

ConsiSpace: Learning Geometric Consistency Matters for Video Spatial Reasoning

Ting Huang², Zhenyu Zhang^{1†}, Wenyuan Huang¹,
Jian Yang¹, and Hao Tang^{2,3†}

¹ School of Intelligent Science and Technology, Nanjing University, Nanjing, China

² School of Computer Science, Peking University, Beijing, China

³ Beijing Academy of Artificial Intelligence, Beijing, China

[†]Corresponding authors: zhangjesse@foxmail.com bjdxtanghao@gmail.com.

<https://believeht029.github.io/ConsiSpace/>

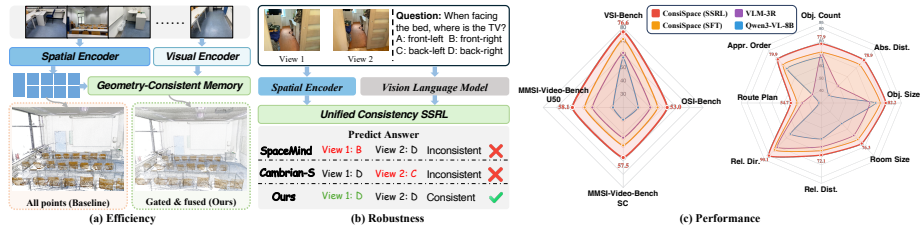


Fig. 1: ConsiSpace learning geometric consistency for video spatial reasoning. (a) **Efficiency:** Geometry-consistent memory reduces redundant spatial evidence through gated writing and fusion. (b) **Robustness:** Unified Consistency SSRL improves answer stability across viewpoints. (c) **Performance:** ConsiSpace consistently improves spatial reasoning performance across benchmarks and VSI sub-tasks.

Abstract. Video spatial reasoning is essential for navigation-oriented perception and long-video question answering, where models must infer spatial relations across long horizons under changing viewpoints. However, existing multimodal large language models (MLLMs) remain largely semantic-centric, and often fail to reliably aggregate consistent spatial evidence from redundant video observations, leading to inefficient or unstable reasoning. To address these issues, we propose **ConsiSpace**, a geometry-consistency-aware framework for geometry-sensitive video spatial reasoning that turns spatial consistency into both an evidence organization principle and an explicit post-SFT learning signal. We build a geometry-consistent memory (GCM) including implicit evidence tokens and explicit geometric cues, and leverage efficient organization strategies to compactly preserve task-related spatial evidence. Furthermore, we utilize unified consistency self-supervised reinforcement learning (UC-SSRL) after supervised fine-tuning to improve cross-view stability, with answer-, metric-, and topology-consistency rewards. Extensive experiments on three spatial-reasoning benchmarks, VSI-Bench, OSI-Bench, and MMSI-Video-Bench, show consistent gains, improving the average score by **12.6** points over the strongest baselines.

Keywords: Video spatial reasoning · Geometric consistency · Self-supervised reinforcement learning

1 Introduction

Video spatial reasoning is a core capability for navigation-oriented perception [10, 16, 24, 27, 28, 39] and video question answering [3, 33, 37, 48]. It requires agents and models to infer object locations, scene connectivity, directions, and spatial relations under partial observability or long video contexts. These demands make video spatial reasoning a practical challenge at the intersection of long-context understanding and physical-space reasoning.

Despite rapid progress in multimodal large language models (MLLMs) [1, 11–13, 29–31, 53], recent spatial benchmarks [23, 44, 45] show that reliable spatial reasoning remains a key bottleneck. Many video-MLLMs are semantics-first: they emphasize recognition and language-conditioned reasoning, while modeling viewpoint geometry and spatial structure only weakly. Recent works address this limitation through spatial data, training objectives, and reasoning formats, such as SpaceR [26], ViLaSR [43], Cambrian-S [46], and RoboBrain [15]. Another direction incorporates explicit spatial cues into MLLMs. Feed-forward geometry foundation models, such as DUS3R [36], MAST3R [17], and VGGT [35], make pose, depth, and structure signals available from RGB videos. Built on these geometry priors, Spatial-MLLM [42], VLM-3R [7], SpaceMind [50], and GeoThinker [20] fuse geometry-aware tokens into MLLMs. Other methods exploit explicit 3D representations or maps for stronger grounding and interpretability, including Video-3D LLM [52], N3D-VLM [41], GS-Reasoner [4], and Map2Thought [8].

However, video spatial reasoning still faces two practical challenges: efficiency and robustness under viewpoint changes. Adjacent frames often contain redundant spatial evidence, so storing or attending to more frames can increase memory and computation without proportional gains, as shown in Fig. 1(a). Moreover, redundant or mismatched evidence may cause inconsistent spatial grounding and unstable answers across viewpoints, as shown in Fig. 1(b). We argue that both challenges can be addressed by exploiting *spatial consistency*: under smooth viewpoint changes of a static scene, correct spatial relations should remain invariant, providing a natural signal for compact evidence organization and stable reasoning.

