LEARNING PERSISTENT WORLD REPRESENTATIONS FOR SPATIAL REASONING FROM PARTIAL OBSERVATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial reasoning over videos and multiple views requires integrating spatial evidence distributed across partial observations. However, existing approaches mainly aggregate visual evidence without explicitly constructing a representation of the observed world. We argue that spatial reasoning should be mediated by a query-independent persistent world representation, causally constructed from partial observations before downstream reasoning. Following this formulation, we propose the *Internal Spatial World Encoder* (ISWE), which augments a vision-language model with a causal world-construction process that enables world-informed visual enrichment. ISWE preserves observation-specific evidence in local states and causally integrates it into persistent World representations through asymmetric Local-to-World updates. The constructed world is maintained through complementary temporal and persistent memory carriers, and is accessed to enrich visual tokens by combining local and cross-observation information before query-dependent answer generation. We further propose *Scope-Aligned Spatial Internalization*, which aligns spatial supervision with the validity scopes of local, world, and memory representations. ISWE-8B achieves a macro-average score of 69.6 on VSI-Bench. It further demonstrates strong generalization on ViewSpatial-Bench and MMSI-Bench, achieving 58.8 and 38.3, respectively.

1 INTRODUCTION

Spatial reasoning from videos and multiple views requires more than extracting evidence from individual observations. Each observation reveals only part of an environment, whereas queries about relative location, distance, route planning, observation order, or changing reference frames can depend on evidence accumulated across time and viewpoints (Yang et al., 2025; 2026c). A model must therefore retain both what was observed and how partial observations relate to one another. We ask: *how can a VLM construct a query-independent persistent world representation from partial observations before the downstream query is known, and subsequently leverage this representation for spatial reasoning?*

Recent approaches primarily improve the amount or quality of spatial evidence available for downstream reasoning. Long-video methods extend context with more frames, retrieve relevant segments, or compress visual information into memory (Chen et al., 2025; Kim et al., 2025a; Man et al., 2025), while spatially enhanced models incorporate reconstructed geometry, camera-motion cues, and spatial instruction tuning (Fan et al., 2026). However, these approaches do not explicitly organize partial observations into a persistent latent representation of the same environment. More evidence or richer geometry can strengthen perception, but cannot replace constructing a persistent world representation that captures relationships across observations. This motivates separating query-independent world construction from downstream query-dependent reasoning.

We therefore propose a new spatial reasoning paradigm, where *a query-independent persistent latent world representation is causally constructed from partial observations before downstream reasoning and subsequently mediates spatial reasoning*. Rather than introducing the query during world construction, the model first organizes partial observations into a query-independent representation of the environment and subsequently leverages this representation for downstream reasoning. This

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

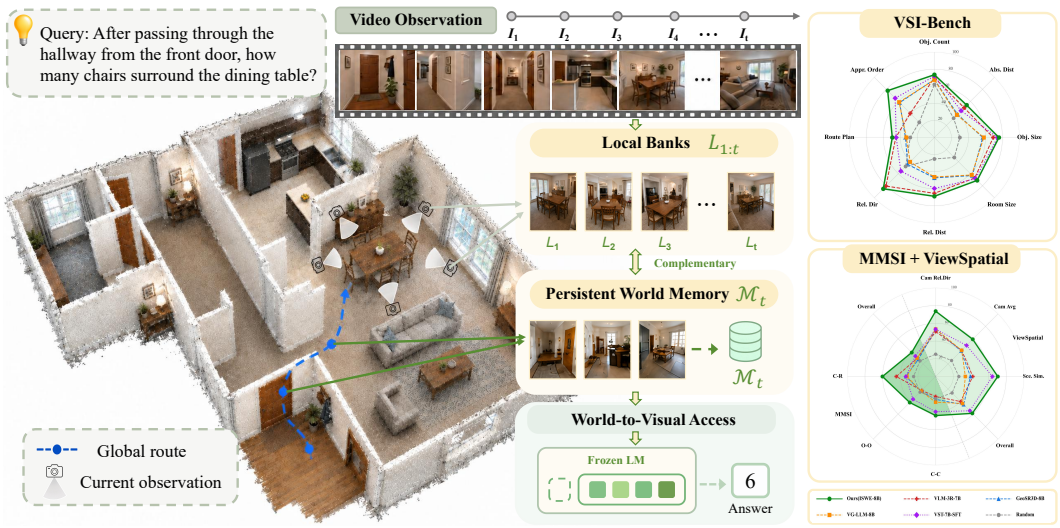


Figure 1: **Overview of ISWE.** ISWE first constructs a query-independent persistent world representation from partial observations and then performs query-dependent reasoning through world-informed visual representations. The world representation integrates observation-specific and cross-observation spatial evidence via Local Banks and Persistent World Memory. Right: per-task comparisons on VSI-Bench and the joint MMSI/ViewSpatial-Bench radar.

formulation separates world construction from query-dependent answer generation, enabling spatial reasoning over a persistent representation of the observed environment.

To instantiate this formulation, we propose the *Internal Spatial World Encoder* (ISWE). ISWE enables world-mediated spatial reasoning by constructing a query-independent latent world before downstream reasoning. This intermediate world representation organizes spatial evidence across observations and provides world-informed visual representations for subsequent reasoning. We further introduce *Scope-Aligned Spatial Internalization* to facilitate structured spatial knowledge acquisition. Experiments show that ISWE-8B achieves 69.6 on VSI-Bench, and obtains strong performance on ViewSpatial-Bench and MMSI-Bench.

