Towards Spatial Supersensing in the Wild

Tianjun Gu[1*], Tianyu Xin[1*], Kuan Zhang[1*], Bowen Yang[1], Kok-Chung Chua[1], Peize Li[4], Xinran Zhang[1], Yupeng Chen[1], Qiyue Zhao[1], Qinlei Xie[1], Jianhang Liu[1], Yucheng Lu[1], Yinan Han[5], Marco Pavone[2,3], and Yiming Li[1†]

[1] Tsinghua University, China [2] NVIDIA, USA [3] Stanford University, USA
[4] King's College London, UK [5] Technical University of Darmstadt, Germany
{Grady10086, tianyuxin1024, yimingli9702}@gmail.com
https://vsi-super-wild.github.io/

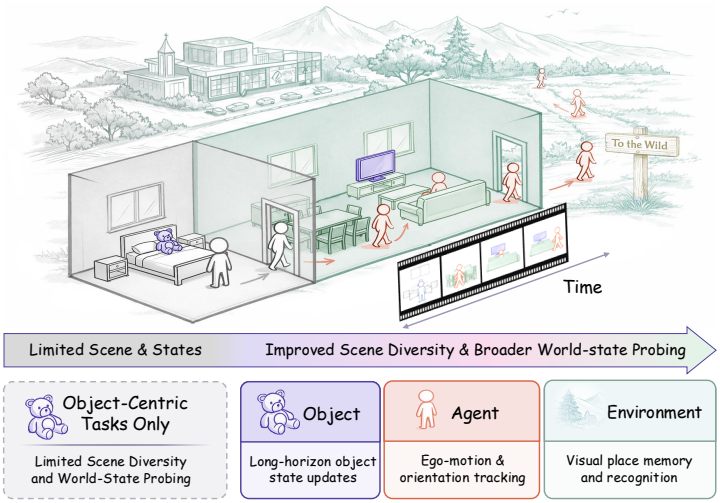


Fig. 1: We advance spatial supersensing beyond household object modeling to world modeling across three anchors—agent, object, environment—using long-form, in-the-wild videos from diverse real-world scenes.

Abstract. Humans can efficiently parse continuous sensory streams, from hours to years, scaffolding an internal world model that grounds spatial reasoning and prediction. To mimic this capacity, spatial supersensing [59] challenges multimodal models to move beyond linguistic understanding toward true world modeling. However, their benchmark relies on synthetic long videos—formed by concatenating random short clips—and is mostly limited to household scenes, leaving real-world continuity and diversity underexplored. To address the gap, we introduce **VSI-Super-Wild**, a large-scale spatial supersensing benchmark built from genuinely long, in-the-wild videos across diverse scenarios. Notably, inspired by cognitive studies on how humans structure experience, we systematically probe the full triad of world state: the agent (observer), objects (scene items), and the environment (places and global layout).

[*] Equal contribution. [†] Corresponding author.

In total, VSI-Super-Wild comprises 442 real-world long-form videos across 8 scene categories and 6,980 human-verified question-answer pairs. Evaluating multimodal models on VSI-Super-Wild exposes a fundamental disconnect: despite advances in static image understanding, models consistently fail at tasks that require coherent world-state tracking over time. We characterize how performance degrades with world-state complexity and temporal horizon, and diagnose four failure modes: spatial collapse, semantic shortcuts, insufficient update, and instance confusion. This taxonomy reveals that models lack the mechanisms to bind objects, agents, and environments into a unified spatial world model—a fundamental gap that defines the path forward for spatial supersensing.

Keywords: Spatial Supersensing · MLLMs · Video Understanding

1 Introduction

Humans bind what we see (*e.g.*, objects and places) with what we do (*e.g.*, ego motion and actions) into a unified world model from continuous visual experience, enabling us to answer, days or months later, where we left our keys or whether we turned off the stove [22, 23]. Cognitive science suggests that this capability does not arise from *recording a complete, verbatim stream of visual input*, but from *constructing and maintaining internal representations* that integrate external information (*e.g.*, object features and environmental geometry) with self-related information (*e.g.*, ego pose and motion), thereby modeling how the world evolves over space and time [5, 10, 11, 45].

How close are existing multimodal large language models (MLLMs) to building internal world models from streaming visual inputs? Despite rapid progress on video benchmarks, many tasks can still be solved with sparse-frame evidence, local matching, or textual priors—leaving true world modeling unexplored [9,24,27,30,37]. To push beyond language-only understanding, Cambrian-S introduces *spatial supersensing* [59]: the ability of MLLMs to construct, update, and predict with an implicit 3D world model from continual sensory experience. Its benchmark, VSI-Super, defines two challenging tasks, *i.e.*, spatial recall (VSR) and continual counting (VSC), to test whether MLLMs can construct and maintain world states across long horizons and changing viewpoints.

Despite its pioneering contributions, VSI-Super leaves several key dimensions of spatial supersensing underexplored. *First*, it predominantly includes household scenes, with limited exposure to in-the-wild diversity [63]. *Second*, its long videos are synthesized by concatenating random short clips with in-frame editing, departing from fully natural sensory streams. *Third*, it emphasizes object-centric probing while underrepresenting agent- and environment-centered aspects of world state. These limitations motivate a new benchmark moving spatial supersensing toward real streams, real scenes, and full world-state coverage.

To bridge these gaps, we introduce **VSI-Super-Wild**, a large-scale video benchmark for evaluating spatial supersensing in the wild. It comprises 442 real-world videos (up to 4+ hours) across 8 scene categories, with 6,980 human-

verified Q&A pairs. As illustrated in Fig. 1, VSI-SUPER-WILD advances beyond prior work along two axes: *real-world diversity* and *world-state coverage*.

- For *real-world diversity*, we curate long-form in-the-wild videos spanning diverse scene categories, camera movements, and object semantics—with natural continuity and no generative editing or clip concatenation.
- For *world-state coverage*, we broaden evaluation beyond objects through a cognitively grounded, multi-anchor design inspired by neuronal vector coding [8]. We view spatiotemporal world states as organized around three complementary anchors—*agent*, *objects*, and *environment*.

Specifically, VSI-SUPER-WILD includes four tasks that target different anchors of world state: **VMR** (motion orientation recall) probes the *agent* by testing camera motion orientation; **VPO** (place temporal ordering) probes the *environment* by testing memory of place order under varying viewpoints; **VOO** (object temporal ordering) probes *objects* by requiring temporal reasoning about when specific items appear; and **VOC** (continuous object counting) integrates across anchors by requiring persistent tracking of objects as the agent moves through the environment. Together, these tasks require true world modeling that can unify agent, object, and environment across space and time.

