LongSpace: Exploring Long-Horizon Spatial Memory from Perception to Recall in Video

Anonymous ACL submission

Abstract

Multimodal Large Language Models (MLLMs) have advanced image and video understanding and can increasingly handle longer visual inputs. Long-horizon tasks such as autonomous driving and robotic navigation require more than recognizing the current view, as models must remember and retrieve previously observed spatial layouts, routes, viewpoint changes, and object states. To evaluate this capability, we introduce LongSpace-Bench, a room-tour video benchmark for long-horizon spatial memory, covering scene perception, spatial relations, and spatial memory. In this work, we further propose LongSpace, a memory framework for long-video spatial reasoning. LongSpace models long videos as sequential chunks, incorporates 3D structural cues into early decoder layers, and constructs layer-aware memory for question-guided retrieval. Experiments on multiple spatial reasoning benchmarks show that LongSpace improves long-video spatial understanding, further demonstrating explicit spatial memory as a key capability for long-horizon video MLLMs.

1 Introduction

Recent MLLMs are extending visual understanding from static images to longer visual inputs (Zhang et al., 2024; Qian et al., 2024; Zhang et al., 2025; Chen et al., 2025). In continuous visual observations, spatial memory is a central capability. Models must not only recognize visible objects and events, but also maintain an understanding of scene layouts, object relationships, viewpoint changes, and navigable structures over time. In applications such as autonomous driving, robotic navigation, and embodied assistance, later decisions or questions often depend on spatial evidence observed much earlier. Recent studies have evaluated or improved spatial reasoning in multi-image, multi-view, and video settings (Yang et al., 2025c; Xu et al., 2025; Yeh et al., 2025; Lin et al., 2025), but most work still focuses on short-term contexts or local spatial relations.

As the observation horizon extends, spatial reasoning increasingly relies on long-horizon spatial memory, as illustrated in Figure 1. Unlike general temporal understanding, spatial evidence exhibits structural persistence, with layouts, depth cues, orientations, and route relations distributed across different segments that can influence later answers even when absent from the current visual context. Simply increasing the input length is insufficient, as redundant visual tokens may dilute important spatial cues, and unstructured segment-level information throughout the video is difficult to retrieve for subsequent reasoning. Effective long-horizon spatial reasoning thus benefits from models capable of extracting reliable spatial cues from local observations while maintaining these cues over extended temporal intervals for question-guided retrieval. In practice, long-horizon spatial memory represents the capacity to retain, organize, retrieve, and utilize spatial evidence across prolonged observations.

However, existing evaluations remain limited in capturing this capability. They often focus on short videos, multi-image inputs, or local relations (Li et al., 2025b; Zhu et al., 2026), and seldom address both **long-horizon** observations and **multi-dimensional** spatial abilities simultaneously. To address this gap, we introduce LongSpace-Bench, a room-tour video benchmark for long-horizon spatial memory. Constructed from real-world room-tour videos, LongSpace-Bench encompasses continuous indoor layouts, room transitions, object arrangements, and navigation routes, aligning with the requirement to preserve and retrieve long-range spatial evidence outlined above. Its tasks span three levels: scene perception, spatial relations, and spatial memory, including recognition of stable scene semantics, assessment of geometric relations such as distance and orientation, and memory-intensive

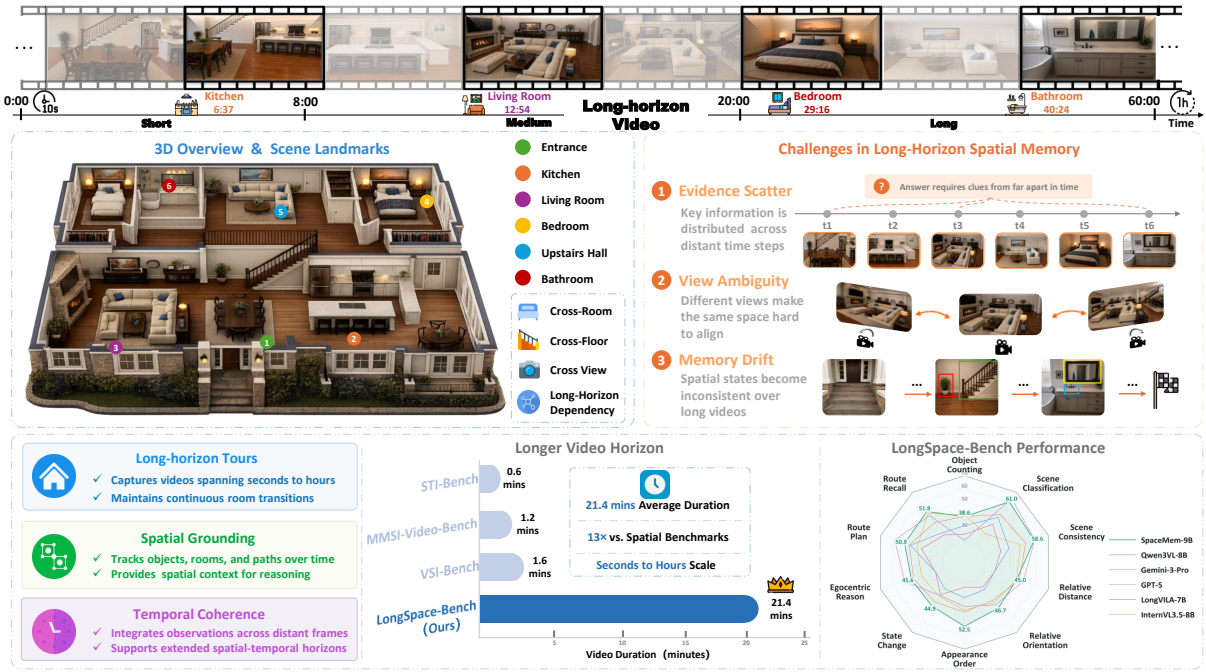


Figure 1: Long-horizon spatial memory require spatial evidence to be retained across distant observations, changing views, and evolving scene states. LongSpace-Bench spans video horizons from seconds to hours and evaluates spatial perception, relations, and memory over continuous room-tour videos.

reasoning over appearance order, state changes, route planning, and route recall. Collectively, these tasks enable LongSpace-Bench to evaluate whether models can retain, organize, retrieve, and utilize spatial information over extended temporal horizons.

To enable long-horizon spatial reasoning and memory, we further propose LongSpace. It directly addresses the modeling requirements outlined above by obtaining reliable spatial cues from local observations and preserving them for retrieval across segments. Prior studies indicate that geometry-enhanced models facilitate the capture of depth, orientation, and layout (Fan et al., 2025; Zheng et al., 2026; Zhao et al., 2025), while research on spatial memory highlights the importance of storing and reusing long-term scene information in a structured form (Yang et al., 2025d; Cai et al., 2025; Hu et al., 2026). Building on these insights, LongSpace integrates geometry-aware perception with retrievable memory. It represents a long video as an ordered sequence of chunks, aligns geometry features within each chunk to strengthen local spatial representations, and constructs retrievable layer-aware memory that maintains structured spatial evidence across chunks. During question answering, the model retrieves relevant evidence from memory to guide generation. LongSpace thus aims not only to increase the number of input frames, but to establish a queryable long-horizon spatial memory throughout the observation sequence. Experimental results demonstrate that LongSpace produces larger improvements on memory-intensive tasks, indicating that explicit spatial memory delivers advantages beyond stronger visual encoders or extended input contexts.

Our contributions are summarized as follows:

- We introduce **LongSpace-Bench**, a benchmark for evaluating long-video spatial reasoning and memory over real-world room-tour videos.
- We propose **LongSpace**, a framework that integrates geometry-aware perception with retrievable long-horizon video memory.
- We conduct a comprehensive evaluation on LongSpace-Bench across proprietary, open-source, and spatial-centric models, providing empirical analysis of long-video spatial memory.

2 Related Work

2.1 Visual Spatial Reasoning

Visual spatial reasoning studies whether a model can form a stable understanding of 3D scene structure, relative spatial relations, and viewpoint changes from images or videos. Recent work

Table 1: Comparison of LongSpace-Bench with representative spatial reasoning and memory benchmarks. MC, NA, and OE denote multiple-choice, numerical-answer, and open-ended formats. Avg. Duration denotes average video duration; – indicates unavailable or not applicable statistics. Multi-Dim. Eval. indicates whether a benchmark covers multiple spatial ability dimensions rather than a narrow relation type.