Our main contributions are summarized as follows:

- We propose the *Internal Spatial World Encoder* (ISWE), which realizes world-mediated spatial reasoning by constructing a query-independent persistent latent world representation from partial observations. By introducing an explicit world-building process before reasoning, ISWE transforms fragmented observations into a structured spatial state for downstream reasoning.
- We introduce *Scope-Aligned Spatial Internalization*, which aligns supervision with the validity scopes of Local, World, and Memory representations to internalize structured spatial knowledge. This design prevents spatial knowledge entanglement across representation levels while enabling teacher-free inference.
- ISWE-8B achieves strong performance across three spatial reasoning benchmarks, obtaining 69.6 on VSI-Bench, 58.8 on ViewSpatial-Bench, and 38.3 on MMSI-Bench, demonstrating strong generalization across diverse spatial reasoning settings.

2 Related Work

2.1 Spatial Reasoning in Vision-Language Models

Recent studies have strengthened spatial reasoning in vision-language models by enriching the spatial evidence available during query answering (Wang et al., 2026; Ma et al., 2026; Yang et al., 2025; Zhang et al., 2025a). Spatially enhanced VLMs introduce cues such as depth, point clouds, camera

parameters, reconstructed 3D features, or additional spatial supervision to improve metric estimation, viewpoint-dependent reasoning, and cross-view consistency (Ma et al., 2025a;b; Yeh et al., 2026; Cheng et al., 2026; Zhao et al., 2026; Zhou et al., 2025). These approaches provide stronger geometric and physical priors than general vision-language pretraining alone, and have substantially advanced spatial understanding across images, videos, and multi-view observations (Song et al., 2025; Qi et al., 2026; Jeong et al., 2026; Wu et al., 2025). Memory-augmented approaches such as VLM^2 (Liu et al., 2026) leverage historical observations for long-horizon reasoning, while our work constructs a query-independent world representation from multi-observation evidence before downstream reasoning.

2.2 Multi-Observation Memory and Persistent Scene Representation

Long-video and multi-observation reasoning has motivated methods that retain, retrieve, or compress frames and visual tokens to preserve useful information over extended contexts (Chen et al., 2025; Kim et al., 2025a; Wang et al., 2025b; Xue et al., 2025; Zuo et al., 2025). Other approaches maintain recurrent or persistent memory states, or progressively aggregate multi-view observations into compact scene representations (Diko et al., 2025; Man et al., 2025; Zeng et al., 2025; Sohn et al., 2026; Huang et al., 2025). These mechanisms improve information retention and cross-observation reasoning, but they face a common trade-off between preserving intermediate observation-dependent states and compressing accumulated information into a bounded representation (Kim et al., 2025b; Lin et al., 2025). Retaining more historical states better preserves temporal or viewpoint-specific evidence, whereas aggressive consolidation improves efficiency but may discard information whose relevance emerges only later. Our work is closely related to this line of research, but combines retention of stage-wise World snapshots with recurrent spatial consolidation in a query-independent world representation.

2.3 Spatial Knowledge Internalization in Vision-Language Models

A growing line of work transfers spatial knowledge from external geometric or perception models into VLM representations during training, reducing the need for explicit spatial inputs at inference time (Bigverdi et al., 2025; Zong et al., 2025; Yeh et al., 2026). Existing approaches commonly distill depth, 3D structure, object-level relations, or other spatial cues into intermediate visual representations, thereby improving spatial perception without permanently retaining the corresponding teacher models (Zhang et al., 2025a; Song et al., 2025; Qi et al., 2026). An adjacent line transfers VLM-generated supervision into specialized spatial models: UZ3DVG (Tan et al., 2026), for example, uses object-wise pseudo-descriptions and reasoning chains generated by a VLM to train a lightweight 3D visual grounding student, rather than internalizing this knowledge into a VLM. Most VLM-based internalization methods, however, optimize a shared representation and do not explicitly distinguish the different observation ranges over which spatial knowledge remains valid. This distinction becomes important in multi-observation settings, where local geometry, cross-view structure, and persistent spatial relations naturally correspond to different information scopes (Wang et al., 2026; Jeong et al., 2026). Our work extends spatial internalization in this direction by aligning spatial supervision with representations of different scopes while retaining teacher-free inference.

3 Internal Spatial World Encoder

Spatial reasoning over partial observations requires more than aggregating visual evidence. It requires organizing observations into a persistent representation of the observed environment. Following this formulation, ISWE treats the latent world as an intermediate state between observation and reasoning. Visual observations are first transformed into a query-independent world representation, which is subsequently accessed to provide spatial context for downstream reasoning. As illustrated in Fig. 2, ISWE constructs this world through complementary Local, World, and Memory representations. We first formulate the latent world representation and then describe how it is constructed and accessed through Local-to-World Update, Temporal World Bank, Persistent World Memory, and World-to-Visual Access.

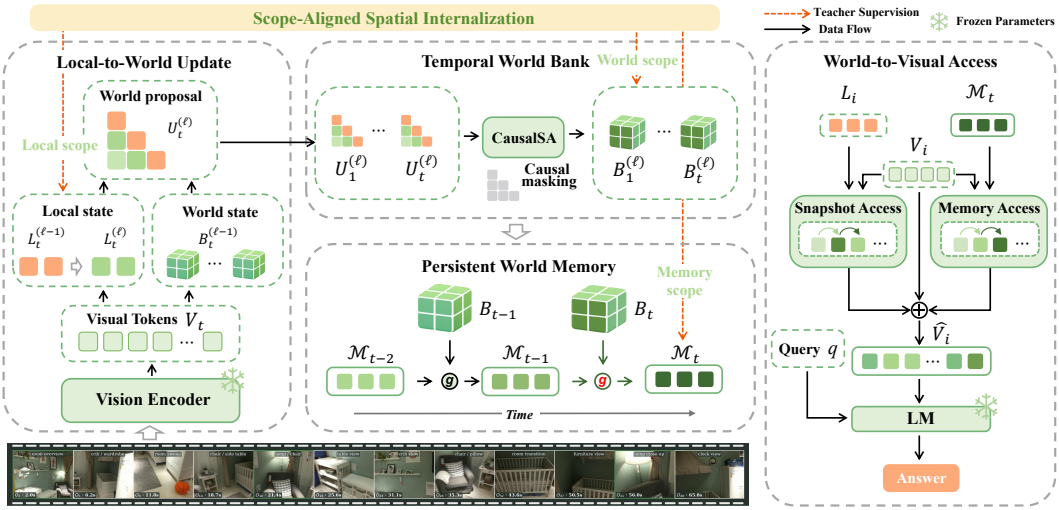


Figure 2: **Overall architecture of ISWE.** One-way Local-to-World updates form the Temporal World Bank $\mathcal{B}_{1:t}$, whose causal World states B_t recurrently update Persistent World Memory $\mathcal{M}_t$. Visual tokens access the corresponding Local Bank L_i and final Persistent World Memory $\mathcal{M}_t$ via Snapshot and Memory Access. Scope-aligned supervision promotes multi-scope spatial knowledge internalization in ISWE, while the query enters only the frozen LM.

