StreamSpatial: A Benchmark and Framework for Streaming 3D Visual-Spatial Reasoning

Junlin Xie[1,2,*], Keyang Zhong[4,5,*], Quanlong Zheng[3,†], Ruifei Zhang[1,2], Kuo Wang[4], Yanhao Zhang[3,†], Haonan Lu[3], Xiang Wan[2], and Guanbin Li[4,5,6,‡]

[1] The Chinese University of Hong Kong, Shenzhen
[2] Shenzhen Research Institute of Big Data
[3] OPPO AI Center, OPPO Inc., China
[4] Sun Yat-sen University
[5] Shenzhen Loop Area Institute
[6] Guangdong Key Laboratory of Big Data Analysis and Processing

junlinxie@link.cuhk.edu.cn, liguanbin@mail.sysu.edu.cn

[*] Equal contribution [†] Project leader [‡] Corresponding author

Abstract. Embodied systems demand continuous streaming perception, yet prevailing offline benchmarks assume global spatial omniscience, neglecting real-time spatial intelligence. We introduce StreamSpatial-Bench, a testbed for continuous visual-spatial reasoning with fine-grained online temporal perspectives, dynamic multi-agent interactions, and comprehensive 3D spatial reasoning. Evaluating 18 MLLMs reveals substantial gaps versus humans, attributed to limited spatiotemporal memory, absent incremental updating, and poor dynamic adaptation. We propose Stream3D-Mem, incorporating a memory buffer with selective consolidation via round-decayed compression to prune redundancy while preserving context, and distilling cognitive map knowledge into compact implicit latent tokens to bypass latency-heavy explicit map generation. Future work can leverage neuromorphic architectures and Benchmark to advance human-like 3D spatial intelligence for robotics and AR/VR.

Keywords: Streaming Spatial Reasoning · Multi-agent Interactions · AR/VR

1 Introduction

Humans are inherently streaming visual learners [6,28]. To emulate this capability, embodied systems necessitate a continuous streaming perception framework. This ability involves **incrementally updating spatial environment** (e.g., tracking shifting object relativity like *'the chair is now to my left'*), **performing predictive world modeling** (e.g., anticipating collision risks or inferring occluded geometry), and **robustly adapting to environmental dynamics driven by multi-agent interactions** (e.g., assessing whether *'the person in red is approaching'*). Ultimately, embodied agents should perform robust real-

time 3D spatial reasoning to effectively execute complex instructions or respond to user queries in ever-changing environments [15, 21].

However, prevailing offline benchmarks [10, 17, 26] (e.g., VSI-Bench [23]) presuppose global spatial omniscience, granting a priori access to complete 3D scene representations, thereby failing to capture spatial intelligence as a continuous process tied to real-time sensory streams. To bridge this gap, we introduce **StreamSpatialBench**, a rigorous testbed for continuous visual-spatial reasoning. At its core, StreamSpatialBench rests on three foundational pillars:

- **Fine-grained Online Temporal Perspective**: anchoring temporal relations (past, current, future) to the query timestamp in contrast to offline post-capture analysis, thereby enabling fine-grained temporal queries (e.g., "seconds ago" vs. "right now") and necessitating adaptation to time-dependent contexts where answers dynamically evolve as the stream progresses;
- **Realistic Dynamic Spatiotemporal Reasoning**: systematically evaluating robust adaptation to environmental dynamics driven by complex multi-agent interactions;
- **Comprehensive 3D Spatial Reasoning**: requiring agents to perform robust real-time 3D spatial reasoning across shifting egocentric and allocentric perspectives while maintaining long-term memory.

We evaluate 18 representative MLLMs—spanning offline image/video models, specialized embodied models, and streaming video models—against human performance on our StreamSpatialBench. Results reveal substantial gaps between all models and human evaluators. To better understand why MLLMs fall short of human-level spatial reasoning in streaming scenarios, we conduct an in-depth analysis across multiple dimensions. Our key findings are: (1) **Limited spatiotemporal memory**: Current models exhibit constrained capacity for long-term spatial relationships and layout memory, with performance degrading significantly as temporal spans increase or spatial regions expand. (2) **Lack of incremental updating**: Models provide invariant answers to the same questions across different timestamps, suggesting they rely on static statistical features rather than dynamically updating spatial understanding over time. (3) **Poor adaptation to dynamic environments**: In multi-agent dynamic scenarios with continuously changing layouts, model performance deteriorates substantially.

Motivated by these, we propose a simple yet effective framework, **Stream3D-Mem**: (1) a memory buffer for storing and retrieving frame tokens over time, and a selective consolidation strategy with round-decayed compression to prune redundant tokens while preserving recent context, merging short-term buffers into long-term engrams. (2) For spatial reasoning, we bypass latency-heavy explicit map generation by distilling cognitive map knowledge into compact implicit latent tokens, reducing latency and token usage while boosting accuracy.

Extensive experiments show that Stream3D-Mem consistently improves general MLLMs on streaming 3D spatial reasoning, delivering real-time inference with bounded memory, lower latency, and competitive or improved accuracy. Together, StreamSpatialBench and Stream3D-Mem offer both a rigorous testbed and a practical solution toward continuous, human-like spatial intelligence.

2 Related Work

Benchmark	Setting	Fine-grained Temporal	Multi-Agent Perspective Dynamic Environment	3D Spatial Reasoning
VSI-Bench [23]	offline	✗	✗	✓
Spinbench [26]	offline	✗	✗	✓
SPAR-Bench [17]	offline	✗	✗	✓
OmniSpatial [10]	offline	✗	✗	✓
TemporalBench [2]	offline	✓	✗	✗
OST-Bench [13]	online	✗	✗	✓
SpatialStream (Ours)	online	✓	✓	✓

Table 1: Comparison of benchmark datasets across key dimensions. Our SpatialStream benchmark provides comprehensive coverage of streaming video understanding with 3D visual-spatial reasoning capabilities.