We evaluate 13 mainstream MLLMs on VSI-SUPER-WILD, where they exhibit consistently weak performance—often near chance, suggesting that they still struggle to *maintain coherent, recallable world states* under the real, diverse in-the-wild settings. To pinpoint where this gap is most pronounced, we conduct *quantitative analyses on task demands and temporal horizon*. Across tasks, models are relatively more reliable on object-centric state queries, but degrade noticeably on queries that require stronger 3D spatial cognition about the environment or agent. Across temporal horizons, performance drops substantially as videos become longer, indicating difficulty in sustaining world-state updates over extended streams and a tendency for errors to accumulate rather than be corrected with additional evidence. We further conduct *qualitative error analyses* to understand why these failures occur, and observe recurring patterns of spatial collapse, semantic shortcuts, insufficient update, and instance confusion, reinforcing that today's MLLMs often default to frame-level semantics and lack stable spatiotemporal representations that support long-horizon reasoning. Taken together, these observations suggest that progress on spatial supersensing in the wild may hinge on more principled spatiotemporal world modeling—i.e., representations that move beyond frame-level semantics while remaining consistent and coherent under long-horizon, continually updated streams.

In summary, our contributions are:

- **In-the-Wild Video Benchmark.** We curate 442 long-form videos across 8 scene categories, with 6,980 human-verified Q&As, providing a large-scale benchmark for spatial supersensing in unconstrained, real-world settings.
- **Multi-Anchor Task Suite.** We design four cognitively grounded tasks that probe world states beyond objects, spanning agent, object, and environment anchors, to systematically evaluate implicit world modeling.

– **Diagnostic Insights.** We benchmark thirteen mainstream MLLMs on **VSI-SUPER-WILD**, delivering task- and horizon-wise quantitative analyses and qualitative error diagnoses that expose recurring failure modes and identify open challenges for spatial supersensing.

2 Related Work

Video Multimodal Large Language Models. Recent Multimodal Large Language Models (MLLMs) signify a strategic paradigm shift from static image comprehension [1,19,29,34,66] to complex video streams understanding [16,21,28,65,67]. This modality shift from image to video naturally introduces *temporality*, creating growing demands for memory and streaming mechanisms that can retain, update, and process information over extended visual sequences [20,42,47,56,64]. Organized as sequences of 2D frames, videos also push models beyond frame-wise understanding and infer an underlying 3D spatial world, motivating recent efforts on spatial understanding and reasoning [12,15,17,35,38,53,61]. Building on progress in temporal modeling and spatial reasoning, several recent works have begun to further explore whether MLLMs can *model the world* by constructing and updating the world states over continuous visual experience [33, 59]. However, current validation of such world-modeling ability remains underexplored on several key dimensions with simplified evaluation settings, leaving the need for more rigorous evaluation of how MLLMs construct and update spatiotemporal world states from real-world continuous visual experience.

Video Understanding Benchmarks for MLLMs. Existing video benchmarks for MLLMs already span a broad landscape, including general video understanding [24, 27, 30, 41, 46], long-horizon video evaluation [14, 50, 54], and 3D spatial reasoning [26,31,43,49,57,58,62]. However, empirical results show that they still highly rely on linguistic priors [44, 55, 60], and often answer QAs correctly by semantic shortcut rather than establishing a coherent world representation [9]. VSI-SUPER therefore marks an important step by evaluating the *spatial supersensing* capability of MLLMs, which explicitly aims to reduce such shortcuts and push evaluation toward implicit 3D world modeling and predictive world-state understanding [59]. Yet several important dimensions remain underexplored, as its evaluation primarily focuses on object-centric states, while its data are built from relatively synthetic and scene-restricted settings rather than diverse real-world continuous videos. These limitations motivate the need for a more rigorous benchmark that evaluates MLLM world modeling over broader world-state coverage and more diverse in-the-wild scenes.

World Modeling over Video Streams. World models, broadly understood as neural systems that internalize and simulate the spatiotemporal dynamics of the physical world, have recently gained renewed attention in both generative paradigms [6, 40, 52, 61] and representation-learning paradigms [2, 7, 36, 39]. In

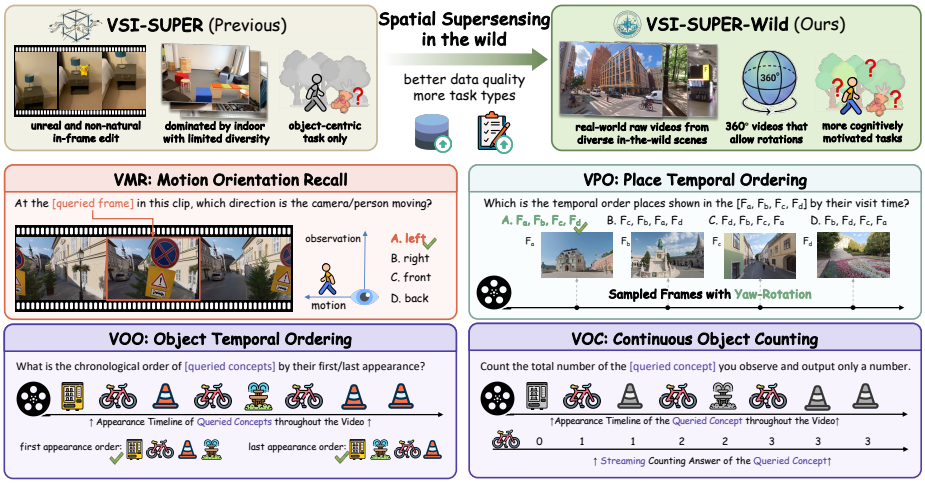


Fig. 2: Overview of VSI-Super-Wild. We extend spatial supersensing from synthetic indoor scenes to real-world, 360° outdoor environments with better data quality and more task types. Our benchmark probes the agent-object-environment triad through four tasks: **VMR** (motion orientation recall), **VPO** (place temporal ordering), **VOO** (object temporal ordering), and **VOC** (continuous object counting). Together, these tasks demand coherent world modeling across diverse, unconstrained scenarios.

parallel, large-scale pretraining could also endow MLLMs with knowledge about spatiotemporal dynamics of the real world, and yet their capability of world modeling has not yet been fully examined. Cambrian-S [59] takes an important step in this direction by introducing MLLM-based predictive world modeling, where models must construct and update world states over continuous video streams. Nevertheless, several important dimensions remain underexplored, as it primarily centers on object-level state modeling and synthetic long-video construction. This raises a natural question: how should MLLM-based world modeling be evaluated more comprehensively? Theories in cognitive science, such as neuronal vector coding [8], suggest that biological spatiotemporal cognition is organized around three anchors: *the agent, the object, and the environment*. This perspective motivates the evaluation of world modeling in MLLMs with richer world states and more diverse real-world scenes.

3 VSI-Super-Wild

To bridge the gap between existing benchmarks and the real-world spatial supersensing, we introduce VSI-Super-Wild, constructed from genuinely long-form, in-the-wild videos. This section details the design and construction of our benchmark. We first motivate the four tasks (§3.1) through the lens of cognitive science. We then describe the semi-automatic data curation pipeline with human-in-the-loop

verification (§3.2) and present comprehensive dataset statistics that highlight the scale, diversity, and complexity of VSI-SUPER-WILD (§3.3).

3.1 Task Motivation and Setup

To evaluate spatial supersensing in unconstrained settings, VSI-SUPER-WILD is designed around the triad of world modeling: *the agent, objects, and the environment*. By leveraging genuinely long-form, unedited panoramic videos and comprehensive tasks, our benchmark enables broader world state probing and improved real-world diversity compared to prior benchmarks. We define four cognitively grounded tasks to systematically assess how well MLLMs can construct and maintain these specific world states.