To this end, we propose **ConsiSpace**, a *geometry-consistency-aware* framework for video spatial reasoning. ConsiSpace uses 3D spatial consistency as both an evidence organization principle and a post-SFT learning objective. It introduces a *Geometry-Consistent Memory* that stores visual-spatial evidence with explicit pose and depth cues, and organizes them through geometry-guided writing, fusion, and retrieval for compact and robust reasoning. To further improve viewpoint stability, we propose Unified Consistency Self-Supervised Reinforcement Learning (UC-SSRL), which refines answer, metric, and topological consistency across paired views after supervised fine-tuning anchors answer correctness. Extensive experiments on **MMSI-Video-Bench** [23], **OSI-Bench** [44], and **VSI-Bench** [45] demonstrate the effectiveness of ConsiSpace across spatial construction, relational, and metric reasoning tasks. Overall, ConsiSpace improves the average score by **12.6** points over the strongest baselines, as illustrated in Fig. 1(c).

Our main contributions are as follows:

- We propose **ConsiSpace**, a geometry-consistency-aware framework for video spatial reasoning that organizes spatial evidence and improves viewpoint stability.
- We introduce a Geometry-Consistent Dual-Memory with geometry-guided writing, fusion, and retrieval, together with UC-SSRL for post-SFT consistency learning.
- Extensive experiments and analyses on three spatial-reasoning benchmarks show that ConsiSpace improves the average score by **12.6** points over the strongest baselines.

2 Related Work

Multimodal large language models. MLLMs [1, 29, 31] have advanced general image and video understanding. Recent long-video MLLMs [5] and memory-augmented agents [38] improve temporal scalability through longer contexts, retrieval, or external memory. However, recent spatial benchmarks [6, 18, 23, 45, 47] show that reliable spatial reasoning remains challenging. General-purpose MLLMs are strong in semantic perception and language-conditioned reasoning, but they often lack explicit mechanisms for maintaining spatial consistency across viewpoints and long temporal contexts. Our work builds on MLLMs, while focusing on geometry-sensitive video spatial reasoning rather than general video understanding.

Spatial-aware MLLMs for spatial reasoning. Prior work improves spatial reasoning through 3D cues, spatial supervision, and reasoning-centric training. Some methods inject explicit 3D information, such as Video-3D LLM [52] and N3D-VLM [41]. Others rely on reconstruction- or consistency-based supervision, such as RoSS3D [34] and VLM-3R [7], or scale spatial data and training, such as Cambrian-S [46]. Reasoning-centric methods further enhance complex spatial reasoning with RL, grounding-aware CoT, or 3D-centric reasoning, including SpatialLadder [21], GS-Reasoner [4], and Think3D [49]. Several methods also design geometry-aware integration modules, such as SpaceMind [50] and GeoThinker [20]. Despite strong progress, these methods mainly focus on representation design or training-time supervision, with limited control over how redundant video evidence is written, fused, and retrieved. In contrast, we use geometric consistency as an explicit criterion for managing the full memory life-cycle.

MLLMs with geometry foundation models. Geometry foundation models provide strong 3D priors from monocular or multi-view imagery, enabling geometry-aware perception without full 3D supervision. Representative examples include DUSt3R-style reconstruction [36], feed-forward geometry with VGGT [35], permutation-equivariant geometry learning (π^3) [40], and depth foundations such as Depth Anything 3 [22]. These models are typically incorporated into MLLMs via feature fusion, distillation, or alignment, as in VG-LLM [51], G²VLM [9], 3DRS [14], Spatial Forcing [19], and Spatial-MLLM [42]. However, geometry