Benchmarking Spatial Reasoning. Visual spatial reasoning benchmarks have evolved from basic perception toward complex cognitive tasks. VSI-Bench-[23] pioneered "thinking in space" by distinguishing static and dynamic spatial understanding. Subsequent efforts probe progressively deeper capabilities: VSR [14] targets spatial-relation verification (e.g., "in front of"), BLINK [8] and SITE [19] assess perception and localization, SPAR-Bench [27] measures metric distance estimation, and OmniSpatial [10] examines spatial-compatibility reasoning that couples geometry with physical commonsense. In contrast, our work addresses the *streaming* nature of real-world spatial reasoning, where objects and viewpoints change continuously and evidence must be integrated across multiple views over time. With fine-grained temporal annotations, our benchmark further enables precise tracing of reasoning chains throughout a stream, fostering models that reason robustly under realistic, evolving conditions.

Fine-grained Temporal Understanding. Contemporary video benchmarks have shifted focus from static frame analysis to fine-grained temporal reasoning, demanding models to comprehend complex event sequences and causal relationships over time. Foundational works like Next-QA [22] and Temporal-Bench [2] explicitly evaluate temporal logic and moment localization, while LongVideoBench [20] challenges models to retrieve evidence across extended contexts. Despite these advances, existing efforts predominantly treat time as a sequence of semantic events or 2D bounding boxes, often neglecting the underlying

3D geometric continuity. Our work bridges this gap by integrating fine-grained temporal annotations with 3D spatial reasoning, ensuring models capture how spatial relationships dynamically evolve rather than merely identifying when events occur.

Embodied Agent Learning and Reasoning. Embodied AI studies agents that perceive, reason, and act within physical or simulated environments through continuous sensorimotor interaction. Recent work has shifted from modular pipelines with hand-crafted rules toward end-to-end frameworks built on vision-language models, improving instruction following and zero-shot generalization in long-horizon planning, memory-augmented navigation, and manipulation under partial observability. Simulation platforms such as Habitat [16] and AI2-THOR [12] have standardized embodied evaluation, and benchmarks such as ALFRED [18] and EmbodiedBench [25] target complex instruction following. On the model side, PaLM-E [7], RT-2 [29], and OpenVLA [11] show that large multimodal models can drive low-level control. However, these efforts largely assume *episodic* inputs and overlook a core difficulty of open-world deployment: maintaining a geometrically consistent world model under *continuous streaming* perception and dynamic scene changes. We address this gap with a streaming 3D spatial reasoning benchmark that explicitly evaluates how embodied agents build, retain, and update spatial memory over time, bridging static embodied QA and real-time, multi-view spatial intelligence.

3 StreamingSpatial Bench

In this section, we detail the construction of **StreamSpatialBench**, designed to materialize the foundational pillars of **streaming 3D visual-spatial intelligence** outlined previously. Figure 1 illustrates the specific sub-categories of the benchmark along with corresponding examples. In Sec. 3.1, we elaborate on the task formulations within this dual-dimensional framework, explaining how temporal contexts and capability requirements intersect to create diverse evaluation scenarios. Subsequently, Sec. 3.2 presents the automated QA generation pipeline and rigorous human-in-the-loop validation procedures employed to guarantee question clarity and answer correctness. Finally , we present the statistics of our dataset.

3.1 Task Formulation

Building on the capability-centric design of StreamSpatialBench, we organize our benchmark around five core competencies that define streaming 3D visual-spatial intelligence. Each task is grounded in one of three temporal contexts relative to the query timestamp: Past (P), Current (C), or Future (F). To explicitly characterize the reasoning dependency, we annotate each capability with an $(A \rightarrow B)$ pair, where A denotes the timestamp of the **evidence** and B indicates the timestamp of the **target answer**. Based on this notation, we delineate the following capabilities:

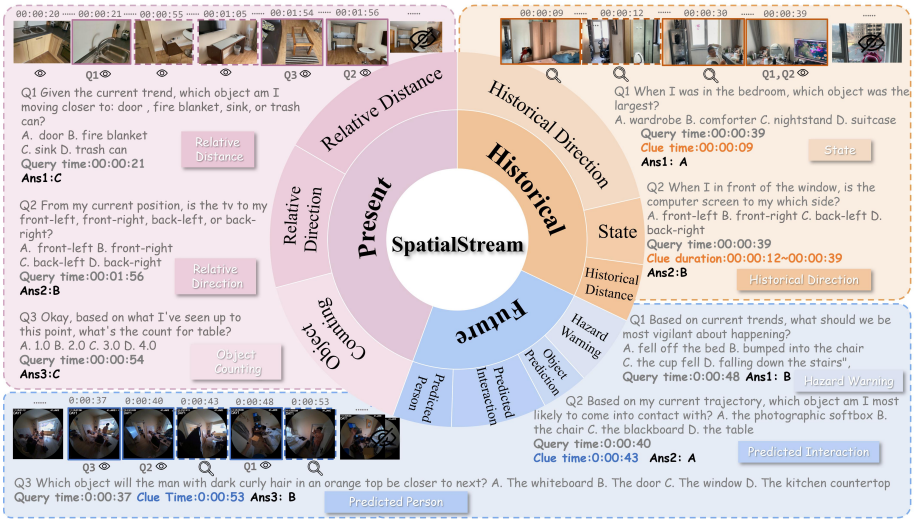

Fig. 1: Diagram of StreamSpatial-Bench, including 3 major categories and 10 subcategories. Here, the *Query time* is defined as the time when a query is raised in streaming scenarios, and *Clue time* denotes the time of the visual clues involved in answering the query.