3.1 World Representation Interface

As shown in Fig. 2, ISWE exposes two complementary representations for downstream visual access: an observation-specific Local Bank for each input observation and a Persistent World Memory that accumulates information across the observed prefix. For an available observation prefix $I_{1:t}$, we define the resulting world representation as

$$\mathcal{W}_t = (L_{1:t}, \mathcal{M}_t), \qquad L_{1:t} = [L_1; \ldots; L_t], \tag{1}$$

where L_i denotes the Local Bank associated with observation i, and $\mathcal{M}_t$ denotes the fixed-capacity Persistent World Memory after integrating observations up to time t. Local Banks preserve observation-specific evidence, while Persistent World Memory consolidates cross-observation information into a compact state. The Temporal World Bank $B_{1:t}$ serves as an intermediate construction state for world-scope supervision and Memory updates, rather than as part of the visual-access interface $\mathcal{W}_t$.

3.2 Causal World Construction

ISWE constructs the persistent world representation through three stages: Local-to-World Update forms observation-wise proposals, causal integration produces the Temporal World Bank, and the resulting World states recurrently update Persistent World Memory. The first two stages are interleaved over J internal steps, with t indexing observation time and ℓ indexing the internal step.

Observation initialization. The frozen visual encoder provides multi-level hidden features in addition to the native visual tokens V_t. We fuse these features into an observation representation X_t and initialize the two streams as

$$L_t^{(0)} = B_t^{(0)} = X_t. \tag{2}$$

This initialization provides the observation-level input for subsequent world construction. The feature alignment and fusion implementation details are provided in Appendix A.2.

Local-to-World Update. At internal step $\ell \in \{1, \ldots, J\}$, the local state $L_t^{(\ell-1)}$ is independently refined within each observation to obtain $L_t^{(\ell)}$, which is then written into the preceding World state $B_t^{(\ell-1)}$ to form the proposal $U_t^{(\ell)}$:

$$L_t^{(\ell)} = \mathcal{F}_{\mathrm{L}}^{(\ell)}\big(L_t^{(\ell-1)}\big), \qquad U_t^{(\ell)} = B_t^{(\ell-1)} + \alpha^{(\ell)} \mathcal{P}^{(\ell)}\big(L_t^{(\ell)}\big). \tag{3}$$

Here, $\mathcal{F}_{\mathrm{L}}^{(\ell)}$ denotes the ℓ-th Local refinement block applied independently within each observation, $\mathcal{P}^{(\ell)}(\cdot)$ aligns the refined Local evidence with the World stream, and $\alpha^{(\ell)}$ scales its contribution. We collect the observation-wise proposals up to t as $U_{1:t}^{(\ell)} = [U_1^{(\ell)}; \ldots; U_t^{(\ell)}]$. The index ℓ tracks progress through the internal iteration. At every internal step, $U_t^{(\ell)}$ enters the causal world update that produces $B_t^{(\ell)}$. Information flows only from local to world, preserving observation-specific local states, and the complete interleaved update schedule is provided in Appendix A.2. After the final internal step, the local state is spatially merged and projected to form the observation-specific Local Bank $L_t = \rho\big(L_t^{(J)}\big)$, where $\rho(\cdot)$ denotes the spatial merge-and-projection operator.

Temporal World Bank. At internal step ℓ, the world stream applies causal self-attention to the proposal prefix $U_{1:t}^{(\ell)}$, producing the updated World states $B_{1:t}^{(\ell)}$. Causal masking ensures that each $B_i^{(\ell)}$ depends only on $U_{1:i}^{(\ell)}$. After J steps, the resulting states constitute the Temporal World Bank:

$$B_{1:t}^{(\ell)} = \mathrm{CausalSA}^{(\ell)}\big(U_{1:t}^{(\ell)}\big), \qquad \mathcal{B}_{1:t} = B_{1:t}^{(J)} = \Big[B_1^{(J)}; \ldots; B_t^{(J)}\Big]. \tag{4}$$

The updated state $B_t^{(\ell)}$ is propagated to the next internal step through Eq. 3, where $B_t^{(\ell-1)}$ denotes the preceding internal state at the same observation index t. We denote the final state $B_t^{(J)}$ as B_t, which integrates the observation prefix $I_{1:t}$, in contrast to the observation-specific Local state L_t. The ordered final World states $\mathcal{B}_{1:t} = [B_1; \ldots; B_t]$ constitute the Temporal World Bank and support world-scope supervision and Persistent World Memory updates, rather than direct visual readout.

Persistent World Memory. As shown in Fig. 2, each final World state B_t recurrently updates the fixed-capacity Persistent World Memory. Given B_t and the previous state $\mathcal{M}_{t-1}$, the model constructs a candidate $\widetilde{\mathcal{M}}_t$ and uses a learned gate g_t to balance newly integrated evidence against retained memory:

$$\mathcal{M}_t = (1 - g_t) \odot \mathcal{M}_{t-1} + g_t \odot \widetilde{\mathcal{M}}_t. \tag{5}$$

A small g_t preserves the previous Memory, whereas a large g_t incorporates more of the candidate state. Applied sequentially over the ordered World states in $\mathcal{B}_{1:t}$, this update consolidates cross-observation information into the fixed-capacity Persistent World Memory $\mathcal{M}_t$. Details of candidate construction and gate parameterization are provided in Appendix A.4.