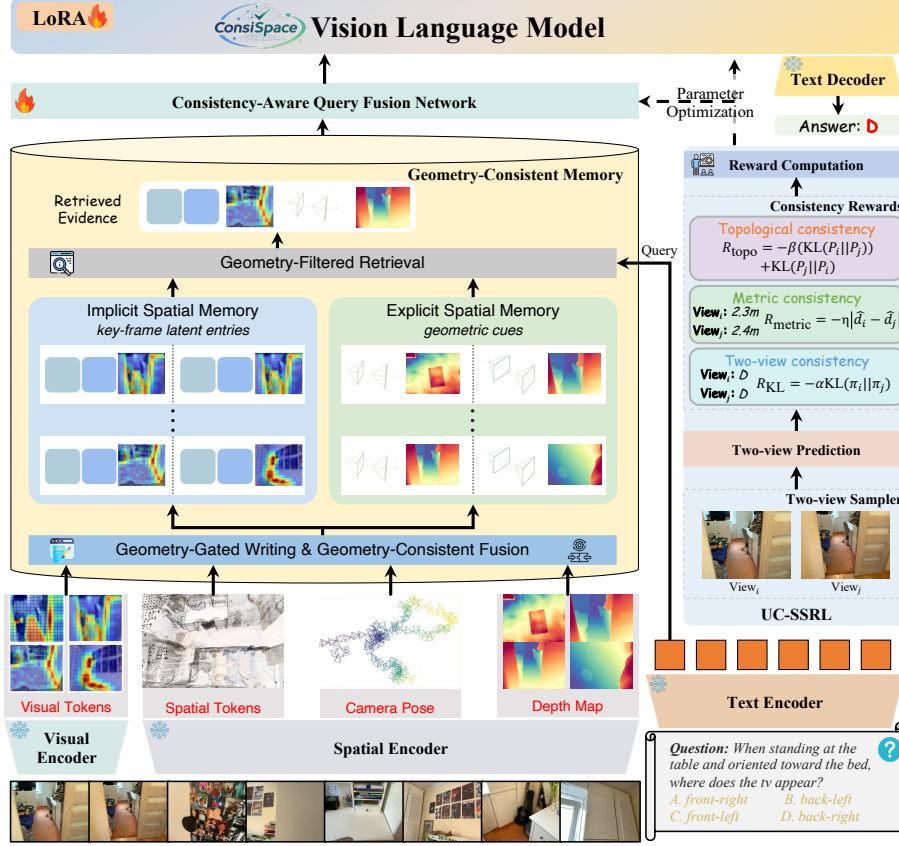


Fig. 2: Overview of ConsiSpace. ConsiSpace organizes visual-spatial tokens and pose-depth cues with a geometry-consistent memory, retrieves query-relevant evidence for answer generation, and uses UC-SSRL to improve viewpoint stability.

priors are mostly used as input features or training signals, while their role in long-video evidence management remains less explored. We instead treat geometry as a consistency criterion for deciding when to write, how to fuse, and what to retain during hierarchical retrieval.

3 Method

3.1 Overview

We consider video spatial reasoning, where a model answers a query q given a video $V = \{I_t\}_{t=1}^T$, producing an output y . This setting requires reliable spatial inference over long temporal horizons, while observations are often redundant and spatial evidence can be sparse or viewpoint-dependent. ConsiSpace addresses this problem by using **geometric consistency** to organize video evidence and stabilize cross-view reasoning. As illustrated in Fig. 2, ConsiSpace consists of dual encoders, a geometry-consistent memory (GCM) with update and read

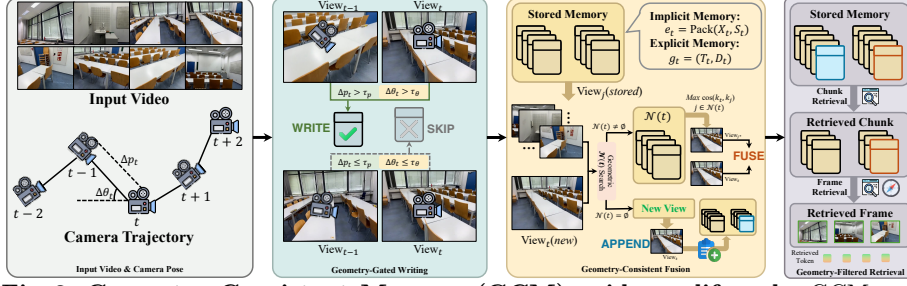


Fig. 3: Geometry-Consistent Memory (GCM) evidence lifecycle. GCM uses geometry-gated writing, geometry-consistent fusion, and geometry-filtered hierarchical retrieval to compact redundant views and retrieve query-relevant evidence.

operations, a query-conditioned fusion module, and a video language model for prediction. During training, we start from the supervised fine-tuned model and apply unified consistency self-supervised reinforcement learning (UC-SSRL) to refine viewpoint stability with self-supervised consistency rewards.

3.2 Feature Extraction

Visual tokens. Given each frame I_t , we extract patch-level visual embeddings using a frozen SigLIP2 [32] vision encoder E_{vis} :

$$X_t = E_{\text{vis}}(I_t), \quad X_t \in \mathbb{R}^{N \times d}, \quad (1)$$

where N denotes the number of visual tokens and d is the embedding dimension. These tokens provide semantic evidence for memory construction and retrieval.

Spatial tokens and geometric cues. We extract geometry-aware representations with a frozen VGGT [35] geometry encoder E_{spa} . Given a frame I_t optionally conditioned on preceding frames $I_{<t}$, VGGT produces spatial tokens and explicit geometric cues:

$$S_t, G_t = E_{\text{spa}}(I_t; I_{<t}), \quad (2)$$

where S_t denotes geometry-aware spatial tokens, and G_t includes camera pose T_t and per-frame depth map D_t . These cues provide geometric consistency signals for memory writing, fusion, and retrieval.

3.3 Geometry-Consistent Memory

ConsiSpace maintains a *Geometry-Consistent Memory* $\mathcal{M}$ with two complementary banks (Fig. 2). The Implicit Spatial Memory $\mathcal{M}_{\text{imp}}$ stores per-frame evidence tokens, including visual semantics and spatial tokens, for query-relevant retrieval. The Explicit Spatial Memory $\mathcal{M}_{\text{exp}}$ stores camera poses and depth cues, which provide geometry-based consistency signals for memory control. This separation decouples evidence representation from geometry-guided operations: $\mathcal{M}_{\text{imp}}$ supports evidence retrieval, while $\mathcal{M}_{\text{exp}}$ guides geometry-gated writing, geometry-consistent fusion, and geometry-filtered retrieval (Fig. 3).

Memory entries. For each frame t , we construct an implicit evidence entry by packing visual tokens X_t and spatial tokens S_t :

$$e_t = \text{PACK}(X_t, S_t), \quad g_t = (T_t, D_t), \quad (3)$$

where T_t is the camera pose and D_t is the per-frame depth map predicted by the spatial encoder. $\text{PACK}(\cdot)$ concatenates and projects visual and spatial tokens into a fixed-length evidence representation. We store e_t in $\mathcal{M}_{\text{imp}}$ and the aligned geometric cues g_t in $\mathcal{M}_{\text{exp}}$ with the same timestamp.