Benchmark	Modality	Scale			Benchmark Design			Long Video	Multi-Dim. Eval.
		Samples Num	QA Pairs	Avg. Duration	Annotation	Tasks	Answer		
SpatialRGPT (Cheng et al., 2024)	Image	1,406	1,406	–	Auto	2	OE+NA	✗	✗
SpatialVLM (Chen et al., 2024)	Image	546	546	–	Auto&Human	2	OE+NA	✗	✗
CVBench (Tong et al., 2024)	Image	2,638	2,638	–	Auto	2	MC	✗	✗
All-Angles-Bench (Yeh et al., 2025)	Image	380	2,132	–	Human	6	MC	✗	✗
MMSI-Bench (Yang et al., 2025c)	Image	1,990	1,000	–	Human	11	MC	✗	✓
SPAR-Bench (Zhang et al., 2026)	Image	14,708	7,207	–	Auto&Human	20	MC+NA/OE	✗	✓
VSI-Bench (Yang et al., 2024)	Video	288	5,000+	1.2mins	Auto&Human	8	MC+NA	✗	✓
STI-Bench (Li et al., 2025b)	Video	369	2,064	0.6mins	Auto&Human	8	MC	✗	✓
MMSI-Video-Bench (Lin et al., 2025)	Video	1,278	1,106	1.6mins	Human	13	MC	✗	✓
LongSpace-Bench (Ours)	**Video**	**445**	**4,073**	**21.4 min**	**Human**	**10**	**MC+NA**	✓	✓

moves beyond appearance-driven inference by introducing stronger spatial inductive bias into video MLLMs. Cambrian-1 (Tong et al., 2024) revisits vision encoder and connector design from a vision-centric perspective, proposing spatially aware aggregation to preserve high-resolution details important for spatial reasoning. Cambrian-S (Yang et al., 2025b) integrates persistent memory, implicit 3D cognition, and predictive sensing into a unified spatial framework. SpaceVista (Sun et al., 2025) extends visual spatial reasoning to all-scale scenarios from millimeters to kilometers, combining structured spatial knowledge, scale-aware modeling, and progressive training for better cross-scene spatial understanding. Another line of work enhances spatial reasoning with RGB-video geometric priors. VLM-3R (Fan et al., 2025) learns implicit 3D tokens through instruction-aligned 3D reconstruction, while VG-LLM introduces an explicit 3D geometry encoder to enrich visual representations with structural cues. SpaceMind (Zhao et al., 2025) improves geometry-language interaction through camera-guided modality fusion. Recent studies focus on making geometry more effective for reasoning. Spatial-R1 (Ouyang, 2025) emphasizes task-specific optimization for video spatial reasoning.

2.2 Memory-Enhanced Spatial Reasoning

Memory-enhanced spatial reasoning studies how a model preserves, updates, and retrieves scene-structured history during continuous observation. Unlike general video memory, this line of work emphasizes spatial consistency, structured organization, and direct support for downstream reasoning. 3D-Mem (Yang et al., 2025d) constructs incrementally updatable scene memory through memory and frontier snapshots, while MTU3D (Zhu et al., 2025b) maintains a dynamic spatial memory bank for grounding, scene representation, and exploration. 3DLLM-Mem (Hu et al., 2026) couples working memory with episodic memory for long-horizon embodied reasoning. OnlineSI (Liu et al., 2026) maintains a finite explicit spatial memory for online 3D understanding from video streams, and HIMM (Li et al., 2026) disentangles episodic and semantic memory for long-horizon exploration and question answering. 3DSPMR (Cai et al., 2025) further emphasize structured reuse of long-term spatial knowledge through spatialized memory retrieval or unified 3D memory from relational, visual, and geometric cues.

3 LongSpace-Bench

Spatial reasoning over extended visual observations requires more than recognizing isolated objects or local relations. A model may need to recall which room appeared earlier, how two areas are connected, whether an object state changed, or the path connecting different viewpoints. Existing spatial benchmarks offer useful evaluations for static scenes, multi-image reasoning, and short-video understanding, but they do not directly assess whether video MLLMs can preserve and utilize spatial evidence across long-horizon observations. We introduce LongSpace-Bench to address this limitation.

LongSpace-Bench is built for long-horizon spatial reasoning and memory. It evaluates the use of spatial evidence in continuous video observations and covers complementary abilities, including

scene perception, spatial relationships, and spatial memory. As shown in Table 1, existing image and multi-image benchmarks mainly assess static spatial understanding, which offers limited insight into spatial evidence that unfolds over time. Existing video benchmarks introduce dynamic inputs, but they often emphasize short clips, local relations, or narrower task scopes. As a result, they provide limited coverage of whether models can retain, retrieve, and combine spatial evidence over longer temporal spans. LongSpace-Bench targets this evaluation gap by testing the use of different types of spatial information across long-horizon observations.

3.1 Task Definition

LongSpace-Bench organizes long-video spatial ability into three levels: scene perception, spatial relationship, and spatial memory. Scene perception measures a model's understanding of global environments and stable scene semantics, including Object Counting, Scene Classification, and Scene Consistency. Spatial relationship focuses on geometric relations, such as Relative Distance and Relative Orientation, and requires models to infer spatial configurations between objects or regions under viewpoint changes. Spatial memory tests whether models can preserve and retrieve spatial evidence over long temporal horizons, covering Appearance Order, State Change, Egocentric Reasoning, Route Planning, and Route Recall.

3.2 Benchmark Statistic

LongSpace-Bench is built from real-world room-tour videos and contains 445 videos, approximately 159 hours of video, and 4,073 question-answer pairs. As shown in Figure 2, it covers short, medium-, and long-horizon videos and includes ten spatial task types. Object counting uses numerical answers, while the other tasks mainly use multiple-choice answers. Additional annotation details, dataset statistics, benchmark comparison, and evaluation protocol are provided in Appendix Sections A.1–A.4.

4 Method

4.1 Overview

Long-horizon spatial reasoning requires models to understand 3D spatial information and retrieve evidence across long temporal ranges. Language-aligned 2D visual features capture semantic appearance but do not explicitly model 3D features. Directly concatenating all video tokens into the language context can exceed the context budget and reduce computational efficiency. LongSpace addresses these issues with spatial geometry-aware perception and layer-wise KV memory. The perception module injects dense 3D features into decoder-side visual tokens, while the memory module compresses selected KV states into layer-wise memories as video segments are encoded temporally. During question answering, LongSpace retrieves query-relevant memories for the final response.

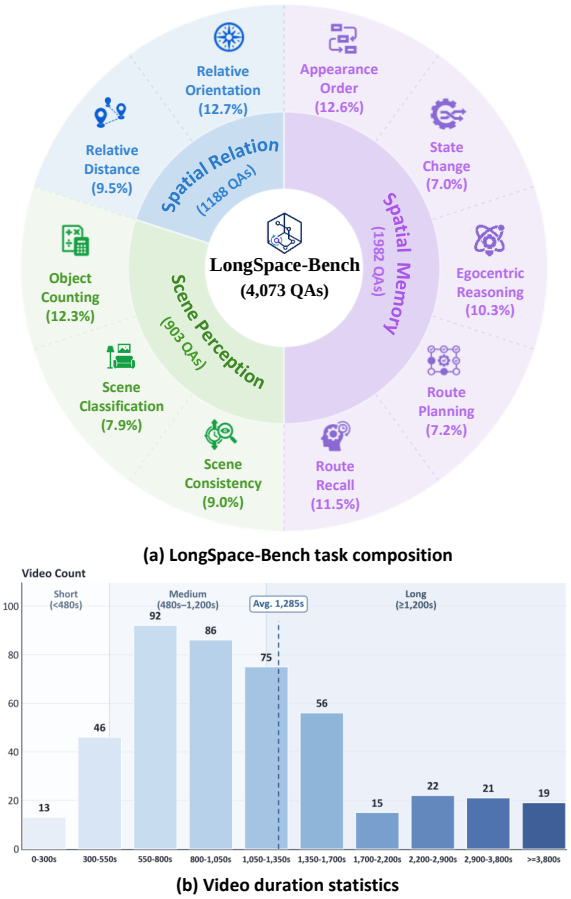


Figure 2: LongSpace-Bench Statistics Showing (a) Distribution of Question Types Across the Benchmark and (b) Distribution of Video Durations.

4.2 Spatial Structure Perception

As shown in Figure 3, SSP fuses Qwen3-VL 2D visual tokens with 3D spatial tokens from the 3D geometry encoder and merger on the decoder side. The 3D spatial features are first pooled, length-aligned, and normalized to match the number of visual tokens n_v:

$$\mathbf{G} = \mathrm{Align}\left(\psi_\theta(\Phi_{3\mathrm{d}}(\mathcal{X})), n_v\right). \quad (1)$$