3.3 World-to-Visual Access

The final module in Fig. 2 implements $\mathcal{A}_\psi(\cdot)$ to enrich the native visual tokens using both components of $\mathcal{W}_t$. For each observation i in the prefix, *Snapshot Access* $\mathcal{A}_\psi^{\mathrm{snap}}(\cdot)$ retrieves observation-specific evidence from L_i, while *Memory Access* $\mathcal{A}_\psi^{\mathrm{mem}}(\cdot)$ retrieves consolidated context from the final Memory state $\mathcal{M}_t$:

$$\widehat{V}_i = V_i + \Big[\mathcal{A}_\psi^{\mathrm{snap}}(V_i; L_i) + \mathcal{A}_\psi^{\mathrm{mem}}(V_i; \mathcal{M}_t)\Big] W_O, \quad i = 1, \ldots, t. \tag{6}$$

Both branches use projected visual tokens as attention queries. $\mathcal{A}_\psi^{\mathrm{snap}}(\cdot)$ accesses only the corresponding Local Bank L_i, whereas $\mathcal{A}_\psi^{\mathrm{mem}}(\cdot)$ accesses the shared final Memory $\mathcal{M}_t$. Their outputs are summed, projected by W_O, and residually added to V_i, yielding the *world-enriched visual tokens* $\widehat{V}_i$. We collect them as $\widehat{V}_{1:t} = [\widehat{V}_1; \ldots; \widehat{V}_t]$. The enriched visual sequence $\widehat{V}_{1:t}$ is then supplied together with the downstream query q to the frozen LM for autoregressive answer generation. Thus, both world construction and World-to-Visual Access are completed before the query enters the frozen LM. More details are provided in Appendix A.5.

3.4 Scope-Aligned Spatial Internalization

We align supervision with the validity scope of each representation. Grounding DINO (Liu et al., 2024) provides Local semantic supervision, while VGGT (Wang et al., 2025a) supplies Local geometric and World relative-pose targets. LingBot-Map (Chen et al., 2026) provides Geometric Context Attention (GCA) supervision for World and Memory states, together with trajectory regression

for Memory. Let $\mathcal{S} = \text{local}, \text{world}, \text{mem}$ denote the supervised scopes, with $\mathcal{H}_s$ and Z^s representing the corresponding features and teacher targets:

$$\mathcal{L}_s = \mathcal{D}_s(\mathcal{H}_s, Z^s), \quad s \in \mathcal{S}, \qquad \mathcal{L}_{\text{scope}} = \sum_{s \in \mathcal{S}} \mathcal{L}_s. \tag{7}$$

Here, $\mathcal{D}_s(\cdot)$ aggregates the supervision losses within scope s without additional scope-level weighting. Details are provided in Appendix A.6.

3.5 Training Objective and Teacher-Free Inference

The complete training objective is

$$\mathcal{L} = \lambda_{\text{LM}} \mathcal{L}_{\text{LM}} + \mathcal{L}_{\text{scope}}, \tag{8}$$

where $\mathcal{L}_{\text{LM}}$ denotes the autoregressive answer-generation loss. Loss composition and weights are detailed in Appendix A.8. All external teachers are removed at inference, enabling RGB-only World construction and visual enrichment before query-dependent answer generation by the frozen LM.

4 Experiments

4.1 Experimental Setup

Benchmarks and metrics. We evaluate on three complementary benchmarks: VSI-Bench (Yang et al., 2025) for video-based spatial reasoning, ViewSpatial-Bench (Li et al., 2026) for camera- and person-centered spatial localization, and MMSI-Bench (Yang et al., 2026c) for multi-image reasoning about spatial relationships, attributes, motion, and multi-step queries. Importantly, the VLM-3R subset used for training is filtered to have no scene overlap with the VSI-Bench evaluation set, preventing direct scene-level contamination. For VSI-Bench, we report mean relative accuracy (MRA) for numerical answers, accuracy for multiple-choice answers, and the macro-average over eight tasks. For ViewSpatial-Bench, we report per-task, camera- and person-group, and overall accuracy. For MMSI-Bench, we report per-category accuracy and overall accuracy across all examples.

Backbone and training data. We use SenseNova-SI-1.5-InternVL3-8B (Cai et al., 2026) as the backbone. We sample 110K spatial instruction examples from the VLM-3R training set (Fan et al., 2026) and train for one epoch with a learning rate of 2×10^{-5} and a global batch size of 32. Training is performed on four NVIDIA H100 GPUs, with a full training run taking approximately 51 hours. The language model and visual encoder remain frozen, while the ISWE world-construction and access modules are trainable. The ISWE configuration used in the main experiments is query-independent. Controlled comparisons use matched training and evaluation settings, differing only in the explicitly modified component.

Baselines and evaluation protocol. Tables 1 and 2 compare ISWE with three baseline groups: proprietary models, open-source general-purpose models, and spatial intelligence models. Representative baselines include GPT-4o (OpenAI, 2024), Gemini-1.5-Pro (Gemini Team Google et al., 2024), Qwen3-VL (Bai et al., 2025), InternVL3 (Zhu et al., 2025), VST (Yang et al., 2026a), Cambrian-S (Yang et al., 2026b), VLM-3R (Fan et al., 2026), GeoSR3D (Zhang et al., 2026), and VG-LLM (Zheng et al., 2025); the complete list is provided in the tables. We use the official VSI-Bench evaluation implementation (Yang et al., 2025) and `lmms-eval` (Zhang et al., 2025b) for ViewSpatial-Bench and MMSI-Bench. * denotes our reevaluation. All scores use a 0–100 scale, with detailed results and controlled comparisons provided in Appendix C and Table 3.