- **Long-term 3D Spatial Memory and Reasoning**: ($P \rightarrow P$) Requires the maintenance of persistent allocentric and egocentric spatial representations over extended temporal spans and region spans. In our benchmark tasks, this includes: (a) Historical State: Properties of objects seen in the past (b) Historical Direction: Spatial relationships between objects from past observations (c) Historical Distance: Proximity of objects to the observer at previous timestamp.
- **Predictive World Modeling** ($P \cup C \rightarrow F$): Projects the evolution of a dynamic scene to anticipate future spatial states and potential hazards. In our benchmark tasks, this includes: (a) Predicted Interaction: Predicting what objects the observer will interact with next. (b) Object Prediction: Predicting what will appear in the observer's view next. (c) Hazard Warning: Any hazard warnings in the observer's current view. (d) Predicted Person: Predicting whom will the observer meet with.
- **Real-time Spatial Reasoning** ($C \rightarrow C \cup P; C$):Requires online parsing and dynamic assimilation of continuously input spatial-temporal streaming data, achieving real-time, incremental spatial reasoning that is immune to historical layout interference. In our benchmark tasks, this includes: (a) Relative Direction: The direction of objects relative to the observer in the current time. (b) Relative Distance: The distance of objects relative to the observer in the current time. (c) Object Counting: How many of a certain type of object the target observer has seen.

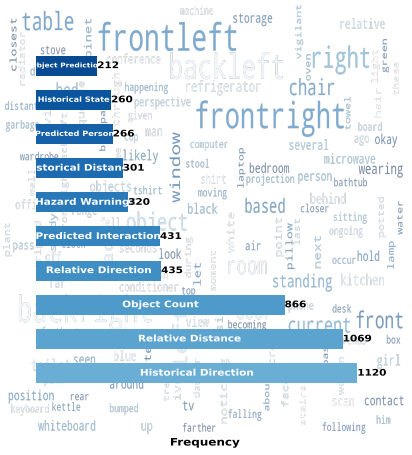

Fig. 2: The name and number of questions for each subcategory, along with word clouds.

Category	Count	Percentage
Video Data		
Total Videos	2003	–
Egolife Videos	671	33.50%
Other Videos	1332	66.50%
Question Data		
Total Questions	5280	–
Avg. Questions per Video	2.64	–
View Perspective		
Egocentric	4474	(84.73)%
Allocentric	806	(15.27)%
Spatial Complexity		
Spanning < 4 Regions	4817	91.23%
Spanning ≥ 4 Regions	463	8.77%
Temporal Complexity		
Spanning < 4 Span	4254	80.57%
Spanning ≥ 4 Span	1026	19.43%

Table 2: Core Statistics of the Streaming Spatial Intelligence Benchmark

– **Multi-View Spatial Co-Registration**: Establishes the ego-centric frame of reference within a global context by resolving self-location and orienting object-to-object relationships across first- and third-person viewpoints.

– **Multi-Agent Spatial Relationship Reasoning:** Maintains robust spatial awareness and reasoning integrity by filtering, disambiguating, and tracking salient entities within a fluid, multi-agent environment. In our benchmark tasks, this includes: Predicted Person: Predicting what kind of person the observer will interact with next

3.2 Benchmark Collection Process

Multi-Source Video Aggregation We curate a comprehensive video corpus by aggregating diverse egocentric datasets, including ScanNet [4], ScanNetV2, and related benchmarks, to ensure foundational coverage of structured indoor environments. To address the limitations of static or short-duration clips in existing benchmarks, we specifically incorporate large-scale, dynamic datasets that capture the complexity of real-world agent movement: ① We leverage Rynnec [5] to introduce substantial environmental variability, spanning 200+ residential structures and 20,832 egocentric videos. Crucially, we select clips featuring single-/dual-/tri-zone trajectories and cross-zone paths, ensuring the benchmark evaluates models under varied lighting conditions, camera heights, and rotational dynamics. ② To capture naturalistic human activities with rich spatial-temporal dependencies, we integrate EgoLife [24], which is pivotal for evaluating long-horizon, multi-person, and multi-view interactions. This source contributes diverse indoor and semi-outdoor scenes, continuous recordings that

require sustained attention, realistic AR-style egocentric capture from smart glasses worn by real users, and complex social interactions involving multiple agents. From this aggregated pool, we manually filter scenes to ensure each sample contains clear spatial events, sufficient multi-view overlap, and distinct object interactions, resulting in a high-quality subset for benchmark construction.

Millisecond-Level Temporal Grounding Annotation. Unlike prior benchmarks that rely on Large Language Models for approximate timestamp generation-often resulting in coarse granularity and temporal drift-we prioritize human-expert precision to establish a reliable ground truth. We established a rigorous, multi-stage manual annotation pipeline: ① Annotators undergo rigorous specialized training to standardize the definition of spatial events and temporal boundaries, ensuring a unified interpretation of complex scenarios. ② Experts meticulously annotate both the *query* occurrence timestamps and the *answer* temporal ranges (start/end), achieving millisecond-level granularity through precise temporal alignment with visual evidence. ③ A strict cross-verification protocol is implemented to check the temporal alignment and logical consistency between the annotated queries and their corresponding answer spans, with any discrepancies exceeding the threshold routed for senior arbitration. This human-centric approach ensures that our benchmark provides a high-fidelity signal for evaluating fine-grained temporal reasoning.

Confusing Distractor Generation Strategy. We systematically generated plausible distractors including spatially adjacent objects, temporally shifted answers, and semantically related but spatially incorrect alternatives.