Geometry-gated writing. Adjacent frames are often redundant, so writing every frame to memory is inefficient. We gate memory writes using geometric changes predicted by the geometry encoder. Let T_t be the camera pose, and let (p_t, v_t) denote the camera position and viewing direction derived from T_t . We compute inter-frame motion and viewpoint changes:

$$\Delta p_t = \|p_t - p_{t-1}\|_2, \quad \Delta \theta_t = \angle(v_t, v_{t-1}), \quad (4)$$

and write an entry only when new spatial evidence is likely to appear:

$$\text{WRITE}(t) = \mathbb{I}[\Delta p_t > \tau_p \vee \Delta \theta_t > \tau_\theta]. \quad (5)$$

This geometry-gated policy suppresses redundant writes under smooth viewpoint changes while preserving informative transitions.

Geometry-consistent fusion. When $\text{WRITE}(t) = 1$, we further consolidate entries that are geometrically consistent. We define a geometric neighborhood for entry t :

$$\mathcal{N}(t) = \{j : \|p_t - p_j\| < r, \angle(v_t, v_j) < \theta\}. \quad (6)$$

If $\mathcal{N}(t) \neq \emptyset$, we select a fusion target based on semantic similarity in the evidence space:

$$j^* = \arg \max_{j \in \mathcal{N}(t)} \cos(k_t, k_j), \quad k_t = W_k e_t, \quad (7)$$

where k_t is a key projected from the evidence entry e_t . We then fuse e_t into e_{j^*} ; otherwise, we append (e_t, g_t) as a new entry. This step reduces repeated evidence while maintaining a geometry-consistent scene memory.

Geometry-filtered retrieval. Given a query q , we obtain a text embedding q_e via a text encoder E_{text} and perform coarse-to-fine retrieval. We first retrieve candidate chunks based on semantic relevance:

$$s_c(j) = \cos(q_e, k_j^c), \quad C_{\text{topk}} = \text{TOPK}(\{s_c(j)\}, K_c), \quad (8)$$

which prioritizes recall at a low cost.

Within candidate chunks, we refine frame-level selection by combining semantic relevance with a lightweight geometry term:

$$s_f(i) = \alpha \cos(q_e, k_i^f) + \beta \cos(Q_{\text{topo}}, e_i^{\text{dir}}), \quad (9)$$

where Q_{topo} is a direction-oriented embedding derived from q , and e_i^{dir} encodes the viewing direction derived from T_i . We select top- K_f frames F_{topk} according to $s_f(i)$ and gather the corresponding evidence.

In addition to the retrieved evidence, we construct lightweight summary tokens to provide global topological and metric context. Given q_e , we derive two query vectors:

$$Q_{\text{topo}} = W_{\text{topo}}q_e, \quad Q_{\text{metric}} = W_{\text{metric}}q_e, \quad (10)$$

where Q_{topo} emphasizes topological cues such as relative direction or connectivity, and Q_{metric} emphasizes metric cues such as distance, size, and ordering. We compute summary tokens via query-conditioned aggregation:

$$z_{\text{topo}} = \text{CA}(Q_{\text{topo}}, \mathcal{B}_{\text{topo}}), \quad z_{\text{metric}} = \text{CA}(Q_{\text{metric}}, \mathcal{B}_{\text{metric}}), \quad (11)$$

where $\text{CA}(\cdot)$ denotes a lightweight query-conditioned aggregation operator. $\mathcal{B}_{\text{topo}}$ contains values from $\mathcal{M}_{\text{imp}}$, and $\mathcal{B}_{\text{metric}}$ contains embedded geometric cues from $\mathcal{M}_{\text{exp}}$.

Finally, we form the context for the video language model by concatenating the query token, summary tokens, retrieved evidence tokens, and aligned geometric cues:

$$Z = \text{Concat}([\text{TOK}(q)], z_{\text{topo}}, z_{\text{metric}}, \{z_i\}_{i \in F_{\text{topk}}}, \{g_i\}_{i \in F_{\text{topk}}}), \quad (12)$$

where $\{z_i\}$ are evidence tokens from $\mathcal{M}_{\text{imp}}$, and $\{g_i\}$ are aligned geometric cues from $\mathcal{M}_{\text{exp}}$. In practice, geometry filtering can be implemented by down-weighting inconsistent entries during scoring or by masking them before concatenation. The resulting sequence Z is projected into the language model embedding space and fed into the video language model for prediction.

3.4 Unified Consistency Self-Supervised Reinforcement Learning

While geometry-consistent memory organizes evidence efficiently, the model may still produce different answers when the same question is observed from different temporal windows or retrieved evidence sets. Such inconsistency indicates brittle spatial grounding under viewpoint changes. We therefore optimize *cross-view geometric consistency* after supervised fine-tuning, where SFT provides task-aligned initialization and UC-SSRL further refines viewpoint stability. Different from Spatial-SSRL [25], which relies on intrinsic pretext tasks, our objective derives self-supervised rewards from multi-view consistency in video spatial reasoning: if the underlying 3D relations are stable, predictions from different views should agree.

