera, would the `{object_1}` be on the left or right?
9. Is the `{object_1}` located on the left side or right side from the camera's angle?
10. Relative to the camera's orientation, is the `{object_1}` positioned left or right?
11. In the camera's field of view, does the `{object_1}` appear in the left region or right region?
12. Would you say the `{object_1}` is on the left or right half of the frame as seen by the camera?
13. Based on the camera's view, which lateral position does the `{object_1}` occupy – left or right?
14. Through the camera's lens, is the `{object_1}` captured on the left or right part of the scene?
15. Is the `{object_1}` situated on the left-hand or right-hand side from the camera's standpoint?
16. When viewing the scene through the camera, does the `{object_1}` appear to the left or to the right?
17. As captured by the camera, is the `{object_1}` positioned on the left or right section of the image?
18. Does the camera show the `{object_1}` on its left side or its right side?
19. If the camera is the reference point, is the `{object_1}` located on the left or right portion?
20. From what the camera sees, is the `{object_1}` positioned on the left or right area?

Table 33: Relative Distance Question Template (Camera Perspective)

Relative Distance Question Template (Camera Perspective)

1. Between `{category_1}` and `{category_2}`, which is closer to the camera?
2. Is `{category_1}` or `{category_2}` nearer to the camera's position?
3. Which is positioned closer to the camera: `{category_1}` or `{category_2}`?
4. From the camera's perspective, which is at a shorter distance: `{category_1}` or `{category_2}`?
5. Compare `{category_1}` and `{category_2}`, which is situated closer to the camera?
6. Which category appears closer to the camera's viewpoint: `{category_1}` or `{category_2}`?
7. In terms of proximity to the camera, which is closer: `{category_1}` or `{category_2}`?
8. Which would you say is nearer to the camera lens: `{category_1}` or `{category_2}`?
9. A `{category_1}` or a `{category_2}`, which is closer from the camera?
10. From the camera's standpoint, which has less distance: `{category_1}` or `{category_2}`?
11. Which category is at a reduced distance from the camera: `{category_1}` or `{category_2}`?
12. When measuring from the camera, which would require less distance to reach: `{category_1}` or `{category_2}`?
13. Between `{category_1}` and `{category_2}`, which one is nearer to where the camera is positioned?
14. Which has the shorter spatial distance from the camera: `{category_1}` or `{category_2}`?
15. In relation to the camera's location, which is more proximate: `{category_1}` or `{category_2}`?
16. Does `{category_1}` or `{category_2}` have greater proximity to the camera?
17. As viewed from the camera's position, which is closer: `{category_1}` or `{category_2}`?
18. Which category is in closer proximity to the camera's placement: `{category_1}` or `{category_2}`?
19. Are `{category_2}` or `{category_1}` positioned nearer to the camera?
20. When measuring from the camera, which requires traveling less distance to reach: `{category_1}` or `{category_2}`?