4.2 Main Results

Tables 1 and 2 examine spatial reasoning under three complementary forms of partial observation: evidence distributed over video, changes in viewpoint and reference frame, and relationships across multiple images. We analyze both overall performance and the task-level patterns most relevant to persistent world representations. All three benchmarks report ISWE with query-independent access.

Table 1: **Performance on VSI-Bench.** Our evaluations follow the official evaluation implementation of *Thinking in Space* (Yang et al., 2025). We report per-task scores on a 0–100 scale, with *Avg.* denoting their macro-average. * marks results evaluated by us under the same evaluation protocol as ISWE. Bold and underlined numbers indicate the best and second-best results, respectively, among open-source general and spatial intelligence models.

Model	Numerical Answer				Multiple-Choice Answer				Avg.
	Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order	
Random Choice (Frequency)	62.1	32.0	29.9	33.1	25.1	47.9	28.4	25.2	34.0
Proprietary Models									
GPT-4o	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5	34.0
Gemini-1.5-Pro	56.2	30.9	64.1	43.6	51.3	46.3	36.0	34.6	45.4
Seed-1.6	43.5	34.4	66.1	52.8	55.1	35.7	44.3	68.0	49.9
Gemini-2.5-Pro	46.0	37.4	68.7	54.4	62.0	43.9	47.4	68.8	53.6
GPT-5	53.3	34.5	73.3	47.5	63.7	48.7	50.3	68.9	55.0
Open-source General Models									
Bagel-7B-MoT	30.1	29.2	35.5	25.8	34.9	41.4	30.4	24.1	31.4
Qwen3-VL-8B-Instruct	67.6	47.0	**76.3**	61.9	58.0	51.0	35.1	66.3	57.9
InternVL3-8B	66.1	34.9	43.6	47.5	48.0	39.3	26.3	31.4	42.1
VILA-1.5-40B	22.4	24.8	48.7	22.7	40.5	25.7	31.5	32.9	31.2
LongVA-7B	38.0	16.6	38.9	22.2	33.1	43.3	25.4	15.7	29.2
LLaVA-NeXT-Video-72B	48.9	22.8	57.4	35.3	42.4	36.7	35.0	48.6	40.9
Spatial Intelligence Models									
ViLaSR-7B	58.1	33.9	61.4	28.9	45.1	46.5	29.9	53.2	44.6
VST-7B-SFT	<u>72.0</u>	44.4	74.3	<u>68.3</u>	59.7	55.8	44.9	65.2	60.6
Cambrian-S-3B	70.7	40.6	68.0	46.3	64.8	61.9	27.3	**78.8**	57.3
VLM-3R-7B	70.2	<u>49.4</u>	69.2	67.1	<u>65.4</u>	<u>80.5</u>	<u>45.4</u>	40.1	<u>60.9</u>
GeoSR3D-8B*	68.1	38.4	56.9	63.0	47.3	44.5	34.0	57.8	51.3
VG-LLM-4B*	67.6	37.3	54.8	53.3	48.6	43.4	29.4	35.1	46.2
VG-LLM-8B*	67.6	37.5	58.1	62.1	46.3	40.4	33.0	58.6	50.4
ISWE-8B	**74.0**	**53.8**	<u>75.7</u>	**72.4**	**69.4**	**85.3**	**49.0**	<u>77.5</u>	**69.6**

VSI-Bench: integrating spatial evidence over video. ISWE-8B achieves the highest macro-average score (69.6) in Table 1, outperforming the strongest spatial intelligence baseline in that table, VLM-3R-7B (60.9), and leading on five of eight tasks: object counting, absolute distance, room size, relative distance, and relative direction. It also outperforms SenseNova-SI, reproduced SpatioLM, and LoRA in Table 3. These gains are consistent with improved integration of scene-level and inter-object evidence, although route planning and appearance order remain below GPT-5 and Cambrian-S-3B, respectively.

ViewSpatial-Bench and MMSI-Bench: reasoning across observations. ISWE-8B demonstrates strong generalization on cross-view and cross-image spatial reasoning. On ViewSpatial-Bench, ISWE-8B achieves 58.8 overall, surpassing VST-7B-SFT by 4.1 points and obtaining the highest camera-group score (59.4; Table 2), with notable gains in camera-centered relative direction (73.7) and person-centered scene simulation (70.4). On MMSI-Bench, ISWE-8B reaches 38.3 overall, outperforming VST-7B-SFT (32.5) and Qwen3.5-9B (36.5) (Appendix Table 10). These results demonstrate the ability of ISWE to integrate spatial evidence across observations.

4.3 Analysis of Persistent World Representation Learning

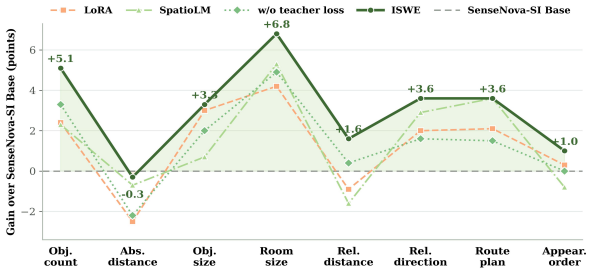
Table 3 compares ISWE with spatial adaptation baselines under same backbone, data, optimization, and evaluation settings. SenseNova-SI-1.5 (Cai et al., 2026) is used as the unadapted base model, while LoRA and reproduced SpatioLM (Wu et al., 2026) serve as adaptation baselines. SpatioLM is reproduced using the same SenseNova-SI-1.5 backbone and training subset while retaining its method-specific DA3 (Lin et al., 2026) supervision. Unlike ISWE, it does not explicitly construct a query-independent persistent latent world representation. ISWE with answer-generation supervi-

Table 2: **Performance on ViewSpatial-Bench and MMSI-Bench.** Scores are reported on a 0–100 scale. Camera/Person denote group accuracies, the first three MMSI columns are macro-averages, and MSR denotes multi-step reasoning. †GPT and ‡Gemini denote proprietary models. Bold and underlined numbers indicate the best and second-best non-proprietary results.