Statics. StreamSpatialBench contains 5,280 question–answer pairs constructed from 2,003 video, with an average of 2.64 questions per video as illustrated in Table 2. As illustrated in Figure 2, the benchmark comprises 10 sub-tasks. Notably, 8.77% of the questions involve spatial reasoning across four or more regions, and 19.43% require inference over extended temporal contexts; these subsets represent the higher-complexity scenarios within the benchmark.

3.3 Stream 3D-Mem Framework

Stream3D-Mem proposes two core designs to enable streaming capabilities: (1) a dynamic memory system that ingests continuous visuals via a transient buffer, extracts distinctive frames as spatial engrams, and merges similar ones through learned consolidation for persistent 3D spatial reasoning with constant memory footprint. (2) an implicit reasoning paradigm that replaces the conventional process of generating multiple explicit cognitive map tokens prior to answer generation with a streamlined approach that utilizes concise implicit tokens to directly derive answers, effectively enhancing spatial reasoning accuracy while reducing inference latency.

Spatial Sub-space Formation To construct a dynamic memory from continuous video streams, we segment inputs into semantically coherent units. For each frame $\mathbf{I}_t$, we fuse 2D identity tokens $\mathbf{T}_t^{id}$ ("what") and 3D spatial tokens $\mathbf{T}_t^{3d}$ ("where") from VGGT:

$$\mathbf{X}_t = \mathbf{T}_t^{id} + \mathbf{T}_t^{3d} \in \mathbb{R}^{1\times HW\times C}. \tag{1}$$

Fused tokens are stored in a fixed-capacity FIFO buffer $\mathcal{B}$. Upon saturation, semantic transitions are detected via adaptive thresholding. We compute temporal smoothed similarity $\bar{s}_t = \frac{1}{w}\sum_{i=0}^{w-1}\cos(\mathbf{X}_{t-i}, \mathbf{X}_{t-i+1})$ and identify a transition when $\bar{s}_t < \tau$, where $\tau = \mu_s - \beta\cdot\sigma_s$ adapts to the buffer's similarity distribution. Frames preceding a transition form a *visual episode*, which is compressed into a *Spatial Prototype* $\mathbf{M}_{proto} \in \mathbb{R}^{N\times P\times D}$ and added to the long-term memory $\mathcal{L}$.

Incremental Construction of Long-Term Memory To manage unbounded spatial experiences within limited capacity, we employ a consolidation mechanism inspired by map hierarchies. When $\mathcal{L}$ reaches capacity, a spatial region fusion process is triggered. First, existing N sub-regions are pooled into global feature vectors $\mathbf{f}_i \in \mathbb{R}^D$ via average pooling over the positional dimension P:

$$\mathbf{f}_i = \frac{1}{P}\sum_{p=1}^{P}\mathbf{M}_{proto}^{(i,p,:)}. \tag{2}$$

These vectors are partitioned into K spatial regions (e.g., functional zones) using K-means. We then adopt an **intra-cluster similarity-first** fusion strategy: the largest cluster is selected to identify the most redundant sub-region pair $(\mathbf{M}_i, \mathbf{M}_j)$ based on cosine similarity:

$$\mathrm{sim}(\mathbf{M}_i, \mathbf{M}_j) = \frac{\mathbf{f}_i\cdot\mathbf{f}_j}{\|\mathbf{f}_i\|\|\mathbf{f}_j\|}. \tag{3}$$

The selected pair is merged using Token Merging (ToMe) [1], which reduces redundancy while preserving semantic integrity. The original nodes are replaced by the fused node, enabling continuous integration of spatial experiences without increasing memory burden.

3.4 Latent Cognitive Map Reasoning

Map Encoding & Compression The cognitive map is serialized into text and encoded via the LLM embedding layer into E_{map}. E_{map} is compressed into N vectors $V_{\text{gt}} \in \mathbb{R}^{N\times D}$ using uniform pooling, serving as fixed-length Ground Truth targets.

Latent Reasoning & Alignment Upon triggering the special token `<|map_start|>`, the model enters the latent reasoning mode. Unlike standard text generation, the model autoregressively generates N implicit hidden states $H_{\text{latent}} =$

$\{h_1, h_2, ..., h_N\}$ without emitting discrete text tokens. To ensure these latent states encode accurate spatial semantics, they are optimized to reconstruct the ground truth map vectors V_{gt}. Specifically, an alignment projector $\phi(\cdot)$ (e.g., an MLP) maps H_{latent} to the dimension of V_{gt} if necessary. The reconstruction is supervised via a Mean Squared Error (MSE) loss:

$$\mathcal{L}_{\text{latent}} = \frac{1}{N} \sum_{t=1}^{N} \|\phi(h_t) - v_t\|_2^2, \tag{4}$$

where $v_t \in V_{\text{gt}}$ represents the target spatial features. This loss is combined with the standard Next-Token Prediction (NTP) loss $\mathcal{L}_{\text{NTP}}$ to form a joint objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{NTP}} + \lambda \cdot \mathcal{L}_{\text{latent}}, \tag{5}$$

where λ balances the semantic reconstruction and language generation signals.

Transition to Text Generation Upon generating the special trigger token (e.g., `<|map_start|>`), the model enters the latent reasoning mode. In this mode, the model autoregressively generates N hidden states without predicting any discrete text tokens. After completing the preset N steps, the model exits the latent mode and passes the final hidden state h_N generated at the N-th step as context input to the text decoder to generate the first answer token.