VMR: Motion Orientation Recall. After observing a long video, **VMR** requires MLLMs to infer the motion orientation relative to its viewing direction at a queried moment, which is specified by the frame sampled there. Such task *cannot* be solved by frame-level matching on the queried moment, since a single frame is often insufficient to determine motion orientation. Instead, the model must recall the latent world state at that moment to decode the corresponding motion orientation. The VMR benchmark therefore probes spatial supersensing by *requiring implicit world modeling* over long-horizon observations, without which such implicit motion states cannot be reliably recalled.

VPO: Place Temporal Ordering. After observing a long panoramic video, VPO asks the model to determine the temporal order of four query frames sampled from different moments representing different places. Each query frame is yaw-rotated to a different viewing direction, preventing trivial frame matching solution. We construct VPO benchmark from panoramic videos so that these rotations are lossless and realistic. With the same place presented under different headings, VPO thus probes spatial supersensing by testing whether models can maintain implicit, heading-invariant place representations and use them to reconstruct the observer's experience over space and time.

VOO: Object Temporal Ordering. After observing a long video, VOO queries a set of objects and asks the model to output their temporal order conditioned on the queried occurrence type, i.e., the *first* or *last* time each object becomes visible. Compared to the VSR benchmark in VSI-SUPER, VOO differs in two key aspects: (i) it uses real, unedited object observations, whereas VSR relies on generative, non-natural objects inserted via in-frame editing; and (ii) it explicitly requires distinguishing first-versus-last occurrences of queried objects within a continuous stream. Consequently, VOO probes spatial supersensing by testing whether models can perform *self-updating* world modeling—maintaining a dynamically updated world state of object encounters over time and space to support occurrence-conditioned queries.

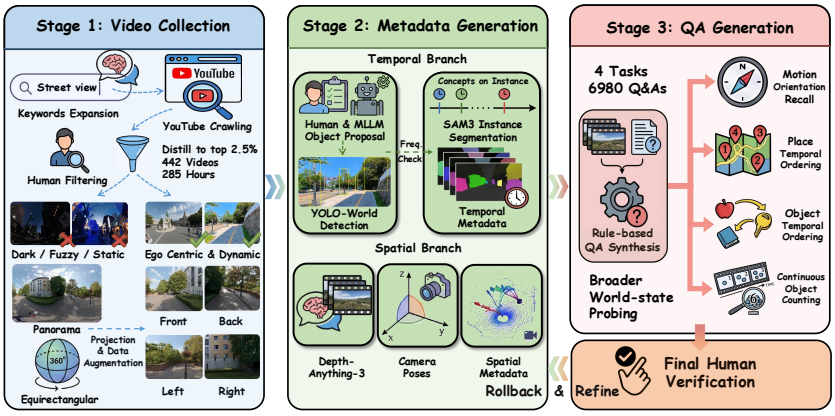


Fig. 3: Data construction pipeline of VSI-SUPER-WILD. We design a semi-automatic pipeline including video collection, metadata generation, Q&A generation, with human-in-the-loop verification.

VOC: Continuous Object Counting. After observing a long video stream, VOC queries an object category and requires the model to output its unique-instance counting number. This task follows the VSC setup in VSI-SUPER, but is instantiated on real, unedited long in-the-wild videos, where the visual evidence for counting is more natural and video streams are more coherent. Consequently, VOC probes spatial supersensing by testing whether models can perform long-horizon, *self-updating* world modeling that maintains a consistent latent state of object counts over time and grounding it into an accurate numerical answer.

3.2 Dataset Construction

As shown in Fig. 3, the construction of our dataset follows a *scalable*, *semi-automatic* pipeline with *human-in-the-loop* verification, consisting of three stages: video collection, metadata generation, and question-answer generation. We describe each stage in detail below.

Video Collection. We crawl in-the-wild panoramic videos from YouTube spanning eight scene categories (culture & entertainment, industry, medical, office & education, residential space, retail space, street view, and transportation hubs). To ensure data quality, we manually filter the crawled videos and retain predominantly egocentric and dynamic recordings. As panoramic videos are typically stored in an equirectangular format, we project each panorama into four orthogonal perspective views (Front/Back/Left/Right) and treat perspective videos as independent inputs for subsequent processing. Unless otherwise specified, the term "video" hereafter refers to these perspective videos.

Metadata Generation. From each video, we generate temporal and spatial metadata to support downstream Q&A construction. First, *temporal metadata* cap-

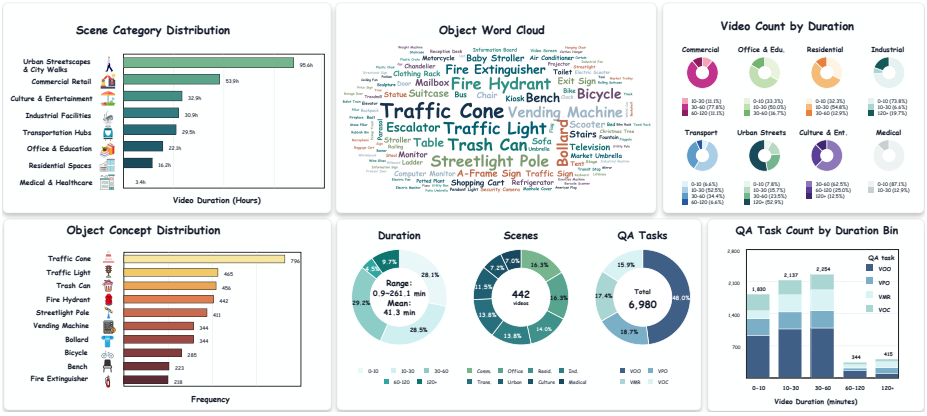


Fig. 4: Statistics of VSI-Super-Wild. The figure reports scene-category duration statistics, object-frequency statistics from VOO/VOC tasks, video-count distributions over duration bins for each scene category, object-concept frequencies, overall single-video duration and QA task distributions, and QA task counts by duration bin.

tures object appearances over time by (i) video-specific candidate objects proposed by human experts and an MLLM, (ii) filtering candidates with YOLO-World [18] based on their temporal occurrence statistics, and (iii) applying Segment-Anything-3 (SAM3) [13] to produce timestamped instance masks for the retained objects. Second, *spatial metadata* consists of time-varying camera poses estimated across the video using Depth-Anything-3 (DA3) [32].

Question-Answer Generation. Given the videos and their temporal and spatial metadata, we synthesize Question–Answer (Q&A) pairs for four tasks: **VMR** (motion orientation recall), **VOO** (object ordering), **VPO** (place ordering), and **VOC** (object counting), where the prefix "V" denotes VSI-Super-Wild. Our Q&A construction is rule-based, with answers provided in either *multiple-choice (MC)* or *numeric* formats. Finally, we incorporate human verification to ensure Q&A correctness, rolling back to refine metadata or Q&A when necessary.

3.3 Benchmark Statistics

After we detail the rigorous curation and synthesis pipeline, the intrinsic value of VSI-Super-Wild lies in the empirical complexity, multi-dimensional distribution, and cognitive difficulty of the video data and tasks. In this section, to demonstrate the scale, diversity, and complexity of VSI-Super-Wild, we present a statistical analysis of the VSI-Super-Wild dataset (see Fig. 4). We deconstruct the data to provide a more comprehensive understanding, including video & meta information, question-answer pairs.