Model	ViewSpatial-Bench			MMSI-Bench				
	Camera	Person	Overall	Pos. Rel.	Geom. Attr.	Motion Perc.	MSR	Overall
Random Choice	25.5	27.1	26.3	25.0	25.0	25.0	25.0	25.0
Proprietary Models								
GPT†	33.6	36.3	35.0	45.3	46.0	34.6	36.4	41.9
Gemini‡	33.7	31.5	32.6	**46.5**	**51.8**	**44.6**	37.9	**45.2**
Gemini 2.5 Pro	**49.4**	39.2	**44.2**	39.6	46.4	32.1	33.3	38.0
Grok 4	45.2	40.2	43.2	40.4	22.3	42.0	38.4	37.8
Seed 1.6	45.4	**42.4**	43.9	37.8	40.7	36.6	**40.4**	38.5
Open-source General Models								
InternVL2.5-2B	32.8	37.0	35.0	30.5	29.5	26.6	26.8	29.0
Qwen2.5-VL-7B	40.6	33.4	36.9	25.9	21.6	29.9	25.8	25.9
LLaVA-NeXT-Video-7B	23.8	37.1	30.6	23.6	30.1	21.3	27.8	24.9
LLaVA-OneVision-7B	28.5	26.5	27.5	28.0	27.8	26.6	11.6	24.5
InternVL2.5-8B	46.5	40.2	43.2	29.9	27.0	23.9	30.3	28.7
Llama-3.2-Vision-11B	23.7	33.6	28.8	26.9	20.0	27.3	19.2	25.4
InternVL3-14B	47.1	33.9	40.3	25.6	27.0	28.0	29.3	26.8
InternVL3.5-8B	40.8	39.3	40.0	27.8	27.8	24.6	**36.4**	29.0
Qwen3-VL-8B-Instruct	45.4	39.2	42.2	33.8	26.4	28.0	29.8	31.1
Spatial Intelligence Models								
Cambrian-S-3B	35.0	41.4	38.3	31.0	24.0	26.6	19.2	27.1
VLM-3R-7B	41.8	38.3	40.0	34.1	33.0	21.2	20.7	29.2
VST-7B-SFT	49.6	**59.5**	54.7	35.4	38.4	**36.0**	18.2	32.5
GeoSR3D-8B	40.8	48.8	44.9	29.9	24.6	26.6	24.2	27.5
VG-LLM-4B	36.9	49.9	43.6	27.9	31.6	21.9	28.3	27.6
VG-LLM-8B	41.7	44.5	43.1	30.6	34.6	28.6	24.2	29.5
ISWE-8B	**59.4**	58.2	**58.8**	**46.5**	**39.2**	32.0	21.2	**38.3**

Table 3: We compare ISWE with the base model SenseNova-SI-1.5, LoRA fine-tuning on SenseNova-SI-1.5, and SpatioLM reproduced using the same backbone with its original DA3 (Lin et al., 2026) supervision. ISWE (answer loss only) retains the full ISWE architecture while removing auxiliary spatial supervision. Left: Num., MC, and Avg.; Right: per-task gains of full ISWE over SenseNova-SI-1.5.

Method	Num.	MC	Avg.
SenseNova-SI-1.5	65.3	67.8	66.5
LoRA ($r = 48$)	67.0	68.7	67.9
SpatioLM (reproduced)	67.2	68.9	68.0
ISWE (answer loss only)	67.3	68.7	68.0
ISWE	**69.0**	**70.3**	**69.6**
Δ vs. SenseNova-SI-1.5	**+3.7**	**+2.5**	**+3.1**



sion alone achieves 68.0, matching SpatioLM and outperforming the base model and LoRA. Introducing the proposed auxiliary spatial objectives further improves ISWE to 69.6, demonstrating the effectiveness of the complete ISWE training design. Detailed per-task comparisons are provided in Appendix C.2.

Qualitative examples of world-informed reasoning. Fig. 3 illustrates the behavior of the trained ISWE model under three inference-time interventions with all parameters fixed. Zero Memory removes the Memory Access contribution, Nearest-Video Memory substitutes memory from another