From the perspective of probabilistic modeling, the generation condition of subsequent text y_t can be expressed as:

$$p(y_t \mid y_{<t}, h_{1:N}), \tag{6}$$

That is, the text generation process is conditioned on the hidden state sequence $h_{1:N}$ produced during the latent reasoning phase as a conditional prefix. By producing significantly fewer latent tokens than explicit "CoT" tokens (cognitive map), the method substantially reduces response latency.

Inference During inference, fixed-step decoding is adopted, executing exactly N implicit steps before resuming text generation. It provides a stable computational budget and ensures consistent reconstruction quality across diverse queries.

4 Can MLLMs Reason 3D Space-Time in Dynamic Long-Horizon Streams?

4.1 Evaluation Setup

Benchmark Models. We conduct extensive experiments across a diverse range of multimodal models with varying architectures and capabilities. Our evaluation includes: (1) Proprietary models: Gemini-2.5-pro and Doubao-Seed-1.6; (2) Open-source models from the DeepSeek series (DeepSeek-VL2-Small and

Methods	Size	Avg.	Rel. Dist.	Rel. Dir.	Obj. Count	Pred. Person	Pred. Interaction	Obj Pred.	Hazard Warning	Historical Dist.	Historical Dir.	Historical State
			Present Question			Future Question				Historical Question		
Perormance Against Human(250 Q&As)												
Human (250 Q&As)												
Human Level	–	87.08	88.50	86.30	95.70	78.65	82.66	97.25	97.66	84.39	77.35	82.31
Gemini-2.5-pro	–	43.23	66.12	36.87	43.72	19.62	31.27	23.00	72.52	31.12	34.17	73.84
Qwen3-VL-235B-A22B	235B	40.85	57.24	21.54	32.13	22.68	36.27	26.89	61.37	52.17	23.34	74.88
Uniformly Sampled, fps=2												
Proprietary MLLMs												
GPT-4o	–	38.24	55.42	31.11	29.40	23.02	39.85	30.19	58.70	44.13	26.05	60.95
Gemini-2.5-pro	–	44.92	65.94	37.24	39.03	19.32	39.95	27.96	73.44	44.19	31.75	69.23
Doubou-Seed-1.6	–	42.28	53.70	27.27	44.74	20.31	39.81	34.12	71.87	35.00	23.09	89.00
Open-Source Vdieo MLLMs												
DeepSeek-VL2-Small	2.8B	24.92	33.13	25.06	8.78	24.62	20.79	18.40	37.81	25.58	21.15	33.85
DeepSeek-VL2	4.5B	18.81	32.40	29.20	1.04	10.61	17.06	13.68	25.62	13.29	21.77	23.46
InternVL3.5-8B	8B	37.01	48.52	34.48	38.67	25.76	23.36	21.70	69.50	30.57	24.46	53.08
Qwen2.5-VL-7B	7B	34.49	36.00	31.72	31.00	27.65	18.69	23.02	58.31	29.90	35.29	53.28
Qwen2.5-VL-72B	72B	38.41	49.20	50.36	31.41	25.38	23.36	26.89	64.37	30.57	25.00	57.58
Qwen3-VL-8B	8B	42.56	62.23	35.17	30.48	26.89	32.94	29.72	76.56	52.16	29.39	67.69
Qwen3-VL-32B	32B	42.89	64.09	45.52	21.25	24.24	31.78	26.42	77.19	57.81	32.89	65.77
Qwen3-VL-235B-A22B	235B	43.24	60.78	27.82	24.36	25.00	38.08	33.96	75.00	59.47	29.87	73.85
LLaVA-Video-Qwen2-7B	7B	38.45	46.34	32.18	31.29	28.03	28.74	31.60	59.69	38.21	29.21	58.85
LLaVA-Video-Qwen2-72B	72B	44.01	59.96	34.48	32.33	29.92	37.62	36.32	62.19	42.86	31.36	73.08
Spatial-MLLM	7B	33.28	35.20	28.80	44.46	35.61	20.47	29.25	44.62	30.33	28.28	35.77
Streaming video input at 1 fps												
Open-Source Streaming MLLMs												
Vispeak	8B	36.63	45.07	27.59	45.84	26.52	36.79	25.94	48.75	31.23	28.39	50.19
Dispider	8B	30.63	27.61	17.24	38.45	36.74	24.30	24.06	47.50	27.57	21.52	41.31
Flash-Vstream	7B	33.55	40.66	27.59	37.88	29.55	25.93	32.55	48.12	31.89	25.54	35.77
Videollm-Online	8B	15.15	14.14	17.01	8.89	11.36	24.30	11.79	14.37	17.28	15.00	17.31
Ours	4B	41.11	45.19	34.48	53.35	28.41	32.39	31.60	62.66	32.89	31.31	58.85

Table 3: Evaluation results for MLLMs. We consolidate performance from both closed-source and open-source MLLM evaluations.

DeepSeek-VL2), Qwen series (Qwen2.5-VL-72B, Qwen2.5-VL-7B and Qwen3-VL-8B), and InternVL series (InternVL3.5-8B and InternVL3.5-38B); (3) Video understanding models: LLaVA-Video-Qwen2-7B and LLaVA-Video-Qwen2-72B; (4) Spatial-specialized model: Spatial-MLLM; and (5) Streaming-capable models: Vispeak, Flash-VStream, Dispider, and Ours. This comprehensive selection allows us to systematically analyze performance across different model paradigms in streaming spatial intelligence reasoning tasks.

Evaluation Settings. To evaluate models that do not yet support streaming video input, we adopt two complementary evaluation settings:

- **Sliding Window Setting (Offline):** Following [9], we perform offline sliding-window evaluation on advanced MLLMs by taking a fixed-length temporal window before each question-asking timestamp t_i (e.g., 32 seconds) and sampling frames at a fixed rate (e.g., 2 fps). This setting provides temporal context for response evaluation while adapting streaming video inputs to non-streaming Video-LLMs.
- **Streaming Setting:** All frames sampled at 1 fps from the start of the clip up to the question timestamp ($0\,\mathrm{s} \rightarrow \mathit{query\_time}$) are provided to the model, simulating real-time inference performance.