Video & Meta Information. In total, VSI-Super-Wild collects 442 high-quality view-level videos, yielding 284.52 hours (17,071.32 minutes) of egocentric experience in the wild. To ensure that the benchmark accurately reflects open-world

Table 1: Performance on VSI-Super-Wild. We report task scores as mean ± standard error of the mean (SEM) for **VMR**, **VPO**, **VOO** (accuracy, %), and **VOC** (MRA, %). **Overall** denotes the question-level aggregate score across all tasks. (**Bold**: best, underline: second best)

Model	Base LM	VMR	VPO	VOO	VOC	Overall
Proprietary Models						
GPT-5.4	UNK.	**37.76±3.02**	24.87±2.63	37.62±3.13	32.94±1.11	34.52±1.68
Gemini-3.1-Pro	UNK.	34.21±2.26	**63.84±4.46**	**42.54±1.35**	**38.16±5.13**	**44.36±1.39**
Gemini-3.1-FlashLite	UNK.	28.96±2.30	23.81±1.75	23.17±0.85	20.35±1.80	23.85±0.72
Open-Source Models						
Cambrian-S-0.5B	Qwen2.5-0.5B	24.36±1.23	10.14±0.84	20.87±0.70	33.30±0.95	21.46±0.46
Cambrian-S-3B	Qwen2.5-3B	25.68±1.25	25.42±1.21	38.63±0.84	34.06±0.96	33.18±0.54
Cambrian-S-7B	Qwen2.5-7B	26.91±1.27	25.58±1.21	40.36±0.85	33.81±0.97	34.22±0.54
Cambrian-S-7B-LFP*	Qwen2.5-7B	28.62±1.20	26.06±2.46	39.26±3.69	26.17±7.98	32.86±2.24
Cambrian-S-7B-LFP	Qwen2.5-7B	26.58±1.27	26.11±1.22	40.27±0.85	28.36±0.94	33.35±0.54
InternVL3.5-8B	Qwen3-8B	27.90±1.29	24.65±1.19	35.13±0.82	36.74±0.96	32.18±0.53
Qwen2-VL-7B	Qwen2-7B	27.41±1.28	23.66±1.18	30.21±0.79	18.72±7.98	26.67±1.37
Qwen2.5-VL-7B	Qwen2.5-7B	31.52±1.33	26.11±1.22	35.61±0.83	22.11±0.87	30.98±0.53
Qwen3-VL-8B	Qwen3-8B	25.60±1.25	24.35±1.19	40.36±0.85	32.73±7.98	33.59±1.37
Qwen3.5-9B	Qwen3.5-9B	25.68±1.25	25.19±1.20	41.25±0.85	30.76±0.95	33.87±0.54
Spatial-TTT-nano*	Qwen3-2B	24.53±1.23	24.12±1.19	27.82±0.77	35.93±0.96	27.85±0.51

complexity, the videos span 8 distinct scene categories. Commercial retail accounts for 16.29% of videos by count and 18.95% by duration, while transportation hubs account for 13.80% by count and 10.38% by duration. Unlike highly edited indoor clips, our videos have an average duration of 38.62 minutes (ranging from 0.87 to 261.08 minutes), preserving the natural temporal continuity required for implicit 3D world modeling.

Question-Answer Pairs. Our pipeline yielded a total of 6,980 high-quality Q&A pairs. To ensure a comprehensive evaluation of the agent-object-environment triad, the questions are distributed across the four proposed tasks: VMR (1,215 pairs), VPO (1,302 pairs), VOO (3,350 pairs), and VOC (1,113 pairs). The VOO split contains two balanced subtypes, with 1,675 first-appearance recall questions and 1,675 last-appearance recall questions. Furthermore, we strictly controlled the answer distributions to prevent models from exploiting statistical shortcuts.

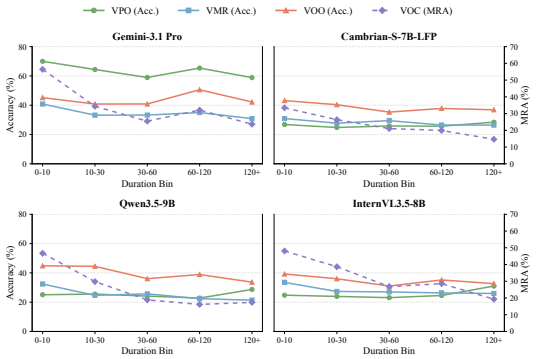
4 Experiment

4.1 Experimental Setup

Models. We benchmark 13 representative MLLMs on VSI-Super-Wild to assess their spatial supersensing capabilities in diverse real-world settings. The evaluated models include 3 proprietary MLLMs, namely GPT-5.4 [48], Gemini-3.1-Pro [25], and Gemini-3.1-FlashLite [25]. We also evaluate 11 open-source MLLMs: Cambrian-S-0.5B/3B/7B(-LFP) [59], Qwen2-VL-7B, Qwen2.5-VL-7B [4], Qwen3-VL-8B [3], Qwen3.5-9B, InternVL3.5-8B [51], and Spatial-TTT-nano [33].

Table 2: Performance across temporal horizons on VSI-Super-Wild. The left panel reports **overall** scores grouped by video duration (minutes). The right panel visualizes per-task scores of four representative models using line charts. (**Bold**: best, underline: second best)

Model/Duration (min)	0–10	10–30	30–60	60–120	120+
Proprietary Models					
GPT-5.4	32.8	39.7	30.8	35.2	35.1
Gemini-3.1-Pro	**51.9**	**42.6**	**40.1**	**46.9**	**41.2**
Gemini-3.1-FlashLite	26.9	24.2	21.8	23.3	20.2
Open-Source Models					
Cambrian-S-0.5B	23.2	21.4	21.5	20.5	14.3
Cambrian-S-3B	37.3	34.4	30.1	29.7	28.7
Cambrian-S-7B	38.5	34.9	31.5	31.3	28.9
Cambrian-S-7B-LFP*	36.8	33.3	30.4	30.7	27.9
Cambrian-S-7B-LFP	39.4	32.7	30.0	32.2	29.7
InternVL3.5-8B	37.5	32.8	28.2	30.5	28.2
Qwen2-VL-7B	29.5	26.0	26.0	21.3	25.7
Qwen2.5-VL-7B	35.9	31.1	27.3	31.5	28.3
Qwen3-VL-8B	36.6	34.7	31.7	32.6	25.9
Qwen3.5-9B	40.1	35.5	29.4	29.3	26.0
Spatial-TTT-nano*	30.5	27.6	26.6	26.4	25.3
Average	35.0	31.3	28.4	28.7	26.3



Evaluation Protocol. Following the evaluation protocol in VSI-Super [59], we uniformly sample 128 frames at 1080P from each video for all models by default. Models marked with * in Table 1 are evaluated with streaming inputs instead of the 128-frame sampled input. Query prompts for VMR, VPO, VOO, and VOC are included in the task examples shown in the supplementary material.