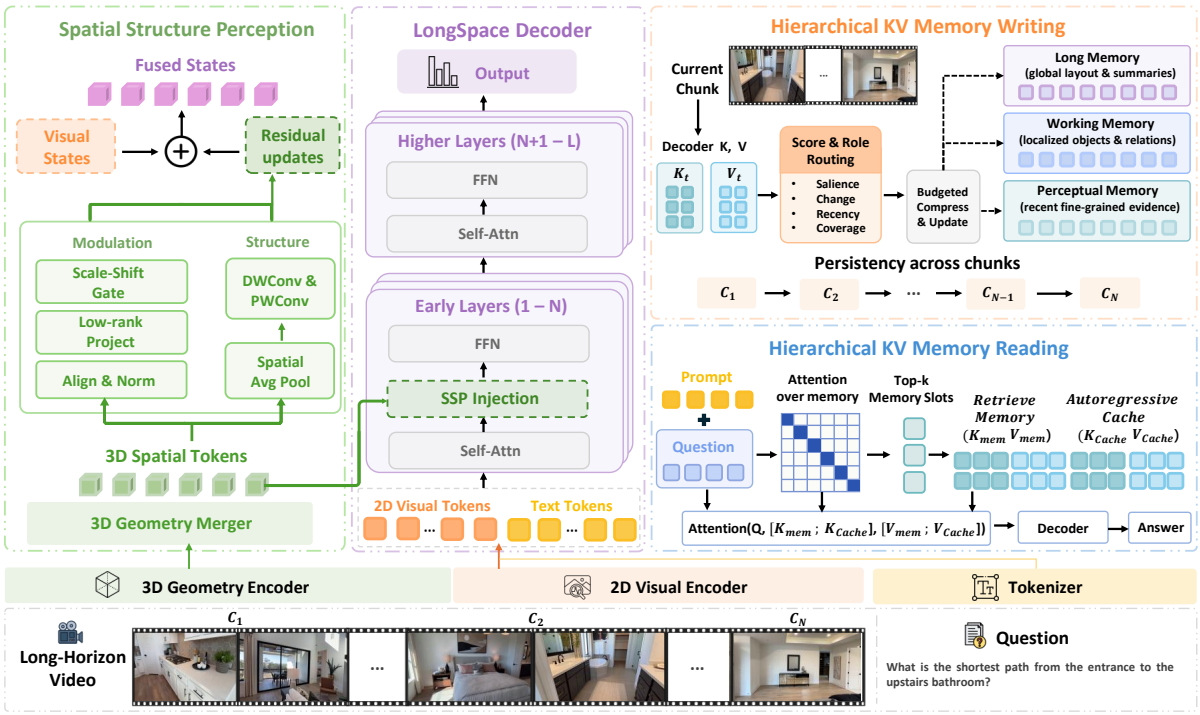


Figure 3: Overview of LongSpace. Spatial Structure Perception fuses 3D spatial tokens with 2D visual representations and injects them into the decoder. Hierarchical KV Memory organizes evidence from sequential video chunks into multi-level memories, which are retrieved according to the question for long-horizon spatial reasoning.

Here, $\mathbf{G}$ denotes the aligned 3D spatial tokens, and $\Phi_{3\mathrm{d}}$ is instantiated with π^3 (Wang et al., 2025b). SSP maintains $\mathbf{G}$ as an independent spatial-structure stream, rather than merging it only once into the input embeddings.

SSP gathers the current visual states using the visual mask. The visual states and 3D spatial tokens are then normalized and projected into a low-rank bottleneck, which produces a scale term γ, an offset term β, and a fusion gate that controls the strength of 3D spatial updates for each visual token. In parallel, SSP extracts a lightweight structural residual by average-pooling the spatial feature map and applying depthwise and pointwise convolutions.

The two paths are then combined as a residual update and written back only to visual-token positions:

$$\mathbf{H}_l^{\Omega} \leftarrow \mathbf{H}_l^{\Omega} + \Delta_l(\mathbf{H}_l^{\Omega}, \mathbf{G}, \mathbf{q}_l). \qquad (2)$$

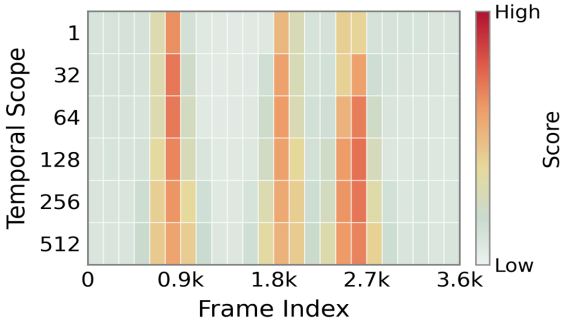
Here, $\mathbf{H}_l^{\Omega}$ denotes the hidden states at visual-token positions in layer l, $\mathbf{q}_l$ denotes the query context from text states, and Δ_l denotes the bounded residual from the modulation and structure branches. The updated visual states correspond to the Fused States in the figure and are passed to subsequent layers.

4.3 Hierarchical KV Memory

Hierarchical KV Memory (HKM) serves as the inference-time memory substrate of LongSpace for preserving long-video evidence under a bounded context. Instead of expanding the final prompt with all video tokens, LongSpace encodes temporally ordered video segments and materializes their evidence as compact hierarchical KV states inside decoder layers. This design enables the final question-answering stage to access cross-segment evidence without keeping the full visual sequence in the language context.

At each decoder layer, HKM turns the current attention states into reusable memory entries, including key-value states, position indices, and hidden features for later selection and compression. These layers capture evidence at different scopes. Sensory layers keep fine-grained visual and spatial evidence, while working layers provide a short-term workspace that binds objects, local spatial relations, and recent changes within the current segment into contextual states. Long-memory layers distill each segment into stable temporal anchors and scene cues, so the final question can locate relevant evidence across long temporal gaps.

For layer l and segment t, HKM represents memory update as role-conditioned evidence selection and budget-constrained compression. It first se-



Question: What is the sequence in which the following recreational items are shown?
A. The pool table, the poker table, the putting green, B. The pool table, the putting green, the poker table, C. The putting green, the pool table, the poker table, D. The poker table, the putting green, the pool table",
Answer: B

Figure 4: Visualization of LongSpace evidence localization. The heatmap shows how LongSpace identifies sparse question-relevant evidence within hour-level video frames.

lects candidate evidence from the current segment according to the layer role:

$$\mathcal{A}_{t,l} = \mathrm{Select}_{\rho(l)}(\mathbf{K}_{t,l}, \mathbf{V}_{t,l}, \mathbf{F}_{t,l}). \quad (3)$$

Here $\rho(l)$ denotes the layer role, $\mathbf{K}_{t,l}$ and $\mathbf{V}_{t,l}$ are the KV states produced by the current segment, and $\mathbf{F}_{t,l}$ is the hidden feature used for scoring.

It then merges the selected evidence with the previous memory and compresses it under the corresponding role-specific budget:

$$\mathcal{M}_{t,l} = \mathrm{Compress}_{\rho(l)}(\mathcal{M}_{t-1,l} \cup \mathcal{A}_{t,l}; B_{\rho(l)}). \quad (4)$$

Here $B_{\rho(l)}$ is the role-specific memory budget. Since each memory entry retains its position id, the compressed memory preserves the original temporal order of the video.

During memory update, Select treats the KV states, hidden feature, and position id at each video position as a candidate memory entry. It assigns each candidate a priority score using four normalized signals: feature norm for salience, adjacent feature difference for state change, uniformly sampled temporal anchors for coverage, and recency for recent evidence. The signals are combined with role-specific weights from $\rho(l)$. When the number of candidates exceeds the budget, Compress keeps recent or high-scoring entries in raw form and groups the remaining entries in temporal order. Within each group, score-softmax weights are used to pool KV states, features, and positions into compact segment entries. Long-range entries also store segment summaries and segment ids to support retrieval.

4.4 Memory Retrieval and Decoding

After encoding all video segments, LongSpace encodes the final question as a query and writes it to each HKM layer as the read condition. For sensory and working memories, it performs sparse top-k reading over stored positions. Candidate KV evidence is ranked by both query relevance and memory scores, and the corresponding local KV states are returned. For long memory, LongSpace adopts a segment-to-token coarse-to-fine read: it first matches the query with segment summaries or spatial prototypes to find relevant segments, and then reads compact KV entries from them. This process supplies the decoder with question-relevant hierarchical video evidence without expanding the full video memory. Figure 4 shows that LongSpace localizes sparse evidence among many irrelevant hour-level video frames.

During decoding, LongSpace keeps the attention operator unchanged and injects the retrieved HKM KV states as a frozen memory prefix. This prefix provides compressed video evidence to each layer, while only the autoregressive cache is updated as new tokens are generated.

5 Experiments

5.1 Implementation Details

Setting. LongSpace is built on Qwen3-VL-8B (Bai et al., 2025a) with π^3 (Wang et al., 2025b) as the 3D geometry encoder, and the geometry-aware module is inserted into the first eight decoder layers. We jointly optimize the language backbone, geometry module, multimodal projector, and language modeling head for one epoch with a global batch size of 64. We use AdamW with a learning rate of 1×10^{-5}, a cosine schedule, and a warmup ratio of 0.03. Each video contains at most 32 frames, and all experiments are conducted on 8 NVIDIA A100 80G GPUs.

Training Datasets. The training data consist of VSI-590K (Yang et al., 2025b), the instruction data introduced by VLM-3R (Fan et al., 2025), and a sampled subset of SPAR-7M (Zhang et al., 2026). These samples cover object-level perception, distance and direction estimation, route reasoning, and appearance-order reasoning.

5.2 Main Results

Standard Spatial Reasoning. We report the results on VSI-Bench (Yang et al., 2024) in Table 3.

Table 2: Quantitative comparison of performance on LongSpace-Bench. The best and second-best results among non-proprietary models are highlighted in **bold** and underlined, respectively.