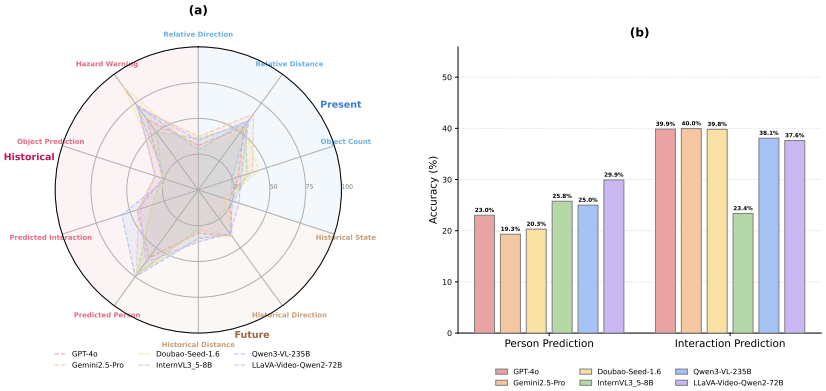

Fig. 3: (a) comparing the fine-grained accuracy of six multimodal models across ten sub-tasks grouped. (b) highlighting model performance on dynamic sub-tasks.

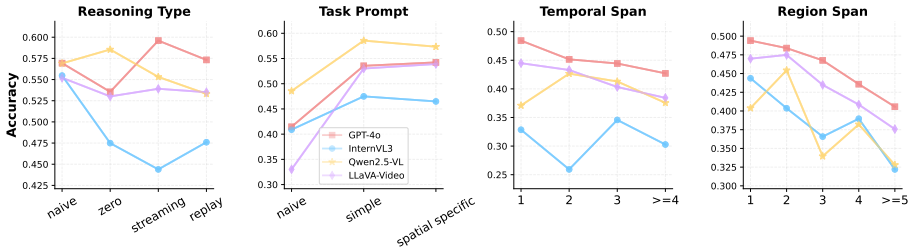

Fig. 4: Accuracy of four multimodal models across different reasoning types, task prompts, temporal spans, and region spans. Temporal span indicates the elapsed time from the video start to the query timestamp (e.g., 1: 0–15s, 2: 15–45s, 3: 45s–2min, $\geq$4: over 2min), and region span denotes the number of distinct spatial regions involved in answering the question.

Human Evaluation. We constructed a stratified random sample of 250 questions from our Bench, ensuring balanced coverage across all 10 task categories (25 questions per category). To obtain reliable human judgments, we recruited five trained annotators and labeled this subset under a blinded protocol: each question was independently answered by three annotators, with non-majority cases resolved through senior arbitration. For comparability, Gemini-2.5-Pro and Qwen3-VL-72B were evaluated on the same subset under identical inference settings.

4.2 Main Results on StreamSpatialBench

Human Performance Baseline. Humans demonstrate superior performance overall, particularly in incrementally updating real-time 3D spatial representations. However, maintaining long-term 3D spatial memory under streaming conditions remains a fundamental challenge.

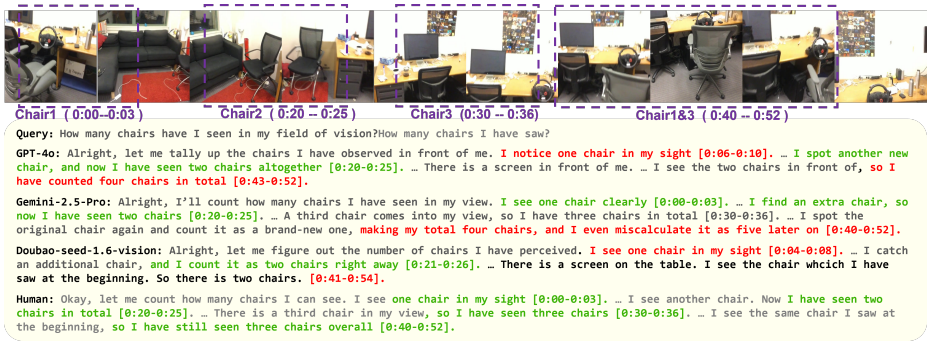

Fig. 5: Case study of models counting object with time, models often misplace objects in time or lose track of their identities, leading to double-counting or missed sightings.

Offline MLLMs with Uniform Temporal Sampling. We evaluate both spatial-specialized and general-purpose offline MLLMs adapted to streaming scenarios through uniform temporal sampling. Spatial-specialized models such as Spatial-MLLM-7B achieve 33.28% accuracy, while the general-purpose InternVL-3.5-8B reaches 37.01%. Notably, general-purpose models show competitive, and in some cases slightly stronger, performance than spatially specialized models. This suggests that diverse pretraining data may provide broader visual-semantic representations that are beneficial for complex streaming spatial tasks.

Streaming Models Comparison. As shown in Table 3, existing 7B–8B streaming models achieve average spatial accuracies ranging from only 15.15% to 36.63%, with substantial variation across subtasks. Their performance drops sharply on geometry-aware and relation-intensive tasks, often falling below 20%. This indicates that conventional streaming paradigms, without explicit spatial supervision, still struggle to develop robust spatial understanding.