Metrics. We evaluate all models on four tasks in VSI-Super-Wild: **VMR**, **VPO**, **VOO**, and **VOC**. For **VMR**/**VPO**/**VOO**, each query is formulated as a four-way multiple-choice (MC) question, where models select exactly one answer from four candidate options. We report *Accuracy* (%), defined as the fraction of queries whose selected option matches the ground-truth answer. For **VOC**, each query is formulated as a numeric prediction problem, where models directly output an integer count. We report *Mean Relative Accuracy (MRA)*, defined as $\text{MRA} = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{|\hat{y}_i - y_i|}{\max(y_i, 1)}\right)$, where y_i and $\hat{y}_i$ denote the ground-truth and predicted counts, respectively. This metric equals 1.0 for exact matches and decreases smoothly with proportional counting errors.

4.2 Overall Performance and Key Challenges

Overall Performance. As shown in Tab. 1, current MLLMs still struggle substantially with spatial supersensing in the wild. The strongest model, Gemini-3.1-Pro, reaches an overall score of 44.36, while GPT-5.4 and the strongest open-source model, Cambrian-S-7B, reach 34.52 and 34.22, respectively. In particular, the numeric task VOC remains challenging even for strong models, with GPT-5.4 reaching 32.94 and the strongest open-source model, InternVL3.5-8B, reaching 36.74. Overall, these results indicate that current MLLMs still struggle to robustly construct, maintain, and query broader spatiotemporal world states under *improved real-world diversity*.

Challenges from World State Complexity. Beyond overall performance, the results vary systematically across tasks with different world-state anchors. For a controlled comparison, we focus on VOO, VPO, and VMR, which are all multiple-choice tasks evaluated with the same metric and respectively probe object-, environment-, and agent-centric world states. As shown in Tab. 1, object-state probing is often more tractable for open-source models: for example, Cambrian-S-7B achieves 40.36 on VOO, compared with 26.91 on VMR and 25.58 on VPO, while Qwen3.5-9B reaches 41.25 on VOO but only 25.68 and 25.19 on VMR and VPO. These results suggest that *agent- and environment-state modeling are systematically harder than object-state modeling* for current MLLMs. Moreover, prior spatial supersensing benchmarks, which focus on object-centric probing, may underexplore this systematic challenge from world-state complexity.

Challenges over Longer Temporal Horizons. We further analyze how performance varies with temporal horizon. As shown in Tab. 2, longer videos are generally more challenging for spatial supersensing: the open-source average score drops from 35.0 in the 0–10 min bin to 26.3 in the 120+ min bin, with the highest average performance occurring in the shortest videos and the lowest in the longest ones. This trend suggests that maintaining coherent world states becomes *increasingly difficult as temporal horizon increases*. The degradation is not perfectly monotonic for every model, but the overall trend shows that longer video streams lower performance and expose weaknesses in different types of world-state modeling. In particular, the drop across temporal bins suggests that maintaining and updating object-, agent-, and environment-related states becomes harder as the observation horizon expands.

4.3 Failure Modes of Spatial Supersensing

Our results show that current MLLMs perform poorly across VSI-Super-Wild, suggesting that *spatial supersensing in the wild* remains challenging under diverse scenes and longer temporal horizons. To better understand limitations and provide a roadmap for future research, we conduct both qualitative and quantitative analysis and summarize *four recurring failure modes* that hinder coherent representations of the agent, objects, and environment.

Spatial Collapse. Spatial supersensing requires models to construct a stable spatial world state in the *3D* space, rather than relying on frame-level *2D* semantics. We examine this through the viewpoint robustness of VPO, which tests whether models can recognize the same place under different viewpoints. Specifically, we compare accuracy between the original and rotated views. Fig. 6 shows that SOTA models exhibit a clear performance drop after viewpoint change (e.g., GPT-5.4: 26.0%→24.9%; Qwen3-VL-8B: 25.7%→24.4%; Cambrian-S-7B: 26.2%→25.6%), even though the underlying physical location remains unchanged. Such degradation reflects *spatial collapse*: models fail to maintain a coherent spatial world state across views, and instead fall back to view-specific *2D* frame matching rather than *implicit 3D* world modeling.

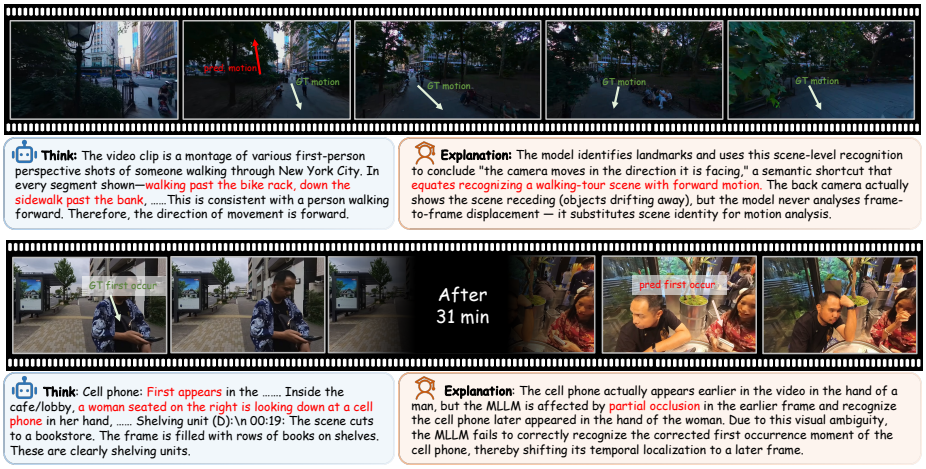


Fig. 5: Top (*semantic shortcut*): The model predicts that the observer is moving forward based on frame-level semantic cues (e.g., a typical walking-tour street scene), rather than inferring motion from spatiotemporal evidence across frames. **Bottom** (*instance confusion*): Under visual challenges such as motion blur and partial occlusion in the wild, the model misses an earlier appearance of the queried object, leading to an incorrect first-occurrence prediction and unstable instance-level tracking over time.

Semantic Shortcut. Existing models often rely on frame-level visual semantics instead of continuous spatiotemporal reasoning, using semantic cues from individual frames as shortcuts for world-state inference. We illustrate this failure mode with VMR, where motion orientation requires constructing an agent-centric state from the video stream, yet models can still be misled by local semantic cues and produce plausible-but-incorrect answers. As shown in the top of Fig. 5, the agent is actually moving backward relative to its viewing direction, but the model predicts "forward" because the straight road ahead provides a strong semantic cue. This failure suggests that models *exploit semantic shortcuts* instead of inferring motion from spatiotemporal evidence and updating an agent-centric world state.

Insufficient Update. Beyond initial recognition, spatial supersensing requires continuously updating world states as video streams arrive. We probe this using VOO first- vs. last-occurrence ordering (Fig. 6), where first-occurrence mainly tests registering objects at their initial encounter, while last-occurrence requires updating world states to reflect later evidence. Across models, accuracy is substantially higher for first-occurrence than for last-occurrence ordering (e.g., GPT-5.4: 31.3% vs. 22.1%; Qwen3-VL-8B: 50.2% vs. 30.5%; Cambrian-S-7B: 50.7% vs. 30.0%). This gap suggests that models can preserve early world states relatively well, but struggle to *sufficiently update world states* as new evidence arrives.