Two-view sampler. Given a training sample (V, q) , we construct two observations for the same question, e.g., two temporal windows, two retrieval contexts, or two evidence sets:

$$o_i = \mathcal{S}(V, q; \xi_i), \quad o_j = \mathcal{S}(V, q; \xi_j), \quad (13)$$

where $\mathcal{S}(\cdot)$ denotes the two-view sampler and ξ_i, ξ_j are stochastic sampling choices. Feeding (o_i, q) and (o_j, q) into the model yields two predicted answer distributions π_i, π_j .

Unified consistency rewards. We design three complementary self-supervised rewards to refine answer-level, metric, and relational stability across paired views.

Two-view answer consistency. We encourage agreement between the two predicted answer distributions by penalizing their divergence:

$$R_{\text{KL}} = -\alpha \text{KL}(\pi_i \parallel \pi_j), \quad (14)$$

which stabilizes answers across different views.

Metric consistency. For metric-related questions, such as distance and size, let $\hat{d}_i$ and $\hat{d}_j$ denote the distance estimates inferred from the two views. We enforce consistency via

$$R_{\text{metric}} = -\eta |\hat{d}_i - \hat{d}_j|. \quad (15)$$

Topological consistency. For relational reasoning, let P_i and P_j denote normalized relational distributions derived from the two views. We encourage symmetric agreement:

$$R_{\text{topo}} = -\beta (\text{KL}(P_i \parallel P_j) + \text{KL}(P_j \parallel P_i)). \quad (16)$$

We combine them into a unified reward:

$$R = R_{\text{KL}} + R_{\text{metric}} + R_{\text{topo}}, \quad (17)$$

and optimize the model with a policy-gradient style objective. UC-SSRL is initialized from the SFT model and refines consistency without additional human annotations; in our implementation, only LoRA parameters are updated (Fig. 2), making it lightweight and complementary to memory-based evidence organization.

4 Experiments

4.1 Evaluation Benchmarks

VSI-Bench. VSI-Bench [45] evaluates visual-spatial intelligence from indoor-scene videos, requiring models to infer spatial relations and measurements from continuous observations. It contains over 5,000 question-answer pairs from 288 real-world indoor videos, covering configurational, measurement-estimation, and spatiotemporal tasks. Following its protocol, we report **Accuracy** for multiple-choice answer (MCA) tasks and **Mean Relative Accuracy (MRA)** for numerical-answer (NA) tasks. For prediction $\hat{y}$ and ground truth y , MRA is computed over tolerance thresholds $\mathcal{C} = \{0.50, 0.55, \dots, 0.95\}$:

$$\text{MRA} = \frac{1}{|\mathcal{C}|} \sum_{\theta \in \mathcal{C}} \mathbb{I} \left(\frac{|\hat{y} - y|}{y} < 1 - \theta \right). \quad (18)$$

OSI-Bench. OSI-Bench [44] targets open-world spatial reasoning from pedestrian-perspective videos with metrically precise 3D ground truth from synchronized

Table 1: VSI-Bench results. “Avg.” denotes the overall mean. **Bold** and underline denote the best and second-best results.

Method	Avg.	Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
		Numerical (MRA)				Multiple-Choice (Acc.)			
<i>Proprietary Models (API)</i>									
GPT-4o	34.0	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5
Gemini-1.5 Flash	42.1	49.8	30.8	53.5	54.4	37.7	41.0	31.5	37.8
Gemini-1.5 Pro	45.4	56.2	30.9	64.1	43.6	51.3	46.3	36.0	34.6
<i>Open-source Models</i>									
LLaVA-Video-7B	35.6	48.5	14.0	47.8	24.2	43.5	42.4	34.0	30.6
LLaVA-Video-72B	40.9	48.9	22.8	57.4	35.3	42.4	36.7	35.0	48.6
LLaVA-OneVision-7B	32.4	47.7	20.2	47.4	12.3	42.5	35.2	29.4	24.4
Qwen2.5-VL-7B	33.0	40.9	14.8	43.4	10.7	38.6	38.5	33.0	29.8
LLaVA-OneVision-72B	40.2	43.5	23.9	57.6	37.5	42.5	39.9	32.5	44.6
Qwen3-VL-8B-Instruct	57.4	66.7	38.9	74.1	53.5	59.0	65.4	32.0	69.4
<i>Specialized Models</i>									
Spacer [26]	45.5	57.8	28.2	59.9	47.1	40.1	45.4	33.5	52.1
ViLaSR [43]	45.4	63.5	34.4	60.6	30.9	48.9	45.2	30.4	49.2
Spatial-MLLM [42]	48.4	65.3	34.8	63.1	45.1	41.3	46.2	33.5	46.3
GeoThinker [20]	50.5	69.5	38.5	57.9	62.2	45.2	46.2	31.4	52.6
VLM-3R [7]	60.9	70.2	49.4	69.2	67.1	65.4	80.5	45.4	40.1
Map2Thought [8]	61.0	70.8	55.0	70.1	69.4	56.9	69.8	38.1	57.4
Cambrian-S-7B [46]	67.5	73.2	50.5	74.9	72.2	<u>71.1</u>	76.2	41.8	80.1
SpaceMind [50]	69.6	<u>73.3</u>	61.4	77.3	<u>74.2</u>	67.2	<u>88.4</u>	44.3	70.6
ConsiSpace _{SFT}	<u>71.2</u>	71.0	<u>73.8</u>	<u>77.7</u>	71.7	68.5	83.6	<u>49.8</u>	73.8
ConsiSpace _{UC-SSRL}	76.6	77.9	78.9	82.2	76.3	72.1	90.1	54.7	<u>79.9</u>