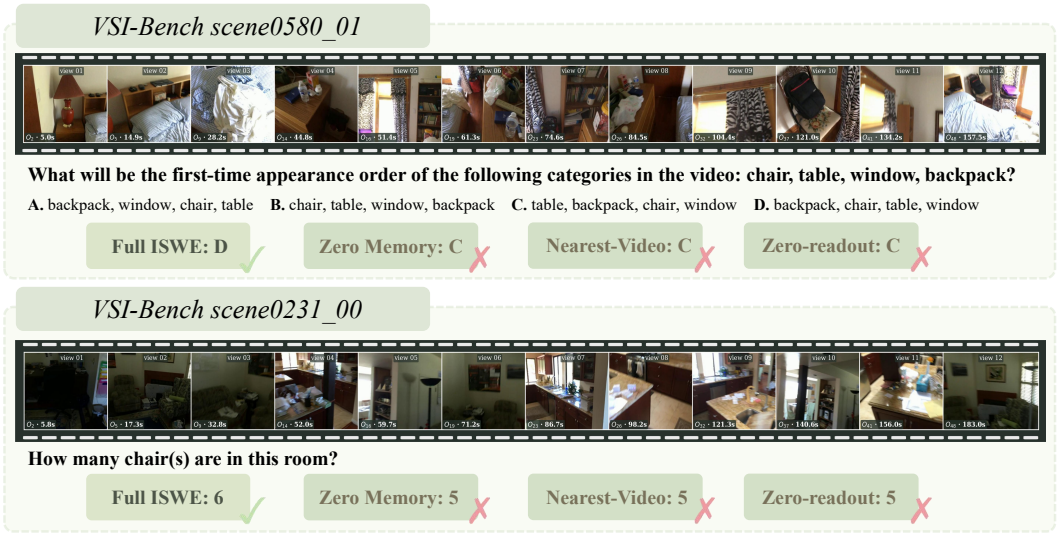


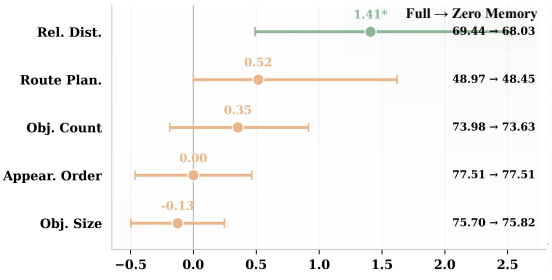
Figure 3: **Qualitative examples of world-informed reasoning on VSI-Bench.** Full ISWE produces the annotated correct answer in the two examples shown, whereas the corresponding inference-time intervention settings produce different, incorrect predictions.

video with the most similar representation, and Zero-readout removes both Local and Memory enhancement (see Appendix A.5). Full ISWE answers both examples correctly, whereas the interventions yield incorrect predictions.

Inference-time diagnostic of Persistent World Memory. Table 4 examines whether the trained ISWE model functionally uses Persistent World Memory at inference. With all parameters fixed, we remove only the Memory contribution from the visual readout. The resulting changes are task-dependent: Relative Distance decreases by 1.41 points, Route Planning by 0.52, and Object Counting by 0.35, whereas Appearance Order is unchanged and Object Size changes negligibly. These patterns are consistent with Persistent World Memory providing complementary cross-observation context while Local Banks retain observation-specific evidence.

Table 4: **Frozen-parameter diagnostic of Persistent World Memory.** With all parameters fixed, we remove only the inference-time Memory contribution from the trained ISWE model.

Task	Evidence Type	Full	Zero M	Drop
Rel. Dist.	Layout	69.44	68.03	**1.41**
Route Plan.	Layout + Traj.	48.97	48.45	0.52
Obj. Count	Layout	73.98	73.63	0.35
Appear. Order	Temporal	77.51	77.51	0.00
Obj. Size	Local / Depth	75.70	75.82	−0.12



5 Conclusion

We presented ISWE, a framework for learning query-independent persistent world representations from partial observations. ISWE causally organizes observation-specific and cross-observation information through asymmetric Local-to-World updates, while Scope-Aligned Spatial Internalization aligns supervision with Local, World, and Memory scopes under teacher-free RGB-only inference. ISWE-8B achieves strong performance across three public spatial reasoning benchmarks, with extensive qualitative analyses further supporting the effectiveness of persistent world representation learning across diverse tasks and challenging partial-observation reasoning settings.

AI USE STATEMENT

In this work, we used generative AI tools for manuscript editing, readability improvement, LaTeX formatting, citation organization, limited code editing assistance, and summarizing already-computed experimental results. We did not use generative AI tools to generate datasets or experimental results, formulate mathematical proofs, design the methodology, or determine the experimental protocol. The research questions, core methodology, and experimental design were developed by the authors. All AI-assisted materials were reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We describe the model architecture and training objectives in Section 3. Appendix A provides details on data composition, model configuration, optimization settings, supervision targets, loss definitions, random seeds, hardware, and training time. Evaluation protocols and fine-grained results are reported in Section 4 and Appendix C. We plan to release the code, configurations, and trained models upon acceptance, subject to applicable licensing requirements.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Mahtab Bigverdi, Zelun Luo, Cheng-Yu Hsieh, Ethan Shen, Dongping Chen, Linda G. Shapiro, and Ranjay Krishna. Perception tokens enhance visual reasoning in multimodal language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3836–3845, June 2025.

Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Tongxi Zhou, et al. Scaling spatial intelligence with multimodal foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7879–7890, 2026.

Lin-Zhuo Chen, Jian Gao, Shangzhan Zhang, Yihang Chen, Ka Leong Cheng, Yipengjing Sun, et al. LingBot-Map: Geometric context transformer for streaming 3D reconstruction. In *European Conference on Computer Vision (ECCV)*, 2026.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, et al. Longvila: Scaling long-context visual language models for long videos. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations (ICLR)*, volume 2025, pp. 18227–18246, 2025.

An-Chieh Cheng, Yang Fu, Yatai Ji, Ligeng Zhu, Guanqi Zhan, Zhuoyang Zhang, Zhaojing Yang, Song Han, Yao Lu, et al. Grounded 3d-aware spatial vision-language modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16688–16700, June 2026.

Anxhelo Diko, Tinghuai Wang, Wassim Swaileh, Shiyan Sun, and Ioannis Patras. Rewind: Understanding long videos with instructed learnable memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13734–13743, June 2025.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Peihao Wang, Huaizhi Qu, Shijie Zhou, Dilin Wang, et al. VLM-3R: Vision-language models augmented with instruction-aligned 3D reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 31054–31065, 2026.

Gemini Team Google et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Zeyi Huang, Yuyang Ji, Xiaofang Wang, Nikhil Mehta, Tong Xiao, Donghyun Lee, et al. Building a mind palace: Structuring environment-grounded semantic graphs for effective long video analysis with llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24169–24179, June 2025.

Suchae Jeong, Jaehwi Song, Haeone Lee, Hanna Kim, Jian Kim, Dongjun Lee, Dong Kyu Shin, Changyeon Kim, et al. Learning multi-view spatial reasoning from cross-view relations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2570–2581, June 2026.

Junho Kim, Hyunjun Kim, Hosu Lee, and Yong Man Ro. Salova: Segment-augmented long video assistant for targeted retrieval and routing in long-form video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3352–3362, June 2025a.