Instance Confusion. For spatial supersensing in the wild, models need to robustly recognize objects under challenging real-world conditions, such as motion blur,

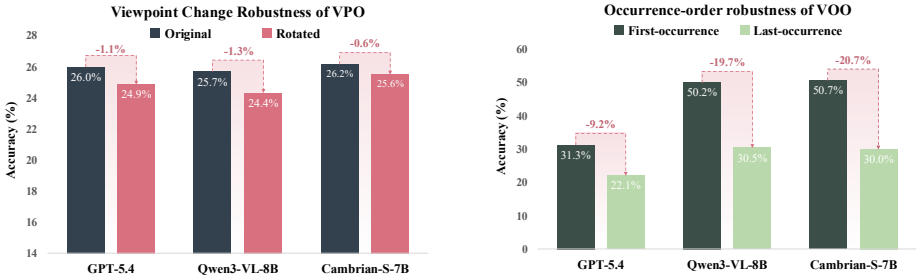


Fig. 6: Left (*spatial collapse*): We compare VPO accuracy between original and rotated views to test whether models can recognize the same place under viewpoint changes. The drop after rotation suggests that models fail to model a consistent world state after changing views despite the spatially unchanged location. **Right** (*insufficient update*): We compare VOO accuracy between first- and last-occurrence ordering. The lower performance on last-occurrence queries indicates that models struggle to update world states as later evidence arrives.

partial occlusion, and viewpoint change. However, as shown in the bottom of Fig. 5, the model misses an earlier appearance of the queried object under blur and partial occlusion, leading to incorrect ordering. This indicates weak object-level identity tracking and unstable object-centric state modeling under in-the-wild visual noise and dynamics. Overall, these failures underscore that *in-the-wild* visual complexity remains highly challenging for spatial supersensing.

4.4 Long-Video Benchmark: VSI-SUPER-WILD vs. VSI-SUPER under Shuffle24

To examine whether VSI-SUPER-WILD introduces additional difficulty beyond existing long-video benchmarks, we compare it with VSI-SUPER under **Shuffle24**. For each multiple-choice question, we evaluate all 24 permutations of the four answer options and report metrics that are invariant to option order. This protocol reduces dependence on a specific option arrangement and reveals whether model predictions are stable under semantically equivalent choices.

Fig. 7 reports four complementary views. **Top-K** measures the fraction of questions with at least one correct prediction in the first K attempts, while **Pass@K** measures the fraction with at least K correct predictions among all 24 permutations. For the same K, VSI-Super-Wild is consistently lower than VSI-Super on both metrics and remains closer to the random-guess baseline, indicating greater difficulty in recovering the correct answer even across shuffled trials. We analyze option-order sensitivity: Fig. 7(c) shows higher **option-order bias** on VSI-Super-Wild, measured by the standard deviation of accuracy across permutations, and Fig. 7(d) shows a larger gap between the best and worst permutations. Overall, VSI-SUPER-WILD is more challenging and less stable under option shuffling, making Shuffle24 necessary to avoid overestimation.

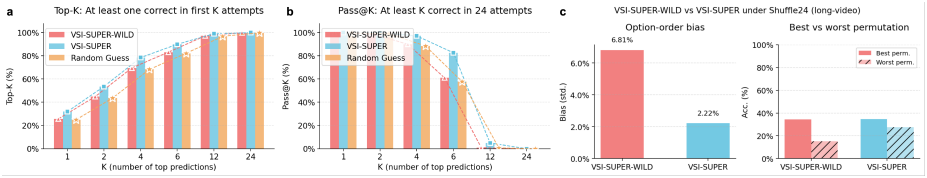


Fig. 7: VSI-SUPER-WILD vs. VSI-SUPER under Shuffle24 (long-video). (**a**) Top-K: fraction of questions with at least one correct in the first K attempts. (**b**) Pass@K: fraction with at least K correct in 24 attempts. (**c**) Option-order bias: standard deviation of accuracy across 24 option permutations, and best vs. worst permutation accuracy. VSI-SUPER-WILD shows lower Top-K and Pass@K, higher bias, and a larger best–worst gap than VSI-SUPER, confirming it is more challenging and that Shuffle24 evaluation is essential.

5 Conclusion

To assess whether MLLMs can conduct human-like implicit world modeling under real-world in-the-wild streams, we introduce **VSI-Super-Wild**, a large-scale benchmark for evaluating *spatial supersensing in the wild*. Compared with existing supersensing benchmarks, VSI-Super-Wild advances evaluation along two axes: *real-world diversity* and *broader world-state probing*. For real-world diversity, we curate 442 unconstrained online videos, spanning 8 scene categories and lasting over 4 hours at maximum, without generative editing or clip concatenation. For broader world-state probing, we construct 6,980 human-verified Q&As across four task types, forming a cognitively grounded task suite that probes object-, agent-, and environment-centric components of world states.

We benchmark 13 mainstream MLLMs on VSI-Super-Wild and find that current models still fall substantially short of robust spatial supersensing in the wild. A closer look reveals clear *performance disparities* across supersensing conditions: models are relatively more reliable on object-centric state queries, yet struggle on agent- and environment-centric components that require stronger 3D spatial cognition; tasks that demand precise numeric answers under continual state updates, such as continuous counting, are especially challenging. Moreover, performance degrades markedly as the temporal horizon increases, suggesting difficulty in sustaining consistent world-state updates over extended streams. Qualitative error analyses further point to recurring failure patterns—spatial collapse, semantic shortcuts, insufficient update, and instance confusion—which help explain these disparities and indicate that current MLLMs still over-rely on frame-level semantics rather than maintaining stable, queryable spatiotemporal world representations. Together, these findings highlight the key bottlenecks that future models must address to achieve reliable spatial supersensing in the wild.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning. In: Advances in Neural Information Processing Systems. vol. 35 (2022)
2. Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Komeili, M., Muckley, M., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., Arnaud, S., Gejji, A., Martin, A., Robert Hogan, F., Dugas, D., Bojanowski, P., Khalidov, V., Labatut, P., Massa, F., Szafraniec, M., Krishnakumar, K., Li, Y., Ma, X., Chandar, S., Meier, F., LeCun, Y., Rabbat, M., Ballas, N.: V-jepa 2: Self-supervised video models enable understanding, prediction and planning. arXiv preprint arXiv:2506.09985 (2025)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report (2025), https://arxiv.org/abs/2511.21631
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923
5. Bainbridge, W.A., Dilks, D.D., Oliva, A.: Memorability: A stimulus-driven perceptual neural signature distinctive from memory. NeuroImage **149**, 141–152 (2017)
6. Ball, P.J., Bauer, J., Belletti, F., Brownfield, B., Ephrat, A., Fruchter, S., Gupta, A., Holsheimer, K., Holynski, A., Hron, J., Kaplanis, C., Limont, M., McGill, M., Oliveira, Y., Parker-Holder, J., Perbet, F., Scully, G., Shar, J., Spencer, S., Tov, O., Villegas, R., Wang, E., Yung, J., Baetu, C., Berbel, J., Bridson, D., Bruce, J., Buttimore, G., Chakera, S., Chandra, B., Collins, P., Cullum, A., Damoc, B., Dasagi, V., Gazeau, M., Gbadamosi, C., Han, W., Hirst, E., Kachra, A., Kerley, L., Kjems, K., Knoepfel, E., Koriakin, V., Lo, J., Lu, C., Mehring, Z., Moufarek, A., Nandwani, H., Oliveira, V., Pardo, F., Park, J., Pierson, A., Poole, B., Ran, H., Salimans, T., Sanchez, M., Saprykin, I., Shen, A., Sidhwani, S., Smith, D., Stanton, J., Tomlinson, H., Vijaykumar, D., Wang, L., Wingfield, P., Wong, N., Xu, K., Yew, C., Young, N., Zubov, V., Eck, D., Erhan, D., Kavukcuoglu, K., Hassabis, D., Gharamani, Z., Hadsell, R., van den Oord, A., Mosseri, I., Bolton, A., Singh, S., Rocktäschel, T.: Genie 3: A new frontier for world models (2025)
7. Bardes, A., Garrido, Q., Ponce, J., Rabbat, M., LeCun, Y., Assran, M., Ballas, N.: Revisiting feature prediction for learning visual representations from video. arXiv:2404.08471 (2024)
8. Bicanski, A., Burgess, N.: Neuronal vector coding in spatial cognition. Nature Reviews Neuroscience **21**, 453–470 (2020)
9. Brown, E., Yang, J., Yang, S., Fergus, R., Xie, S.: Benchmark designers should "train on the test set" to expose exploitable non-visual shortcuts (2025), https://arxiv.org/abs/2511.04655