multi-sensor capture. It includes both qualitative relational reasoning and quantitative metric or kinematic understanding, making it suitable for testing whether spatial reasoning gains transfer beyond indoor benchmarks. Following its protocol, we report **Accuracy** for MCA tasks and **MRA** for NA tasks, where MRA checks whether the relative error falls below a set of tolerances with a small-value floor for near-zero targets.

MMSI-Video-Bench. MMSI-Video-Bench [23] evaluates video-based spatial intelligence with fully human-annotated questions from diverse video sources. It contains 1,106 questions over 1,278 clips and covers Perception, Planning, Prediction, and Cross-Video Reasoning. Following the official protocol, we report exact-match accuracy under two input settings: **Uniform-50**, which feeds exactly 50 uniformly sampled frames per clip, and **Sufficient-Coverage**, which uses the full set of frames used during human annotation. These settings evaluate spatial reasoning under a fixed frame budget and under evidence-rich inputs, respectively.

4.2 Datasets

Supervised fine-tuning. For SFT, we fine-tune the base video language model on a mixture of indoor and outdoor video spatial reasoning data. For indoor spatial reasoning, we use VSI-590K [46] as the primary instruction-tuning source. To improve outdoor coverage, we additionally construct nuScenes-10K, an outdoor synthetic set of 10K QA instances from nuScenes [2], using an automated

Table 2: OSI-Bench results. “Avg.” denotes the overall mean. **Bold** and underline denote the best and second-best results.

Method	Avg.	Rel. Dis.	Rel. Dir.	Qual. EM	Obj. Loc.	Abs. Dis.	Depth Count	Abs. Displ.	Abs. Speed	Quan. EM
Proprietary Models (API)		Relational (MCA)			Static Metric (NA)			Dynamic Metric (NA)		
GPT-5	29.7	34.4	33.1	49.5	32.5	23.7	20.9	10.5	33.8	30.6
Gemini-2.5-Pro	37.2	50.0	28.1	52.5	37.4	28.1	37.9	26.8	31.1	40.8
Open-source Models										
LLaVA-Video-Qwen2-7B	22.9	37.1	31.2	40.9	17.6	22.1	18.2	17.5	19.0	5.7
LLaVA-OneVision-7B	25.7	35.1	32.7	40.8	16.1	25.5	25.6	16.8	28.3	13.4
Qwen3-VL-8B-Instruct	31.2	38.3	31.2	49.3	21.0	15.1	33.3	21.3	34.3	37.8
Specialized Models										
Spacer [26]	30.2	35.0	24.6	39.0	36.7	14.0	36.4	24.0	28.4	34.1
ViLaSR [43]	30.4	42.6	24.5	39.0	40.3	17.1	36.8	15.7	25.8	32.2
Spatial-MLLM [42]	31.3	36.0	25.0	41.5	41.5	17.3	38.3	23.0	28.4	30.3
VLM-3R [7]	40.3	57.0	43.6	52.3	44.6	24.6	42.0	34.2	38.5	26.2
ConsiSpace _{SFT}	<u>45.8</u>	<u>58.1</u>	<u>44.2</u>	<u>60.1</u>	44.4	<u>35.5</u>	<u>46.0</u>	<u>34.7</u>	<u>41.7</u>	<u>47.7</u>
ConsiSpace _{UC-SSRL}	53.0	66.5	51.3	68.7	51.6	41.7	52.7	41.8	48.5	54.4

and programmatically verifiable pipeline based on ego poses and 3D annotations. We apply geometry-validity filtering, answer parseability checking, consistency checking, MLLM verification, and de-duplication to 22,518 generated candidates, yielding 10,000 QA pairs. nuScenes-10K has no shared videos or annotations with OSI-Bench, except for generic object-category overlap. Manual verification on 1,000 randomly sampled QA pairs shows 91.6% correctness. Additional details and verification statistics are provided in *Supp. Mat. A*. SFT is performed with standard maximum-likelihood training to map the concatenated context to the final answer.

Unified consistency SSRL. UC-SSRL does not require additional human annotations beyond the SFT data. We reuse the same training pool and apply a two-view sampler to construct paired observations for each (V, q) , such as two temporal windows or two retrieval contexts. The model is then optimized with answer, metric, and topological consistency rewards across paired views. Unless otherwise stated, only lightweight LoRA parameters are updated during UC-SSRL, making it complementary to SFT and efficient to run.

4.3 Implementation Details

Architecture. ConsiSpace is built on **Qwen3-VL-8B-Instruct** [1] with LoRA adaptation and Geometry-Consistent Memory. We use frozen SigLIP2 [32] as the visual encoder and frozen VGGT [35] as the geometry encoder. The vision tower, aligner, and geometry encoder remain frozen, and only LoRA parameters are updated during both SFT and UC-SSRL.