Model	Scene Perception			Spatial Relation		Spatial Memory					Overall
	Obj.Cnt.	Scn.Cls.	Scn.Cons.	Rel.Dist.	Rel.Ori.	App.Ord.	St.Chg.	Ego.Reas.	Rt.Plan	Rt.Recall	
Proprietary Models											
GPT5 (Singh et al., 2025)	38.7	49.0	52.0	46.8	39.5	41.4	41.1	43.7	47.1	41.4	43.5
Gemini-3-Pro (Gemini, 2025)	20.3	64.0	59.2	49.9	43.6	37.2	34.7	50.8	58.7	48.0	45.3
Open-Source Models											
LongVA-7B (Zhang et al., 2024)	14.5	46.0	42.1	40.1	30.2	28.8	35.4	29.9	43.0	31.0	32.7
LongVILA-7B (Zhang et al., 2024)	25.0	39.5	32.1	27.1	26.7	22.4	34.0	26.4	37.5	29.5	29.1
LLaVA-OneVision-1.5-8B (Li et al., 2024)	22.3	48.8	44.2	40.1	31.6	36.8	32.3	34.4	49.5	39.5	37.0
LLaVA-NeXT-Video-7B (Liu et al., 2024)	9.0	39.2	30.8	26.6	30.0	31.7	28.1	33.0	37.5	38.0	29.9
LLaVA-NeXT-Video-72B (Liu et al., 2024)	8.8	37.9	36.4	33.6	29.5	28.0	40.4	30.6	41.6	32.4	30.5
Qwen2.5-VL-7B (Bai et al., 2025b)	26.4	54.2	48.6	38.0	31.2	33.9	38.2	28.7	48.1	33.1	36.7
Qwen2.5-VL-72B (Bai et al., 2025b)	33.1	59.7	57.3	44.7	**46.7**	42.8	44.2	45.4	50.2	46.7	44.2
Qwen3-VL-8B (Bai et al., 2025a)	31.5	46.3	38.6	27.1	28.9	30.5	28.4	26.4	46.4	31.6	32.9
Qwen3-VL-32B (Bai et al., 2025a)	36.5	55.0	54.5	43.4	37.8	41.2	42.1	39.0	48.5	44.8	46.5
InternVL3.5-8B (Wang et al., 2025a)	30.5	47.7	44.8	41.4	39.9	35.6	37.4	41.1	49.8	38.4	40.0
InternVL3.5-38B (Wang et al., 2025a)	31.8	56.9	**58.9**	43.7	39.0	38.9	42.8	39.2	**57.7**	46.5	44.2
Spatial-Centric Models											
Spatial-MLLM-4B (Wu et al., 2026)	29.9	36.5	34.0	34.6	31.2	25.5	30.2	32.1	37.9	27.3	31.4
VG-LLM-4B (Zheng et al., 2026)	23.8	45.0	37.7	37.7	33.7	31.7	38.2	34.2	47.8	39.2	36.0
VG-LLM-8B (Zheng et al., 2026)	37.3	55.3	43.6	38.0	38.2	29.6	42.5	37.1	54.9	38.6	39.2
VST-3B-SFT (Yang et al., 2025a)	16.7	45.2	34.9	36.2	38.6	33.9	31.6	39.7	44.4	33.9	34.9
VST-7B-SFT (Yang et al., 2025a)	16.1	48.0	42.1	36.7	36.8	29.8	38.9	32.8	47.1	34.1	35.0
VLM-3R-7B (Fan et al., 2025)	37.1	42.8	45.3	42.0	36.7	38.6	42.7	40.9	41.6	39.1	40.2
SpatialLadder-3B (Li et al., 2025a)	24.4	43.1	35.2	36.4	39.9	27.8	37.2	41.8	49.5	33.7	36.0
Cambrian-S-3B (Yang et al., 2025b)	**47.2**	44.4	45.8	37.7	33.9	32.5	35.8	30.4	44.4	32.2	37.9
Cambrian-S-7B (Yang et al., 2025b)	43.4	48.5	43.9	40.8	39.7	31.7	42.1	41.6	47.1	33.0	40.5
LongSpace-9B (Ours)	38.6	**61.0**	58.6	**45.0**	**46.7**	**52.5**	**44.9**	**45.6**	50.9	**51.8**	**49.2**

LongSpace achieves an average score of 70.8, outperforming InternVL3-78B (Wang et al., 2025a), Qwen3-VL-8B (Bai et al., 2025a), and Cambrian-S-7B by 22.4, 12.9, and 7.9 points, respectively. This supports the benefit of geometry-aware perception for local spatial estimation, while LongSpace-Bench tests whether such evidence can be preserved across long videos. Figure 4 further shows that LongSpace attends to question-relevant spatial regions across temporally separated frames.

Long-Horizon Spatial Memory. Table 2 compares models on LongSpace-Bench. LongSpace-9B achieves the highest overall score of 49.2, outperforming the strongest open-source baseline, Qwen3-VL-32B (Bai et al., 2025a), and the strongest proprietary baseline, Gemini-3-Pro (Comanici et al., 2025), by 2.7 and 3.9 points, respectively. The gap is larger over spatial-centric models, with gains of 8.7 points over Cambrian-S-7B (Yang et al., 2025b) and 9.0 points over VLM-3R-7B (Fan et al., 2025). At the task level, LongSpace-9B performs best on Appearance Order, State Change, and Route Recall, with scores of 52.5, 44.9, and 51.8, and ties for the best score on Relative Orientation. The gains, however, are not uniform across categories. Gemini-3-Pro remains stronger on Scene Classification, Scene Consistency, Relative Distance, Egocentric Reasoning, and Route Planning. These results suggest that explicit long-horizon memory is most useful when the task requires retaining and retrieving evidence from distant video segments, whereas some scene-level recognition and high-level planning questions still benefit from stronger proprietary models.

Table 4: Effect of the number of geometry-injection layers on benchmark performance. highlights the best result for each metric.

#Layers	VSI-Bench	CV-Bench	SPAR-Bench	Avg.
6	63.6	86.5	65.5	71.9
8	65.0	86.8	65.3	72.4
12	65.0	86.4	65.4	72.3
24	61.3	85.2	67.3	71.3
36	64.9	86.1	65.3	72.1

5.3 Ablation Studies

Moderate Geometry Injection. Table 4 examines how many decoder layers receive geometry injection. All variants are trained on the same 324K training subset. The best average score is achieved with 8 layers, reaching 72.4, while 12 layers gives a comparable score of 72.3. Adding geometry to more layers does not improve performance monotonically. Although using 24 layers yields the highest SPAR-Bench (Zhang et al., 2026) score of 67.3, its average score drops to 71.3 because the scores on VSI-Bench and CV-Bench (Tong et al., 2024) decrease to 61.3 and 85.2, respectively. This trend suggests that geometry injection is most effective when applied to a moderate number of early decoder layers.

Table 3: Main results on the VSI-Bench benchmark. Best and second-best scores are highlighted in **bold** and underlined, respectively. Numerical answers are evaluated by MRA and multiple-choice answers by accuracy.

Model	Numerical Answer				Multiple-choice Answer				Avg.
	Obj.Count	Abs.Dist.	Obj.Size	Room Size	Rel.Dist.	Rel.Dir.	Route Plan	Appr.Order	
Proprietary Models									
GPT-5 (Singh et al., 2025)	53.3	34.4	73.3	47.5	63.7	48.6	50.2	68.9	55.0
Gemini-2.5-Pro (Comanici et al., 2025)	46.0	37.4	68.7	54.4	62.0	43.9	47.4	68.8	53.6
Gemini-3-Pro (Gemini, 2025)	49.0	42.8	71.5	41.8	56.6	57.5	**61.9**	60.0	56.0
Open-Source Models									
LongVA-7B (Zhang et al., 2024)	38.0	16.6	38.9	22.2	33.1	43.3	25.4	15.7	29.2
VILA-1.5-8B (Lin et al., 2024)	17.4	21.8	50.3	18.8	32.1	34.8	31.0	24.8	28.9
VILA-1.5-40B (Lin et al., 2024)	22.4	24.8	48.7	22.7	40.5	25.7	31.5	32.9	31.2
LLaVA-OneVision-72B (Li et al., 2024)	43.5	23.9	57.6	37.5	42.5	39.9	32.5	44.6	40.2
LLaVA-NeXT-Video-72B (Liu et al., 2024)	48.9	22.8	57.4	35.3	42.4	36.7	35.0	48.6	40.9
Qwen2.5-VL-72B (Bai et al., 2025b)	25.1	29.3	54.5	38.8	38.2	37.0	34.0	28.9	37.0
Qwen3-VL-8B (Bai et al., 2025a)	67.5	47.0	76.3	61.9	58.0	50.9	35.0	66.3	57.9
InternVL3-8B (Zhu et al., 2025a)	68.1	39.0	48.4	33.6	48.3	36.4	27.3	35.4	42.1
InternVL3-78B (Zhu et al., 2025a)	71.2	53.7	44.4	39.5	55.9	39.5	28.9	54.5	48.4
Spatial-Centric Models									
Spatial-MLLM-4B (Wu et al., 2026)	65.3	34.8	63.1	45.1	41.3	46.2	33.5	46.3	48.4
VG-LLM-4B (Zheng et al., 2026)	66.0	37.8	55.2	59.2	44.6	45.6	33.5	36.4	47.3
VG-LLM-8B (Zheng et al., 2026)	67.9	37.7	58.6	62.0	46.6	40.7	32.4	59.2	50.7
VST-3B-SFT (Yang et al., 2025a)	69.3	45.4	71.8	62.4	59.0	46.0	38.7	70.2	57.9
VST-7B-SFT (Yang et al., 2025a)	72.0	44.4	74.3	68.3	59.7	55.8	44.9	65.2	60.6
Cambrian-S-3B (Yang et al., 2025b)	70.7	40.6	68.0	46.3	64.8	61.9	27.3	**78.8**	57.3
Cambrian-S-7B (Yang et al., 2025b)	68.2	45.8	72.5	67.6	66.8	69.6	39.2	73.8	62.9
VLM-3R-7B (Fan et al., 2025)	70.2	49.4	69.2	67.1	65.4	80.5	45.4	40.1	60.9
SpatialLadder-3B (Li et al., 2025a)	62.1	35.3	61.9	41.4	45.6	46.4	27.3	38.5	44.8
LongSpace-9B (Ours)	**73.8**	**57.6**	**77.4**	**74.4**	**72.4**	**84.3**	47.4	**78.8**	**70.8**

Table 5: Component-wise ablation of layer-aware memory organization, capacity allocation, and hierarchical memory roles.