10. Burgess, N.: Spatial memory: How egocentric and allocentric combine. Trends in Cognitive Sciences **10**(12), 551–557 (2006)
11. Byrne, P., Becker, S., Burgess, N.: Remembering the past and imagining the future: a neural model of spatial memory and imagery. Psychological Review **114**(2), 340–375 (Apr 2007)
12. Cai, Z., Wang, R., Gu, C., Pu, F., Xu, J., Wang, Y., Yin, W., Yang, Z., Wei, C., Sun, Q., Zhou, T., Li, J., Pang, H.E., Qian, O., Wei, Y., Lin, Z., Shi, X., Deng, K., Han, X., Chen, Z., Fan, X., Deng, H., Lu, L., Pan, L., Li, B., Liu, Z., Wang, Q., Lin, D., Yang, L.: Scaling spatial intelligence with multimodal foundation models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2026)
13. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.H., Zhou, Y., Momeni, L., Hazra, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H.K., Dollár, P., Ravi, N., Saenko, K., Zhang, P., Feichtenhofer, C.: Sam 3: Segment anything with concepts (2025), https://arxiv.org/abs/2511.16719
14. Chandrasegaran, K., Gupta, A., Hadzic, L.M., Kota, T., He, J., Eyzaguirre, C., Durante, Z., Li, M., Wu, J., Li, F.F.: Hourvideo: 1-hour video-language understanding. In: Advances in Neural Information Processing Systems. vol. 37 (2024)
15. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., Xia, F.: Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14455–14465 (June 2024)
16. Chen, Y., Huang, W., Shi, B., Hu, Q., Ye, H., Zhu, L., Liu, Z., Molchanov, P., Kautz, J., Qi, X., Liu, S., Yin, H., Lu, Y., Han, S.: Scaling rl to long videos. arXiv preprint arXiv: 2507.07966 (2025)
17. Cheng, A.C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., Liu, S.: Spatialrgpt: Grounded spatial reasoning in vision-language models. In: NeurIPS (2024)
18. Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., Shan, Y.: Yolo-world: Real-time open-vocabulary object detection. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR) (2024)
19. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.C.H.: Instructblip: Towards general-purpose vision-language models with instruction tuning. In: Advances in Neural Information Processing Systems. vol. 36 (2023)
20. Di, S., Yu, Z., Zhang, G., Li, H., Zhong, T., Cheng, H., Li, B., He, W., Shu, F., Jiang, H.: Streaming video question-answering with in-context video kv-cache retrieval. In: The Thirteenth International Conference on Learning Representations (2025)
21. Feng, K., Gong, K., Li, B., Guo, Z., Wang, Y., Peng, T., Wang, B., Yue, X.: Video-r1: Reinforcing video reasoning in mllms. arXiv preprint arXiv:2503.21776 (2025)
22. Finley, J.R., Brewer, W.F.: Accuracy and completeness of autobiographical memory: evidence from a wearable camera study. Memory **32**(8), 1012–1042 (2024)
23. Finley, J.R., Brewer, W.F., Benjamin, A.S.: The effects of end-of-day picture review and a sensor-based picture capture procedure on autobiographical memory using sensecam. Memory **19**(7), 796–807 (2011)

24. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: CVPR (2025)
25. Google: Gemini 3 – google deepmind. https://deepmind.google/models/gemini/ (Dec 2025), accessed: 2026-03-05
26. Gu, T., Gong, J., Zhang, Z., Xie, Y., Ma, L., Tan, X., V, A.: Vision-language models lag human performance on physical dynamics and intent reasoning (2026), https://arxiv.org/abs/2601.01547
27. Hu, K., Wu, P., Pu, F., Xiao, W., Zhang, Y., Yue, X., Li, B., Liu, Z.: Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos (2025), https://arxiv.org/abs/2501.13826
28. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
29. Li, J., Li, D., Savarese, S., Hoi, S.C.H.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 19730–19742 (2023)
30. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., Wang, L., Qiao, Y.: Mvbench: A comprehensive multi-modal video understanding benchmark (2024), https://arxiv.org/abs/2311.17005
31. Li, Y., Zhang, Y., Lin, T., Liu, X., Cai, W., Liu, Z., Zhao, B.: Sti-bench: Are mllms ready for precise spatial-temporal world understanding? arXiv preprint arXiv:2503.23765 (2025)
32. Lin, H., Chen, S., Liew, J.H., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
33. Liu, F., Wu, D., Chi, J., Cai, Y., Hung, Y.H., Yu, X., Li, H., Hu, H., Rao, Y., Duan, Y.: Spatial-ttt: Streaming visual-based spatial intelligence with test-time training. arXiv preprint arXiv:2603.12255 (2026)
34. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems. vol. 36 (2023)
35. Ma, W., Chou, Y.C., Liu, Q., Wang, X., Melo, C.M.d., Xie, J., Yuille, A.: Spatial-reasoner: Towards explicit and generalizable 3d spatial reasoning. In: Advances in Neural Information Processing Systems. vol. 38 (2025)
36. Maes, L., Le Lidec, Q., Scieur, D., LeCun, Y., Balestriero, R.: Leworldmodel: Stable end-to-end joint-embedding predictive architecture from pixels (2026)
37. Mangalam, K., Akshulakov, R., Malik, J.: Egoschema: A diagnostic benchmark for very long-form video language understanding (2023), https://arxiv.org/abs/2308.09126
38. Mao, Y., Zhong, J., Fang, C., Zheng, J., Tang, R., Zhu, H., Tan, P., Zhou, Z.: SpatialLM: Training large language models for structured indoor modeling. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
39. Mur-Labadia, L., Muckley, M., Bar, A., Assran, M., Sinha, K., Rabbat, M., LeCun, Y., Ballas, N., Bardes, A.: V-jepa 2.1: Unlocking dense features in video self-supervised learning. arXiv preprint arXiv:2603.14482 (2026)
40. OpenAI: Video generation models as world simulators. https://openai.com/research/video-generation-models-as-world-simulators (Feb 2024), accessed: 2026-03-05