Memory and retrieval. We use a two-bank Geometry-Consistent Memory with $M=64$ memory tokens and feature dimension 512. The memory is organized into 4 chunks with an 8-head resampler. Hierarchical retrieval uses top- $K=8$ at both chunk and frame levels, with 8 value tokens per chunk and 4 value tokens per frame. Detailed memory injection, gating, retrieval, and backend settings are provided in *Supp. Mat. B*.

Table 3: MMSI-Video-Bench results. “Avg.” denotes the overall mean. Attr., Inst., Cam., Scen., and Inter. denote Attribute, Instance, Camera, Scene, and Interaction, respectively. MU., and MV. represent Memory Update, and Multi-View Integration, respectively. **Bold** and underline denote the best and second-best results within each setting.

Method	Attr.	Inst.	Inst.-Inst.	Inst.-Scen.	Scen.-Scen.	Cam.-Inst.	Cam.-Scen.	Cam.	Inst.	Inter.	MU.	MV.	Plan.	Pred.	Avg.
(Sufficient-Coverage) Proprietary															
GPT-4o	22.9	30.3	35.1	26.1	26.6	27.5	24.7	28.9	24.7	34.3	25.7	29.8	26.8		28.1
Gemini 2.5 Flash-Thinking	43.4	38.2	41.6	24.6	34.2	36.2	44.1	34.4	35.8	36.3	38.6	36.3	31.7		36.7
(Sufficient-Coverage) Open-source															
Qwen2.5-VL-7B	31.3	21.1	19.5	30.4	19.0	31.2	35.5	36.7	29.6	27.4	30.0	26.6	35.4		28.8
Qwen3-VL-8B	27.8	31.5	33.3	26.9	23.1	22.4	27.3	29.6	41.0	28.4	29.9	26.5	32.0		29.1
(Sufficient-Coverage) Specialized Models															
Spacer [26]	35.3	29.6	39.2	27.9	35.1	27.8	36.4	28.2	33.0	35.9	30.8	28.0	27.3		31.8
ViLaSR [43]	35.2	29.5	39.1	27.8	35.0	27.7	36.3	28.1	32.9	35.8	30.7	27.9	27.2		31.7
Spatial-MLLM [42]	37.4	31.7	41.3	30.0	37.2	29.9	38.5	30.3	35.1	38.0	32.9	30.1	29.4		33.9
VLM-3R [7]	46.1	40.4	50.0	38.7	45.9	38.6	47.2	39.0	43.8	46.7	41.6	38.8	38.1		42.6
ConsiSpace ^{SFT}	<u>51.8</u>	<u>46.3</u>	<u>55.6</u>	<u>44.6</u>	<u>51.6</u>	<u>44.5</u>	<u>52.9</u>	<u>44.9</u>	<u>49.6</u>	<u>52.4</u>	<u>47.4</u>	<u>44.7</u>	<u>44.0</u>		<u>48.4</u>
ConsiSpace ^{UC-SSRL}	61.4	55.2	65.7	53.3	61.1	53.7	62.2	54.0	58.8	62.0	56.8	53.4	52.9		57.5
(Uniform-50) Proprietary															
Gemini 2.5 Flash-Thinking	43.4	25.0	42.9	23.2	32.9	47.5	38.7	28.9	38.3	31.4	28.6	40.3	31.7		35.2
GPT-5	44.6	38.2	44.2	39.1	40.5	41.2	37.6	28.9	42.0	33.3	28.6	34.7	28.1		36.8
Gemini 3 Pro	44.6	39.5	44.2	33.3	29.1	43.8	35.5	40.0	38.3	34.3	37.1	38.7	35.4		38.0
(Uniform-50) Open-source															
LLaVA-Video-7B	27.7	28.9	28.6	30.4	19.0	25.0	31.2	35.6	30.9	36.3	20.0	21.0	35.4		28.5
Qwen2.5-VL-7B	26.5	25.0	29.9	34.8	20.2	37.5	33.3	31.1	34.6	24.5	22.9	34.7	28.1		29.7
Qwen3-VL-8B	33.7	25.0	24.7	26.1	24.1	30.0	35.5	22.2	29.6	31.4	20.0	25.0	29.3		27.6
(Uniform-50) Specialized Models															
Spacer [26]	36.4	31.1	36.3	30.9	32.3	39.5	30.5	32.2	34.0	28.1	29.2	32.1	27.5		32.3
ViLaSR [43]	36.3	31.0	36.2	30.8	32.2	39.4	30.4	32.1	33.9	28.0	29.1	32.0	27.4		32.2
Spatial-MLLM [42]	38.5	33.2	38.4	33.0	34.4	41.6	32.6	34.3	36.1	30.2	31.3	34.2	29.6		34.4
VLM-3R [7]	47.2	41.9	47.1	41.7	43.1	50.3	41.3	43.0	44.8	38.9	40.0	42.9	38.3		43.1
ConsiSpace ^{SFT}	<u>52.9</u>	<u>47.7</u>	<u>52.8</u>	<u>47.5</u>	<u>48.9</u>	<u>55.9</u>	<u>47.1</u>	<u>48.8</u>	<u>50.5</u>	<u>44.8</u>	<u>45.9</u>	<u>48.7</u>	<u>44.2</u>		<u>48.9</u>
ConsiSpace ^{UC-SSRL}	62.6	56.7	62.5	56.8	58.0	65.8	56.2	58.2	59.9	54.0	55.0	58.1	53.2		58.1