Configuration	Scene Perception	Spatial Relation	Spatial Memory	Overall
Layer organization				
Layer-agnostic	38.5	42.7	43.1	41.8
Layer-aware	50.9	46.0	49.5	49.2
Memory budget				
Read cap. $\times 0.5$	36.7	35.7	35.6	35.9
Read cap. $\times 2$	37.0	35.4	35.2	35.8
Bank cap. $\times 0.5$	37.0	35.4	35.2	35.8
Hierarchical role design				
w/o working role	36.0	36.5	35.4	35.8
w/o long role	34.5	35.9	34.6	34.8

Long-Horizon Memory Organization. Figure 5 and Table 5 analyze long-memory inference and layer-aware organization on LongSpace-Bench. Uniform sampling with 32 frames scores 36.1, while recent-window inference gives a modest gain to 37.7. Long-memory inference reaches 49.2, outperforming the two baselines by 13.1 and 11.5 points. The gain grows with video length, improving short, medium, and long videos by 4.8, 12.8, and 15.1 points over uniform sampling. This suggests that preserving cross-chunk evidence matters more when relevant observations are separated by longer temporal intervals. We further examine whether these gains come from memory organization or memory size. The layer-aware design improves the overall score from 41.8 to 49.2, with the largest gain on scene perception. In comparison, changing the read or memory-bank capacity around the default setting keeps the score near 35.8–35.9. Removing hierarchical roles degrades performance, especially without the long-term role, reducing the score to 34.8. These results show that effective long-video memory relies more on layer- and role-specific organization than on memory capacity alone.

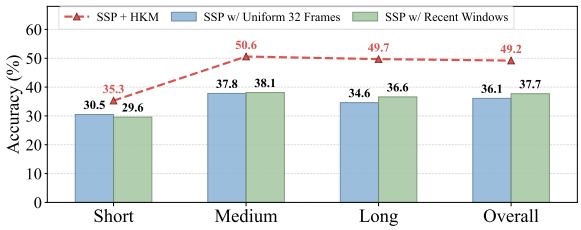


Figure 5: Comparison of different inference settings on LongSpace-Bench across video length levels.

6 Conclusion

We introduce LongSpace-Bench and LongSpace to study spatial reasoning and memory in long videos. LongSpace-Bench is a long-horizon spatial memory benchmark built from real-world room-tour videos, covering scene perception, spatial relations, and spatial memory. LongSpace combines spatial structure perception with hierarchical KV memory to preserve and retrieve question-relevant spatial evidence across video segments. Together, they provide a foundation for evaluating and improving spatial memory in video MLLMs.

Limitations

LongSpace-Bench is built primarily from indoor room-tour videos, which contain rich object layouts, room transitions, and long-range spatial dependencies but do not fully cover complex outdoor environments. And the proposed tasks focus on observation-based spatial memory, requiring the model to remember objects, regions, routes, and temporal order from passively observed videos, while active exploration, interactive navigation, and object manipulation remain outside the current scope.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Zhongyi Cai, Yi Du, Chen Wang, and Yu Kong. 2025. Vision to geometry: 3d spatial memory for sequential embodied mllm reasoning and exploration. *arXiv preprint arXiv:2512.02458*.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. *arXiv preprint arXiv:2401.12168*.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, and 1 others. 2025. Longvila: Scaling long-context visual language models for long videos. In *International Conference on Learning Representations*, volume 2025, pages 18227–18246.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. 2024. Spatialrgpt: Grounded spatial reasoning in vision language models. *arXiv preprint arXiv:2406.01584*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Shijie Zhou, Dilin Wang, and 1 others. 2025. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. *arXiv preprint arXiv:2505.20279*.

Gemini. 2025. Gemini 3 Pro Model Card. Technical report, Gemini. Accessed: 2025-11-18.

Wenbo Hu, Yining Hong, Yanjun Wang, Leison Gao, Zibu Wei, Xingcheng Yao, Nanyun Peng, Yonatan Bitton, Idan Szpektor, and Kai-Wei Chang. 2026. 3dllm-mem: Long-term spatial-temporal memory for embodied 3d large language model. *Advances in Neural Information Processing Systems*, 38:67856–67884.

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

Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. 2025a. Spatialladder: Progressive training for spatial reasoning in vision-language models. *arXiv preprint arXiv:2510.08531*.

Ji Li, Bo Wang, Jing Xia, Mingyi Li, and Shiyan Hu. 2026. Himm: Human-inspired long-term memory modeling for embodied exploration and question answering. *arXiv preprint arXiv:2602.15513*.

Yun Li, Yiming Zhang, Tao Lin, Xiangrui Liu, Wenxiao Cai, Zheng Liu, and Bo Zhao. 2025b. Sti-bench: Are mllms ready for precise spatial-temporal world understanding? *arXiv preprint arXiv:2503.23765*.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26689–26699.

Jingli Lin, Runsen Xu, Shaohao Zhu, Sihan Yang, Peizhou Cao, Yunlong Ran, Miao Hu, Chenming Zhu, Yiman Xie, Yilin Long, and 1 others. 2025. Mmsi-video-bench: A holistic benchmark for video-based spatial intelligence. *arXiv preprint arXiv:2512.10863*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Zixian Liu, Zhaoxi Chen, Liang Pan, and Ziwei Liu. 2026. Onlinesi: Taming large language model for online 3d understanding and grounding. *arXiv preprint arXiv:2601.16538*.

Kun Ouyang. 2025. Spatial-r1: Enhancing mllms in video spatial reasoning. *arXiv e-prints*, pages arXiv–2504.

Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. 2024. Streaming long video understanding with large language models. *Advances in Neural Information Processing Systems*, 37:119336–119360.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Peiwen Sun, Shiqiang Lang, Dongming Wu, Yi Ding, Kaituo Feng, Huadai Liu, Zhen Ye, Rui Liu, Yun-Hui Liu, Jianan Wang, and 1 others. 2025. Spacevista: All-scale visual spatial reasoning from mm to km. *arXiv preprint arXiv:2510.09606*.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, and 1 others. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025a. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. 2025b. π^3: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. 2026. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *Advances in Neural Information Processing Systems*, 38:13569–13597.

Runsen Xu, Weiyao Wang, Hao Tang, Xingyu Chen, Xiaodong Wang, Fu-Jen Chu, Dahua Lin, Matt Feiszli, and Kevin J. Liang. 2025. Multi-spatialmllm: Multi-frame spatial understanding with multi-modal large language models. *arXiv preprint arXiv:2505.17015*.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2024. Thinking in space: How multimodal large language models see, remember, and recall spaces. *arXiv preprint arXiv:2412.14171*.

Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, and 1 others. 2025a. Visual spatial tuning. *arXiv preprint arXiv:2511.05491*.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis L Brown II, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, and 1 others. 2025b. Cambrian-s: Towards spatial supersensing in video. In *The Fourteenth International Conference on Learning Representations*.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. 2025c. Mmsi-bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*.

Yuncong Yang, Han Yang, Jiachen Zhou, Peihao Chen, Hongxin Zhang, Yilun Du, and Chuang Gan. 2025d. 3d-mem: 3d scene memory for embodied exploration and reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17294–17303.

Chun-Hsiao Yeh, Chenyu Wang, Shengbang Tong, Ta-Ying Cheng, Ruoyu Wang, Tianzhe Chu, Yuexiang Zhai, Yubei Chen, Shenghua Gao, and Yi Ma. 2025. Seeing from another perspective: Evaluating multi-view understanding in mllms. *arXiv preprint arXiv:2504.15280*.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, and 1 others. 2025. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.

Jiahui Zhang, Yurui Chen, Yueming Xu, Ze Huang, Jilin Mei, Chunhui Chen, Yanpeng Zhou, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, and 1 others. 2026. From flatland to space: Teaching vision-language models to perceive and reason in 3d. *Advances in Neural Information Processing Systems*, 38.

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

Ruosen Zhao, Zhikang Zhang, Jialei Xu, Jiahao Chang, Dong Chen, Lingyun Li, Weijian Sun, and Zizhuang Wei. 2025. Spacemind: Camera-guided modality fusion for spatial reasoning in vision-language models. *arXiv preprint arXiv:2511.23075*.

Duo Zheng, Yanyang Li, Liwei Wang, and 1 others. 2026. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. *Advances in Neural Information Processing Systems*, 38:20560–20586.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025a. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Rui Zhu, Xin Shen, Shuchen Wu, Chenxi Miao, Xin Yu, Yang Li, Weikang Li, Deguo Xia, and Jizhou Huang.