41. Pătrăucean, V., Smaira, L., Gupta, A., Continente, A.R., Markeeva, L., Banarse, D., Koppula, S., Heyward, J., Malinowski, M., Yang, Y., Doersch, C., Matejovicova, T., Sulsky, Y., Miech, A., Frechette, A., Klimczak, H., Koster, R., Zhang, J., Winkler, S., Aytar, Y., Osindero, S., Damen, D., Zisserman, A., Carreira, J.: Perception test: A diagnostic benchmark for multimodal video models. In: Advances in Neural Information Processing Systems (2023), https://openreview.net/forum?id=HYEGXFnPoq
42. Qian, R., Ding, S., Dong, X., Zhang, P., Zang, Y., Cao, Y., Lin, D., Wang, J.: Dispider: Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
43. Ramakrishnan, S.K., Wijmans, E., Kraehenbuehl, P., Koltun, V.: Does spatial cognition emerge in frontier models? In: International Conference on Learning Representations (2025)
44. Rawal, I.S., Matyasko, A., Jaiswal, S., Fernando, B., Tan, C.: Dissecting multimodality in VideoQA transformer models by impairing modality fusion. In: Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 42213–42244. PMLR (2024)
45. Rensink, R.A., O'Regan, J.K., Clark, J.J.: To see or not to see: The need for attention to perceive changes in scenes. Psychological Science **8**(5), 368–373 (1997)
46. Shangguan, Z., Li, C., Ding, Y., Zheng, Y., Zhao, Y., Fitzgerald, T., Cohan, A.: TOMATO: Assessing visual temporal reasoning capabilities in multimodal foundation models. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=fCi4o83Mfs
47. Shen, Y., Tian, S., Yang, J., Liu, Z.: A simple baseline for streaming video understanding. arXiv preprint arXiv:2604.02317 (2026)
48. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card (2025), https://arxiv.org/abs/2601.03267
49. Wang, Q., Yin, B., Zhang, P., Zhang, J., Wang, K., Wang, Z., Zhang, J., Chandrasegaran, K., Liu, H., Krishna, R., Xie, S., Li, M., Wu, J., Fei-Fei, L.: Spatial mental modeling from limited views (2025)
50. Wang, W., He, Z., Hong, W., Cheng, Y., Zhang, X., Qi, J., Ding, M., Gu, X., Huang, S., Xu, B., Dong, Y., Tang, J.: Lvbench: An extreme long video understanding benchmark. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 22958–22967 (October 2025)
51. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., Wang, Z., Chen, Z., Zhang, H., Yang, G., Wang, H., Wei, Q., Yin, J., Li, W., Cui, E., Chen, G., Ding, Z., Tian, C., Wu, Z., Xie, J., Li, Z., Yang, B., Duan, Y., Wang, X., Hou, Z., Hao, H., Zhang, T., Li, S., Zhao, X., Duan, H., Deng, N., Fu, B., He, Y., Wang, Y., He, C., Shi, B., He, J., Xiong, Y., Lv, H., Wu, L., Shao, W., Zhang, K., Deng, H., Qi, B., Ge, J., Guo, Q., Zhang, W., Zhang, S., Cao, M., Lin, J., Tang, K., Gao, J., Huang, H., Gu, Y., Lyu, C., Tang, H., Wang, R., Lv, H., Ouyang, W., Wang, L., Dou, M., Zhu, X., Lu, T., Lin, D., Dai, J., Su, W., Zhou, B., Chen, K., Qiao, Y., Wang, W., Luo, G.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency (2025), https://arxiv.org/abs/2508.18265
52. World Labs: Marble: A multimodal world model. https://www.worldlabs.ai/blog/marble-world-model (Nov 2025), accessed: 2026-03-05

53. Wu, D., Liu, F., Hung, Y.H., Duan, Y.: Spatial-MLLM: Boosting MLLM capabilities in visual-based spatial intelligence. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
54. Wu, H., Li, D., Chen, B., Li, J.: Longvideobench: A benchmark for long-context interleaved video-language understanding. arXiv preprint arXiv: 2407.15754 (2024)
55. Xiao, J., Yao, A., Li, Y., Chua, T.S.: Can i trust your answer? visually grounded video question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13204–13214 (2024)
56. Xiong, H., Yang, Z., Yu, J., Zhuge, Y., Zhang, L., Zhu, J., Lu, H.: Streaming video understanding and multi-round interaction with memory-enhanced knowledge. In: The Thirteenth International Conference on Learning Representations (2025)
57. Xu, R., Wang, W., Tang, H., Chen, X., Wang, X., Chu, F.J., Lin, D., Feiszli, M., Liang, K.J.: Multi-spatialmllm: Multi-frame spatial understanding with multi-modal large language models. In: CVPR (2026)
58. Yang, J., Yang, S., Gupta, A., Han, R., Fei-Fei, L., Xie, S.: Thinking in Space: How Multimodal Large Language Models See, Remember and Recall Spaces. In: CVPR (2025)
59. Yang, S., Yang, J., Huang, P., Brown, E., Yang, Z., Yu, Y., Tong, S., Zheng, Z., Xu, Y., Wang, M., Lu, D., Fergus, R., LeCun, Y., Fei-Fei, L., Xie, S.: Cambrian-s: Towards spatial supersensing in video. arXiv preprint arXiv:2511.04670 (2025)
60. Yang, Y., Guo, Y., Lu, H., Wang, Y.: Vidlbeval: Benchmarking and mitigating language bias in video-involved lvlms. arXiv preprint arXiv: 2502.16602 (2025)
61. Yang, Y., Liu, J., Zhang, Z., Zhou, S., Tan, R., Yang, J., Du, Y., Gan, C.: Mindjourney: Test-time scaling with world models for spatial reasoning. arXiv preprint arXiv:2507.12508 (2025)
62. Yeh, C.H., Wang, C., Tong, S., Cheng, T.Y., Wang, R., Chu, T., Zhai, Y., Chen, Y., Gao, S., Ma, Y.: Seeing from another perspective: Evaluating multi-view understanding in mllms. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 12000–12008 (2026)
63. Yu, H., Han, Y., Zhang, X., Yin, B., Chang, B., Han, X., Liu, X., Zhang, J., Pavone, M., Feng, C., Xie, S., Li, Y.: Thinking in 360°: Humanoid visual search in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2026)
64. Zhang, H., Wang, Y., Tang, Y., Liu, Y., Feng, J., Jin, X.: Flash-vstream: Efficient real-time understanding for long video streams. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2025)
65. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713 (2024)
66. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. In: The Twelfth International Conference on Learning Representations (2024)
67. Zohar, O., Wang, X., Dubois, Y., Mehta, N., Xiao, T., Hansen-Estruch, P., Yu, L., Wang, X., Juefei-Xu, F., Zhang, N., Yeung-Levy, S., Xia, X.: Apollo: An exploration of video understanding in large multimodal models (2024)