Impact of Model Scale. Larger MLLMs generally demonstrate stronger spatial reasoning ability, with clear scaling trends within model families such as Qwen2.5-VL and LLaVA-Video. Nevertheless, some open-source models can rival or even surpass closed-source counterparts on specific subtasks. For example, Qwen2.5-VL-72B outperforms Gemini-2.5-Pro on relative direction (50.36 vs. 37.24). These results suggest that targeted architectural designs and specialized training strategies can partially compensate for the scale advantages of proprietary models.

4.3 Why Do MLLMs Struggle with Streaming Spatial Understanding?

To systematically investigate the performance bottlenecks of MLLMs on streaming 3D spatial intelligence tasks, we conduct a comprehensive analysis through our proposed benchmark.

Fundamental Spatial Cognitive Analysis

3D Spatial Task Comprehension. We investigated whether performance bottlenecks stem from task misinterpretation rather than 3D spatial reasoning deficits (Fig. 4). Systematic prompt variations—including simplified descriptions and clarified criteria—yielded only marginal improvements (4%). This persistent gap confirms that limitations lie in core spatial cognition, not instruction understanding.

Linguistic Reasoning Injection. Despite CoT's success in general reasoning, our experiments on StreamspatialBench reveal its limited efficacy in spatial cognition. We evaluated three strategies: Zero-CoT, Streaming Checkpoint-Based Spatial CoT, and Memory-Replay Spatial CoT. As shown in Fig. 4, these yielded only marginal improvements over baselines, failing to bridge the human-model performance gap. This suggests that the implicit 3D spatial cognition and cross-view understanding required for spatial supersensing represent fundamental capability gaps unaddressable by prompting alone.

Streaming Spatial Cognitive Analysis

Finding 1: *Models exhibit systematic performance degradation in recalling 3D spatial information as temporal horizon extends from query time.*

To rigorously evaluate long-term 3D spatial memory capabilities, we stratified history-related tasks based on two key dimensions: the number of traversed regions and the temporal span. Specifically, a higher count of traversed regions necessitates the retention of more complex 3D spatial layouts, while a longer temporal span demands sustained memory maintenance over extended durations. As illustrated in Figure 4, model performance exhibits a consistent decline as both metrics increase. This degradation pattern underscores a critical limitation: despite proficiency in short-term semantic understanding, existing MLLMs lack robust mechanisms for selective visual filtering and structured memory accumulation required for unbounded spatial video streams.

Finding 2: *Same question with time-varying answers remains intractable for MLLMs.*

As illustrated in Figure 5, MLLMs struggle severely with streaming scenarios where the same question has different correct answers at different timestamps—a

task intuitive for humans. While humans naturally maintain and update cumulative counts throughout video progression, current models fail to track such temporal states, revealing their dependence on learned statistical patterns rather than genuine temporal reasoning capabilities.

Finding 3: *MLLMs lack predictive world modeling capabilities essential for spatial supersensing.*

While humans effortlessly construct dynamic world models from continuous sensory streams to anticipate future states and potential risks, current MLLMs demonstrate fundamental deficiencies in predictive reasoning. As shown in Table 3, state-of-the-art models exhibit severely limited performance on future-oriented tasks, particularly those requiring trend extrapolation and risk forecasting based on the accumulated spatiotemporal context. This reveals that MLLMs fail to take advantage of observed 3D spatial-temporal patterns for forward inference or hazard anticipation from environmental cues. Rather than developing true predictive world modeling, models default to learned statistical priors from training distributions, employing reactive pattern recognition rather than proactive temporal reasoning.

Dynamic Interference on Spatial Cognition. Dynamic entities (e.g., moving people) in scenes significantly impair models' spatial reasoning capabilities,as in Table 5.

4.4 Evalutation of Streaming3D-Mem

Table 4: Accuracy (%) on the HourVideo benchmark.

Model	Navigation	Perception	Reasoning	Summarization	Avg
Qwen2.5-VL-3B	32.14	35.10	31.60	50.57	34.12
Ours	37.33	41.39	35.32	47.30	38.20

Table 5: Average accuracy comparison.

Model	Avg
InternVL2-8B	34.60
LLaVA-NeXT-Video-72B	40.90
LLaVA-OneVision-72B	40.20
SPAR-8B	41.10
Ours	47.30

Main Results on StreamspatialBench. The notable effectiveness of our framework in Table 3 is attributed to that can selectively retain recent frame tokens while pruning redundant historical ones. Moreover, a "latent token-reason-answer" paradigm that distills cognitive map reasoning into compact implicit representations.

Results on Offline 3D SpatialBench. We further evaluate our model on the offline 3D spatial intelligence benchmark VSI-Bench,as in table 5, where it maintains strong performance.

Results on General Video Bench. Additionally, on a long-form general video dataset, HourVideo [3], encompassing both 2D generic video QA and 3D spatial reasoning tasks, our approach achieves consistently competitive results across extended temporal horizons; as shown in Table 4, while a marginal gap persists on generic video questions, the enhancement of spatial understanding incurs only negligible compromise to general multimodal capabilities.

5 Conclusion

This work identifies critical gaps in MLLMs' streaming spatial reasoning through StreamSpatialBench, revealing deficiencies in spatiotemporal memory, incremental updating, and multi-agent adaptation. Looking forward, this research paves the way for truly embodied AI systems that seamlessly integrate real-time sensory streams with persistent spatial memory. Future developments could leverage neuromorphic architectures and predictive world models to achieve human-like spatial intelligence in robotics and AR/VR applications. Ultimately, advancing streaming spatial reasoning brings us closer to artificial general intelligence that perceives and reasons about dynamic 3D environments as naturally as humans.

Acknowledgements

This work is supported in part by the National Key R&D Program of China (Nos.2024YFB3908503 and 2024YFB3908500), in part by the National Natural Science Foundation of China (No. 62322608) and in part by the Shenzhen Loop Area Institute under grant FPF10120260001.