2026. Video-msr: Benchmarking multi-hop spatial reasoning capabilities of mllms. *arXiv preprint arXiv:2601.09430*.

Ziyu Zhu, Xilin Wang, Yixuan Li, Zhuofan Zhang, Xiaojian Ma, Yixin Chen, Baoxiong Jia, Wei Liang, Qian Yu, Zhidong Deng, and 1 others. 2025b. Move to understand a 3d scene: Bridging visual grounding and exploration for efficient and versatile embodied navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8120–8132.

A Appendix

A.1 LongSpace-Bench Data

Table 6 summarizes the core data properties of LongSpace-Bench. The benchmark is built from Youtube real indoor room-tour videos and keeps clips that contain layout, object-position, viewpoint-change, and route evidence. Each question-answer sample is grounded in video evidence and is assigned to one of three ability groups: scene perception, spatial relationship, and spatial memory.

Table 6: Data card of LongSpace-Bench.

Item	Description
Video source	YouTube real-world indoor room-tour videos
Number of videos	445
Total duration	Approximately 159 hours
Number of QA pairs	4,073
Task categories	Scene perception, Spatial relationship, Spatial memory
Number of subtasks	10
Length subsets	Short: 280 QA; Medium: 2,102 QA; Long: 1,691 QA
Answer format	Numeric answers for counting; multiple choice otherwise
Evaluation metric	MRA for counting; accuracy for multiple choice

Table 7 further reports the sample count, answer format, and evidence scope for each task type. Object Counting uses numerical answers, while the remaining tasks mainly use multiple-choice answers. This mixed format separates fine-grained quantitative perception from categorical spatial reasoning and memory retrieval. The QA examples for each LongSpace-Bench subtask are presented in Figures 7–16.

Table 7: Task taxonomy and sample statistics of LongSpace-Bench. NA denotes numerical answer and MC denotes multiple choice.

Ability	Task	QA	Evidence focus
Scene perception	Object Counting	500 / NA	Countable objects or regions
	Scene Classification	367 / MC	Room or scene category
	Scene Consistency	321 / MC	Stable scene identity and layout
Spatial relationship	Relative Distance	387 / MC	Object or region distance
	Relative Orientation	516 / MC	Left/right/front/back relation
Spatial memory	Appearance Order	514 / MC	Temporal order of observations
	State Change	285 / MC	Updated scene or object state
	Egocentric Reasoning	421 / MC	Viewer-centered direction and location
	Route Planning	293 / MC	Path-level spatial decision
	Route Recall	469 / MC	Remembered navigation trajectory

A.2 Annotation and Filtering Protocol

Figure 6 shows the construction pipeline of LongSpace-Bench. We start from raw YouTube videos and clip them into relevant segments by annotating their start and end timestamps and removing portions unrelated to spatial reasoning. The retained clips are arranged as temporally ordered observation sequences, so that questions can refer to evidence across different rooms, viewpoints, and time steps.

We provide question templates and descriptions of common indoor objects and regions. Annotators write questions, candidate answers, and ground-truth answers according to the corresponding task type. During review, we check whether the supporting evidence can be located in the video, whether each answer is unique, and whether the referring expressions are clear. We remove samples if the answer cannot be uniquely determined from the video, the question requires external commonsense knowledge, the target object is heavily occluded, or the options conflict ambiguously with the video evidence.

A.3 Answer Format and Quality Control

Object Counting uses numerical answers and is evaluated with mean relative accuracy (MRA), which limits the impact of small counting deviations while still penalizing large errors. The remaining tasks mainly use multiple-choice answers and are evaluated by accuracy. These tasks cover relation judgment, state recall, route reasoning, and temporal-order reasoning.

Quality control aims to reduce shortcuts and ambiguity. For numerical questions, annotators check whether the target to be counted is clearly defined and visible in the video. For multiple-choice questions, distractors are written to be plausible in the room context but inconsistent with the video evidence. This design encourages models to retrieve the relevant spatial observation instead of relying on language priors or local static cues.

A.4 Evaluated Models and Input Protocol

Table 8 summarizes the input protocol used for each evaluated model. We follow the practical input constraints of each model interface and provide temporally distributed evidence whenever possible. For proprietary models, GPT-5 is evaluated with 50 uniformly sampled frames, while Gemini-3-Pro is evaluated with video input sampled at 1 fps. For open-source video MLLMs, the Qwen series is evaluated with 512 uniformly sampled frames over the full video. InternVL3.5, LLaVA-OneVision-1.5, LLaVA-NeXT-Video, LongVA, and LongVILA are evaluated with 64 uniformly sampled frames. For spatial-centric MLLMs, we follow the default configurations of each model: Spatial-MLLM-4B uses 16 frames, while VG-LLM, VST, VLM-3R, SpatialLadder, and Cambrian-S use 32 frames. LongSpace uses chunked memory inference, where videos are processed at 1 fps and divided into 32-frame chunks

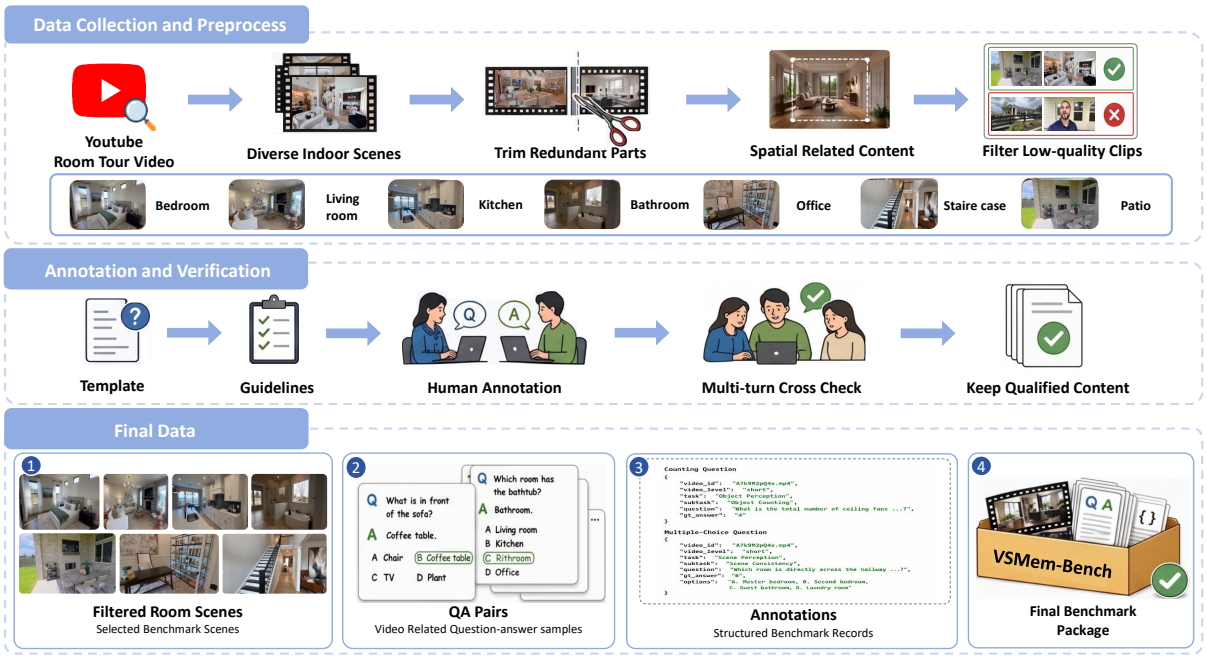


Figure 6: Construction pipeline of LongSpace-Bench. The pipeline collects room-tour videos, removes clips unrelated to spatial reasoning and memory, and produces verified question-answer pairs through taxonomy-guided annotation and manual review.

Table 8: Evaluated models and input protocols.

Model	Input protocol
GPT-5	50 uniformly sampled frames
Gemini-3-Pro	1 fps video input
Qwen series	512 uniformly sampled frames
InternVL3.5	64 uniformly sampled frames
LLaVA-OneVision-1.5	64 uniformly sampled frames
LLaVA-NeXT-Video	64 uniformly sampled frames
LongVA	64 uniformly sampled frames
LongVILA	64 uniformly sampled frames
Spatial-MLLM-4B (Wu et al., 2026)	16 frames under the default setting
VG-LLM-4B (Zheng et al., 2026)	32 frames under the default setting
VG-LLM-8B (Zheng et al., 2026)	32 frames under the default setting
VST-3B-SFT (Yang et al., 2025a)	32 frames under the default setting
VST-7B-SFT (Yang et al., 2025a)	32 frames under the default setting
VLM-3R-7B (Fan et al., 2025)	32 frames under the default setting
SpatialLadder-3B (Li et al., 2025a)	32 frames under the default setting
Cambrian-S-3B (Yang et al., 2025b)	32 frames under the default setting
Cambrian-S-7B (Yang et al., 2025b)	32 frames under the default setting
LongSpace	1 fps chunked memory inference; 32 frames per chunk; 4-frame overlap

with a 4-frame overlap. For multiple questions from the same video, LongSpace constructs the video memory once and reuses it across different question prompts. This avoids repeated encoding of the full long video and keeps the evaluation protocol consistent with the long-memory inference setting.

A.5 Training Data Sources

Our spatial instruction tuning data are drawn from three complementary sources: VSI-590K, VLM-3R, and a sampled subset of SPAR-7M, denoted as SPAR-234K. This data mixture follows the training data setting of VG-LLM, with VSI-590K further included to strengthen video-level spatial supervision. This mixture provides broad spatial supervision, including object-level perception, geometric relation reasoning, and video-level spatial understanding. Overall, it contains approximately 1.18M spatial instruction samples.