Minsoo Kim, Kyuhong Shim, Jungwook Choi, and Simyung Chang. InfiniPot-V: Memory-constrained KV cache compression for streaming video understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025b.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, et al. ViewSpatial-Bench: Evaluating multi-perspective spatial localization in vision-language models. In *European Conference on Computer Vision (ECCV)*, 2026.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations (ICLR)*, 2026.

Jingyang Lin, Jialian Wu, Ximeng Sun, Ze Wang, Jiang Liu, Yusheng Su, Xiaodong Yu, Hao Chen, Jiebo Luo, Zicheng Liu, and Emad Barsoum. Unleashing hour-scale video training for long video-language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 17523–17552, 2025.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision (ECCV)*, pp. 38–55. Springer, 2024.

Zuntao Liu, Yi Du, Taimeng Fu, Shaoshu Su, Cherie Ho, and Chen Wang. Vision-language memory for spatial reasoning. In *European Conference on Computer Vision (ECCV)*, pp. 297–316. Springer, 2026.

Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso de Melo, Jianwen Xie, and Alan Yuille. SpatialReasoner: Towards explicit and generalizable 3D spatial reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 140751–140774, 2025a.

Wufei Ma, Luoxin Ye, Nessa McWeeney, Celso M. de Melo, Alan Yuille, and Jieneng Chen. Spatial-llm: A compound 3d-informed design towards spatially-intelligent large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17249–17260, June 2025b.

Xianzheng Ma, Tao Sun, Shuai Chen, Yash Bhalgat, Jindong Gu, Angel Chang, Iro Armeni, et al. Do 3d large language models really understand 3d spatial relationships? In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations (ICLR)*, volume 2026, pp. 32591–32610, 2026.

Yuanbin Man, Ying Huang, Chengming Zhang, Bingzhe Li, Wei Niu, and Miao Yin. Adacmˆ2: On understanding extremely long-term video with adaptive cross-modality memory reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8534–8544, June 2025.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Zhangyang Qi, Zhixiong Zhang, Ye Fang, Jiaqi Wang, and Hengshuang Zhao. Gpt4scene: Understand 3d scenes from videos with vision-language models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations (ICLR)*, volume 2026, pp. 157277–157302, 2026.

Tin Stribor Sohn, Maximilian Dillitzer, Jason J. Corso, and Eric Sax. R4: Retrieval-augmented reasoning for vision-language models in 4d spatio-temporal space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 38711–38721, June 2026.

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

Wenbin Tan, Jiawen Lin, Yuan Xie, Yachao Zhang, and Yanyun Qu. Uz3dvg: Unaided zero-shot 3d visual grounding with generated language conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9547–9557, June 2026.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a.

Lan Wang, Yujia Chen, Du Tran, Vishnu Naresh Boddeti, and Wen-Sheng Chu. Seal: Semantic attention learning for long video representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26192–26201, June 2025b.

Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, et al. Spatial mental modeling from limited views. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations (ICLR)*, volume 2026, pp. 82559–82622, 2026.

Jing Wu, Jianhua Wu, Jiayi Guan, Jiahong Chen, Jinghui Lu, Hangjun Ye, Bingzhao Gao, and Long Chen. SpatioLM: Towards general physical spatial intelligence in vision-language models. In *International Conference on Machine Learning (ICML)*, 2026.

Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 143297–143330, 2025.

Zhucun Xue, Jiangning Zhang, Xurong Xie, Yuxuan Cai, Yong Liu, Xiangtai Li, and Dacheng Tao. AdaVideoRAG: Omni-contextual adaptive retrieval-augmented efficient long video understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 6094–6117, 2025.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643. IEEE, 2025.

Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, et al. Visual spatial tuning. In *European Conference on Computer Vision (ECCV)*, 2026a.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, et al. Cambrian-S: Towards spatial supersensing in video. In *International Conference on Learning Representations (ICLR)*, 2026b.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, et al. MMSI-Bench: A benchmark for multi-image spatial intelligence. In *International Conference on Learning Representations (ICLR)*, volume 2026, pp. 157051–157088, 2026c.

Chun-Hsiao Yeh, Shengyi Qian, Manchen Wang, Yi Ma, Joseph Tighe, and Fanyi Xiao. Beyond 3d vqas: Injecting 3d spatial priors into vision-language models for enhanced geometric reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16723–16733, June 2026.

Xiangyu Zeng, Kefan Qiu, Qingyu Zhang, Xinhao Li, Jing Wang, Jiaxin Li, Ziang Yan, Kun Tian, Meng Tian, et al. StreamForest: Efficient online video understanding with persistent event memory. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yu-Jie Yuan, et al. From flatland to space: Teaching vision-language models to perceive and reason in 3D. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025a.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, et al. LMMs-Eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 881–916. Association for Computational Linguistics, 2025b.

Shihua Zhang, Qiuhong Shen, Shizun Wang, Tianbo Pan, and Xinchao Wang. Make geometry matter for spatial reasoning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026.

Ruosen Zhao, Zhikang Zhang, Jialei Xu, Jiahao Chang, Dong Chen, Lingyun Li, Weijian Sun, and Zizhuang Wei. Spacemind: Camera-guided modality fusion for spatial reasoning in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16811–16822, June 2026.

Duo Zheng, Shijia Huang, Yanyang Li, and Liwei Wang. Learning from videos for 3D world: Enhancing MLLMs with 3D vision geometry priors. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, et al. RoboRefer: Towards spatial referring with reasoning in vision-language models for robotics. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 28404–28481, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Yongshuo Zong, Qin Zhang, Dongsheng An, Zhihua Li, Xiang Xu, Linghan Xu, Zhuowen Tu, et al. Ground-v: Teaching vlms to ground complex instructions in pixels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24635–24645, June 2025.

Jialong Zuo, Yongtai Deng, Lingdong Kong, Jingkang Yang, Rui Jin, Yiwei Zhang, et al. VideoLucy: Deep memory backtracking for long video understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 25660–25691, 2025.