References

1. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your vit but faster. arXiv preprint arXiv:2210.09461 (2022)
2. Cai, M., Tan, R., Zhang, J., Zou, B., Zhang, K., Yao, F., Zhu, F., Gu, J., Zhong, Y., Shang, Y., et al.: Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models. arXiv preprint arXiv:2410.10818 (2024)
3. Chandrasegaran, K., Gupta, A., Hadzic, L.M., Kota, T., He, J., Eyzaguirre, C., Durante, Z., Li, M., Wu, J., Fei-Fei, L.: Hourvideo: 1-hour video-language understanding. Advances in Neural Information Processing Systems **37**, 53168–53197 (2024)
4. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
5. Dang, R., Yuan, Y., Mao, Y., Li, K., Liu, J., Wang, Z., Li, X., Wang, F., Zhao, D.: Rynnec: Bringing mllms into embodied world. arXiv preprint arXiv:2508.14160 (2025)
6. De Boer, J., Kommers, P.A., De Brock, B.: Using learning styles and viewing styles in streaming video. Computers & Education **56**(3), 727–735 (2011)

7. Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al.: Palm-e: An embodied multimodal language model. arXiv preprint arXiv:2303.03378 (2023)
8. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: Blink: Multimodal large language models can see but not perceive. In: European Conference on Computer Vision. pp. 148–166. Springer (2024)
9. Huang, Z., Li, X., Li, J., Wang, J., Zeng, X., Liang, C., Wu, T., Chen, X., Li, L., Wang, L.: Online video understanding: Ovbench and videochat-online. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3328–3338 (2025)
10. Jia, M., Qi, Z., Zhang, S., Zhang, W., Yu, X., He, J., Wang, H., Yi, L.: Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. arXiv preprint arXiv:2506.03135 (2025)
11. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al.: Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246 (2024)
12. Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., Deitke, M., Ehsani, K., Gordon, D., Zhu, Y., et al.: Ai2-thor: An interactive 3d environment for visual ai. arXiv preprint arXiv:1712.05474 (2017)
13. Lin, J., Zhu, C., Xu, R., Mao, X., Liu, X., Wang, T., Pang, J.: Ost-bench: Evaluating the capabilities of mllms in online spatio-temporal scene understanding. arXiv preprint arXiv:2507.07984 (2025)
14. Liu, F., Emerson, G., Collier, N.: Visual spatial reasoning. Transactions of the Association for Computational Linguistics **11**, 635–651 (2023)
15. Padmakumar, A., Thomason, J., Shrivastava, A., Lange, P., Narayan-Chen, A., Gella, S., Piramuthu, R., Tur, G., Hakkani-Tur, D.: Teach: Task-driven embodied agents that chat. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2017–2025 (2022)
16. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., et al.: Habitat: A platform for embodied ai research. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9339–9347 (2019)
17. Shi, X., Li, Y., Kou, Q., Yu, L., Xie, J., Zhou, H.: Spar: Scholar paper retrieval with llm-based agents for enhanced academic search. arXiv preprint arXiv:2507.15245 (2025)
18. Shridhar, M., Thomason, J., Gordon, D., Bisk, Y., Han, W., Mottaghi, R., Zettlemoyer, L., Fox, D.: Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10740–10749 (2020)
19. Wang, W., Tan, R., Zhu, P., Yang, J., Yang, Z., Wang, L., Kolobov, A., Gao, J., Gong, B.: Site: towards spatial intelligence thorough evaluation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9058–9069 (2025)
20. Wu, H., Li, D., Chen, B., Li, J.: Longvideobench: A benchmark for long-context interleaved video-language understanding. Advances in Neural Information Processing Systems **37**, 28828–28857 (2024)
21. Xia, F., Zamir, A.R., He, Z., Sax, A., Malik, J., Savarese, S.: Gibson env: Real-world perception for embodied agents. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9068–9079 (2018)

22. Xiao, J., Shang, X., Yao, A., Chua, T.S.: Next-qa: Next phase of question-answering to explaining temporal actions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9777–9786 (2021)
23. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in space: How multimodal large language models see, remember, and recall spaces. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10632–10643 (2025)
24. Yang, J., Liu, S., Guo, H., Dong, Y., Zhang, X., Zhang, S., Wang, P., Zhou, Z., Xie, B., Wang, Z., et al.: Egolife: Towards egocentric life assistant. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28885–28900 (2025)
25. Yang, R., Chen, H., Zhang, J., Zhao, M., Qian, C., Wang, K., Wang, Q., Koripella, T.V., Movahedi, M., Li, M., et al.: Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. arXiv preprint arXiv:2502.09560 (2025)
26. Yao, J., Wang, K., Hsieh, R., Zhou, H., Zou, T., Cheng, Z., Wang, Z., Viswanath, P.: Spin-bench: How well do llms plan strategically and reason socially? arXiv preprint arXiv:2503.12349 (2025)
27. Zhang, J., Chen, Y., Zhou, Y., Xu, Y., Huang, Z., Mei, J., Chen, J., Yuan, Y.J., Cai, X., Huang, G., et al.: From flatland to space: Teaching vision-language models to perceive and reason in 3d. arXiv preprint arXiv:2503.22976 (2025)
28. Zhao, F.: Application of streaming media on demand technology in the visual classroom teaching of college english. In: International conference on Big Data Analytics for Cyber-Physical-Systems. pp. 1626–1631. Springer (2019)
29. Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., et al.: Rt-2: Vision-language-action models transfer web knowledge to robotic control. In: Conference on Robot Learning. pp. 2165–2183. PMLR (2023)