VSI-590K. VSI-590K is a large-scale spatial instruction dataset designed for visual-spatial understanding. We use this dataset to provide video-centric supervision for indoor spatial reasoning, including object perception, distance estimation, relative direction, route reasoning, and appearance-order understanding. These samples help the model learn spatial concepts that are directly tied to observations over video frames.

VLM-3R. VLM-3R introduces instruction data for 3D-aware visual reasoning from monocular video. Its supervision emphasizes spatial context, scene structure, camera motion, and geometry-language alignment. We use the VLM-3R instruction data to strengthen the ability of the model to connect visual observations with implicit 3D structure, which is important for reasoning about layouts and viewpoint changes.

SPAR-234K. SPAR-234K is a sampled subset of SPAR-7M used for spatial reasoning instruction

Algorithm 1: Long-Video Inference with Hierarchical KV Memory (HKM)

Input: Video $\mathcal{V}$, final question q, segment size C, memory budgets $\{B_\rho\}$, read budgets $\{R_\rho\}$
Output: Answer a
Notation: $\mathcal{V}$: input video; q: final question; C: frames per segment; T: number of ordered segments; a: generated answer.
Notation: l: decoder-layer index; $\rho(l)$: role of layer l (sensory, working, or long-memory).
Notation: $\mathcal{M}_{t,l}$: memory at layer l after segment t; $\mathcal{A}_{t,l}$: entries selected from segment t.
Notation: $(\mathbf{K}_{t,l}, \mathbf{V}_{t,l})$: KV states of segment t; $\mathbf{F}_{t,l}$: hidden features used for memory scoring.
Notation: $\mathbf{e}_q$: question embedding; $(\mathbf{K}_l^{\text{mem}}, \mathbf{V}_l^{\text{mem}})$: retrieved KV states; $B_{\rho(l)}/R_{\rho(l)}$: role-specific memory/read budgets.
Sample frames from $\mathcal{V}$ and split them into ordered segments $\{\mathcal{C}_t\}_{t=1}^{T}$ of size C
Initialize HKM states $\mathcal{M}_{0,l} = \emptyset$ for each decoder layer l
for $t = 1$ **to** T **do**
  Encode segment $\mathcal{C}_t$ with Spatial Structure Perception (SSP)
  foreach *decoder layer* l **do**
    Obtain current-segment states $(\mathbf{K}_{t,l}, \mathbf{V}_{t,l}, \mathbf{F}_{t,l})$
    Select position-preserving HKM entries: $\mathcal{A}_{t,l} = \text{Select}_{\rho(l)}(\mathbf{K}_{t,l}, \mathbf{V}_{t,l}, \mathbf{F}_{t,l})$
    Update HKM: $\mathcal{M}_{t,l} = \text{Compress}_{\rho(l)}(\mathcal{M}_{t-1,l} \cup \mathcal{A}_{t,l}; B_{\rho(l)})$
  end
end
Compute query embedding $\mathbf{e}_q$ from the final question q
foreach *decoder layer* l **do**
  Retrieve HKM states: $(\mathbf{K}_l^{\text{mem}}, \mathbf{V}_l^{\text{mem}}) = \text{Read}_{\rho(l)}(\mathcal{M}_{T,l}, \mathbf{e}_q; R_{\rho(l)})$
end
Generate answer a from q using the retrieved HKM states as frozen memory prefixes
return a

tuning. SPAR-7M is built from scenes with 3D ground truth and covers diverse spatial tasks, ranging from basic perception to relation and layout reasoning. Following the subset setting used in VG-LLM, we sample approximately 234K examples to broaden the distribution of spatial relations while keeping the training mixture compact.

A.6 Long-Video Inference with Hierarchical KV Memory

Algorithm 1 provides the Hierarchical KV Memory (HKM) procedure used by LongSpace during inference. The model first samples and splits the video into temporally ordered segments, then encodes each segment with Spatial Structure Perception (SSP). At each decoder layer, HKM selects current KV states and hidden scoring features according to the layer role, and compresses them into the corresponding hierarchical memory.

Notation. Let $\mathcal{V}$ denote the input video, q the final question, C the number of frames per segment, and $\{\mathcal{C}_t\}_{t=1}^{T}$ the resulting sequence of T video segments. Following Section 4.3, l indexes a decoder layer and $\rho(l)$ denotes its role (sensory, working, or long-memory). $B_{\rho(l)}$ specifies the memory budget for that role, while $R_{\rho(l)}$ denotes the number of entries allowed during retrieval. The generated answer is denoted by a.

For segment t at layer l, $\mathbf{K}_{t,l}$ and $\mathbf{V}_{t,l}$ are the KV states produced by the current segment, and $\mathbf{F}_{t,l}$ is the hidden feature used for scoring, consistently with Eqs. 3–4. $\text{Select}_{\rho(l)}$ forms the candidate set $\mathcal{A}_{t,l}$, preserving temporal position information, and $\text{Compress}_{\rho(l)}$ merges it with $\mathcal{M}_{t-1,l}$ to obtain the updated memory $\mathcal{M}_{t,l}$ under budget $B_{\rho(l)}$. After all T segments are written, $\mathbf{e}_q$ denotes the embedding of the final question, and $\text{Read}_{\rho(l)}$ retrieves $(\mathbf{K}_l^{\text{mem}}, \mathbf{V}_l^{\text{mem}})$ from $\mathcal{M}_{T,l}$ under budget $R_{\rho(l)}$. These retrieved KV states remain frozen during generation and are prepended to the autoregressive cache. Thus, HKM changes inference-time memory construction and retrieval without introducing a new attention operator or an additional training objective.

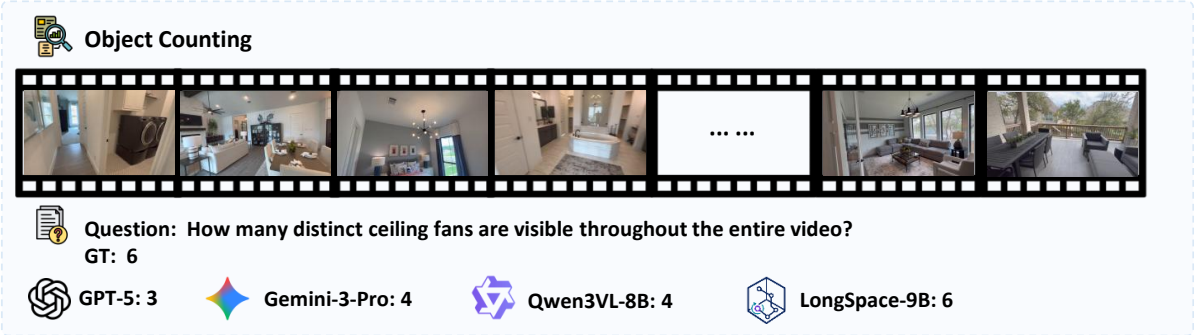


Figure 7: Object Counting QA Example in LongSpace-Bench

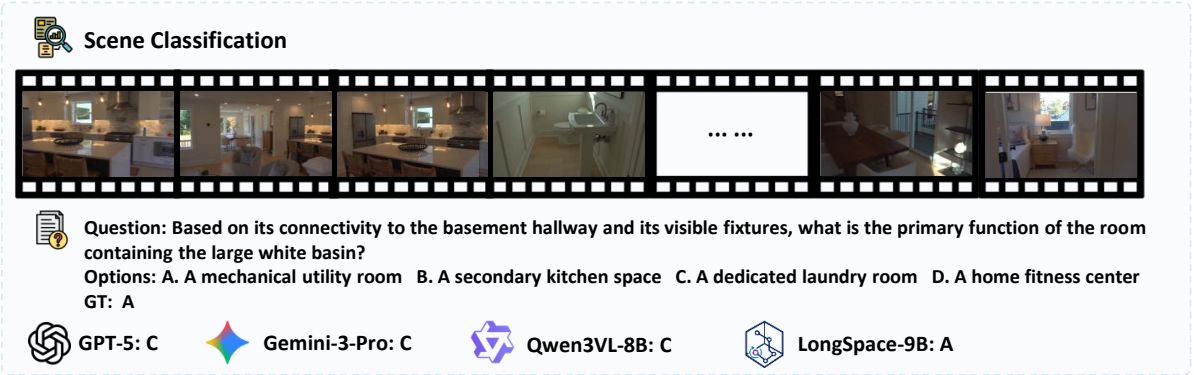


Figure 8: Scene Classification QA Example in LongSpace-Bench

Figure 9: Scene Consistency QA Example in LongSpace-Bench

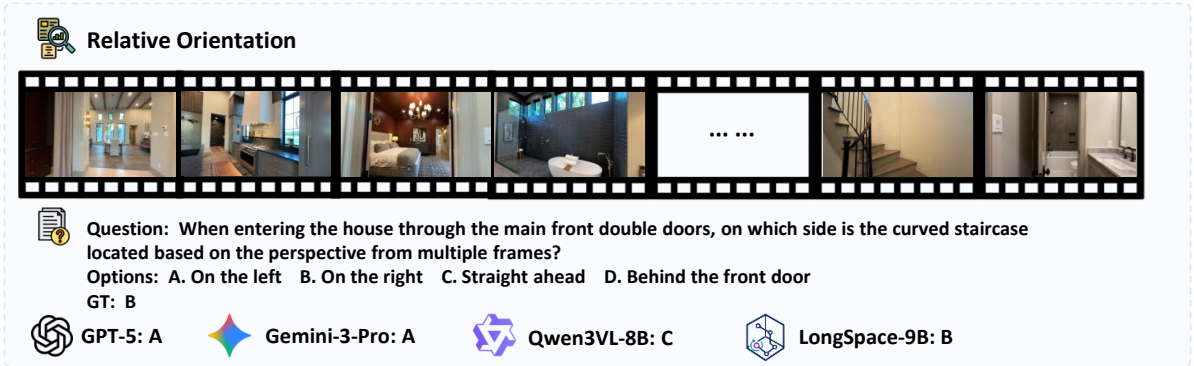


Figure 10: Relative Orientation QA Example in LongSpace-Bench

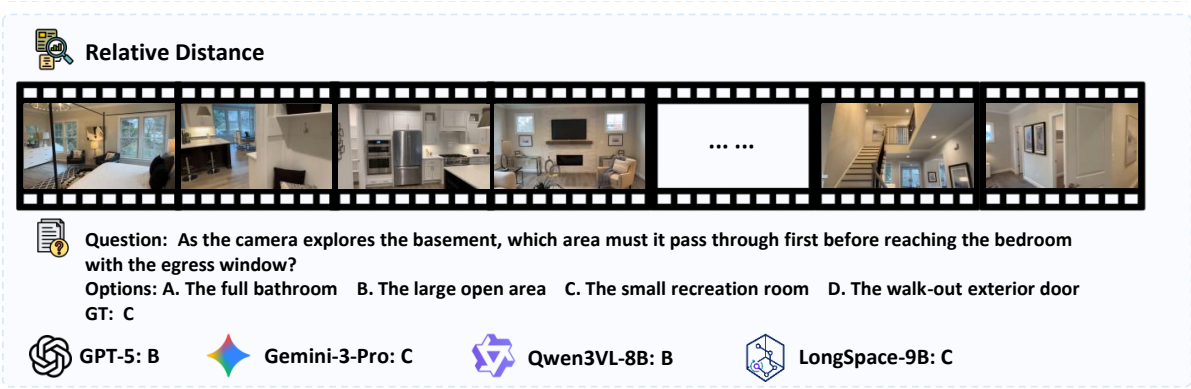


Figure 11: Relative Distance QA Example in LongSpace-Bench

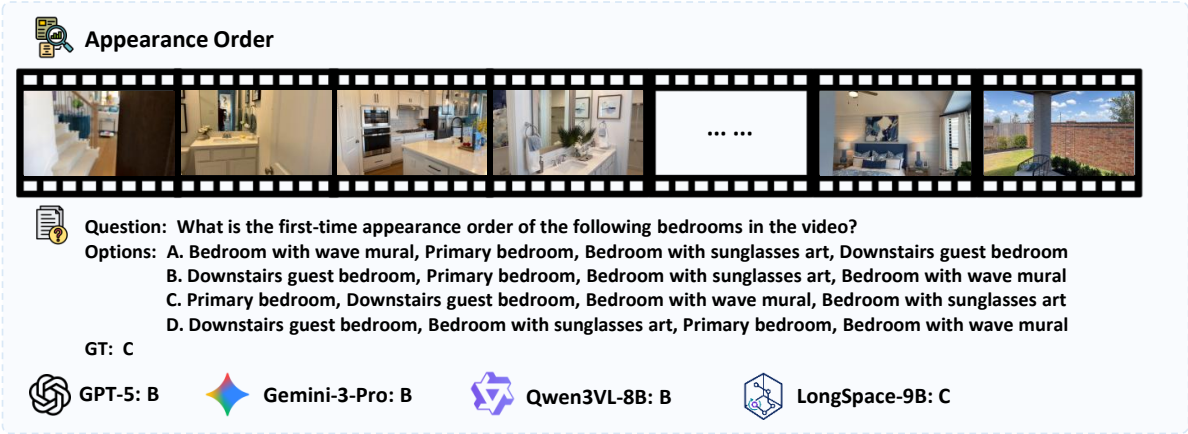


Figure 12: Appearance Order QA Example in LongSpace-Bench

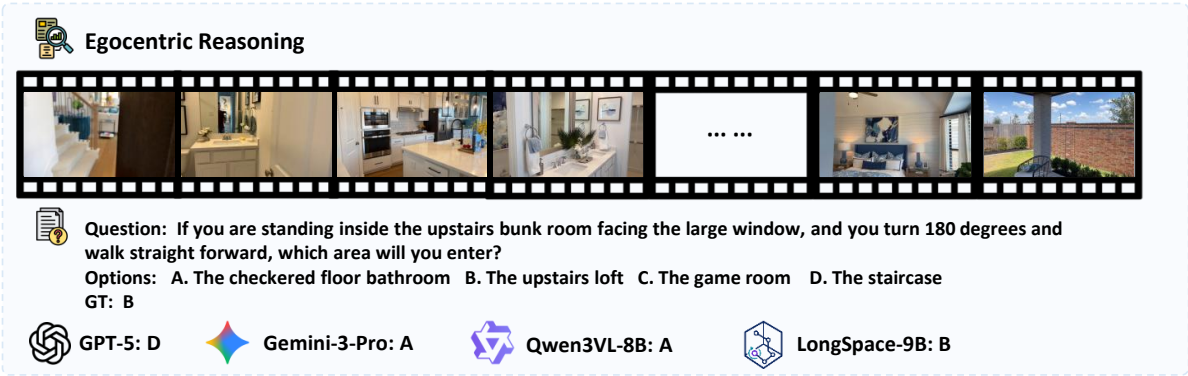


Figure 13: Egocentric Reasoning QA Example in LongSpace-Bench

Figure 14: State Change QA Example in LongSpace-Bench

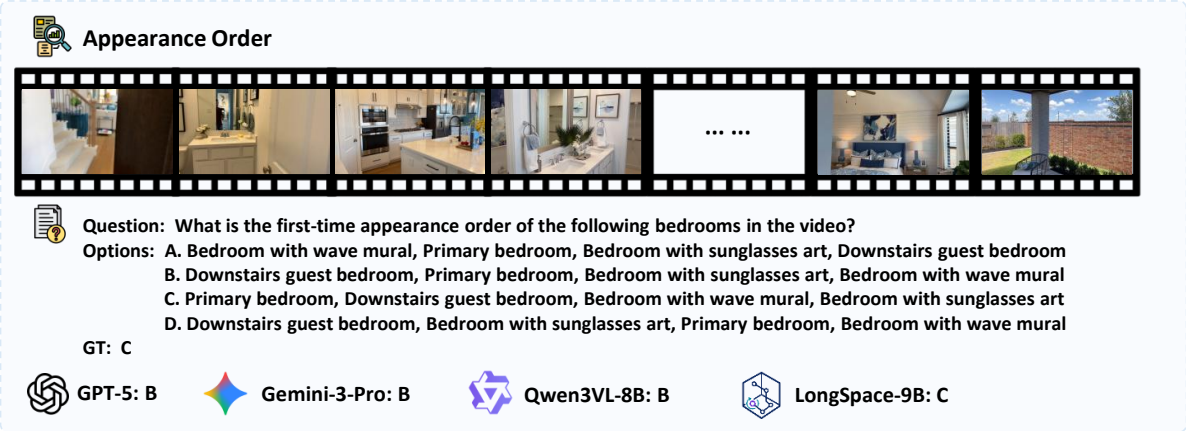


Figure 15: Route Planning QA Example in LongSpace-Bench

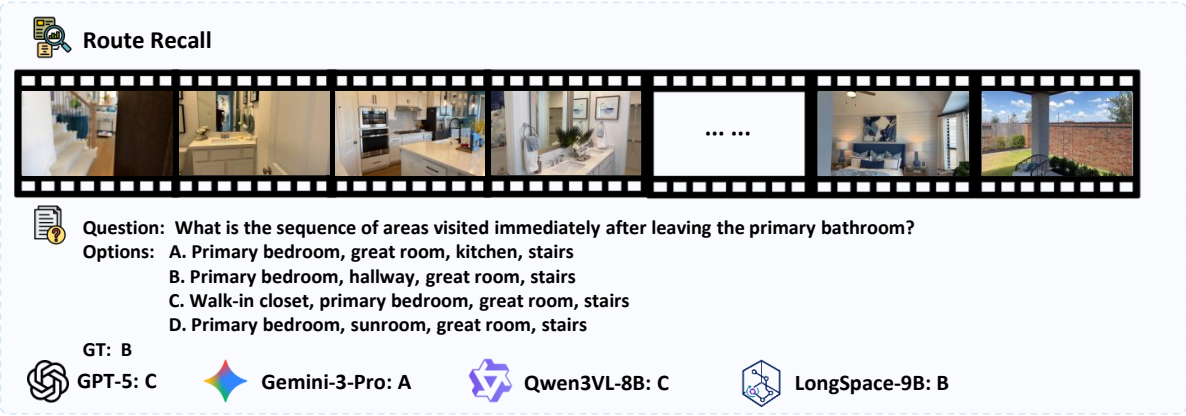


Figure 16: Route Recall QA Example in LongSpace-Bench