Training details. We first fine-tune on VSI-590K and nuScenes-10K with standard maximum-likelihood training. We use LoRA with rank 16 and $\alpha=32$, bfloat16 precision, and DeepSpeed ZeRO-3. The trainable parameters account for only a small fraction of the full model parameters, since the base model and encoders are frozen. We train for 200 steps with learning rate 1×10^{-4} , per-device batch size 1, gradient accumulation 2, and maximum sequence length 5,120. We then initialize UC-SSRL from the SFT checkpoint, keep the same optimization setting for controlled comparison, and set the overall UC-SSRL weight to 10^{-2} . All experiments are conducted on 8×A100 80GB GPUs. Additional optimizer, LoRA, and runtime details are provided in *Supp. Mat. B*

4.4 Main Results

VSI-Bench. As shown in Tab. 1, ConsiSpace achieves 76.6 Avg. with UC-SSRL, outperforming SpaceMind (69.6 Avg.) by +7.0 points and Qwen3-VL-8B-Instruct (57.4 Avg.) by +19.2 points. The gains cover both numerical and multiple-choice tasks, showing improved measurement and relational reasoning.

OSI-Bench. Tab. 2 shows that ConsiSpace reaches 53.0 Avg. with UC-SSRL on OSI-Bench. It improves over VLM-3R (40.3 Avg.) by +12.7 points and Qwen3-VL-8B-Instruct (31.2 Avg.) by +21.8 points. UC-SSRL further improves over SFT by +7.2 points, especially on metric-heavy categories, suggesting better stability under viewpoint changes.

Table 4: End-to-end efficiency and long-video scaling. All results are measured on the same A100 80GB GPU under different input budgets.

Method	Dur.	Frames	Mem. (GB)↓	Inf. (s)↓	VSI↑	OSI↑	MMSI-SC↑	MMSI-U50↑
Qwen3-VL-8B	~10-15s	50	19.4	1.32	57.4	31.2	29.1	27.6
VLM-3R	~10-15s	50	24.8	2.10	60.9	40.3	42.6	43.1
ConsiSpace _{UC-SSRL}	~10-15s	50	22.0	1.65	76.6	53.0	57.5	58.1
Qwen3-VL-8B	~20-30s	100	36.5	3.05	58.2	32.1	29.8	28.5
VLM-3R	~20-30s	100	42.0	4.20	68.0	42.0	48.2	49.1
ConsiSpace _{UC-SSRL}	~20-30s	100	28.5	2.25	71.6	45.8	53.1	53.8
Qwen3-VL-8B	~120-180s	200	71.2	7.45	57.8	31.5	29.3	27.9
VLM-3R	~120-180s	200	78.5	10.80	69.5	43.5	51.1	52.3
ConsiSpace _{UC-SSRL}	~120-180s	200	35.0	3.10	72.4	46.8	54.4	55.1

MMSI-Video-Bench. As shown in Tab. 3, ConsiSpace achieves 57.5 Avg. under *Sufficient-Coverage* and 58.1 Avg. under *Uniform-50*, outperforming VLM-3R by +14.9 and +15.0 points, respectively. UC-SSRL consistently improves over SFT in both settings, indicating that consistency learning benefits both evidence-rich and fixed-budget inputs.

4.5 Efficiency and Scalability Analysis

Beyond accuracy, we evaluate peak GPU memory, inference time, and accuracy under increasing input budgets. All measurements are conducted on the same A100 80GB GPU, as shown in Tab. 4. At the 50-frame setting, ConsiSpace achieves substantially higher accuracy than Qwen3-VL-8B and VLM-3R, while requiring less inference time than VLM-3R. As the input budget increases, VLM-3R accumulates dense evidence and its memory and latency grow rapidly. At 200 frames, VLM-3R requires 78.5GB memory and 10.80s inference time, whereas ConsiSpace uses only 35.0GB and 3.10s while achieving higher accuracy on VSI, OSI, and MMSI. These results show that Geometry-Consistent Memory improves end-to-end long-video scalability, not only retained-entry counts, by selectively organizing spatial evidence instead of densely retaining all observations.

4.6 Navigation-Oriented Validation

To connect our video spatial reasoning setting with the navigation-oriented motivation, we evaluate Habitat ObjectNav next-waypoint prediction. Given ego-centric history and a navigation query, the model predicts the next direction, testing navigation-oriented spatial perception rather than end-to-end embodied control. Tab. 5 shows that ConsiSpace outperforms Qwen3-VL-8B, with further gains from UC-SSRL.

Table 5: ObjectNav next-waypoint. Acc., Succ., and Dist. denote direction accuracy, distance-reduction success, and remaining distance.

Method	Acc.↑	Succ.↑	Dist.↓
Qwen3-VL-8B	42.6	31.8	4.72
ConsiSpace _{SFT}	47.9	36.5	4.31
ConsiSpace _{UC-SSRL}	50.8	39.7	4.05

Table 6: Ablation on GCM. MMSI-Video is averaged over *Sufficient-Coverage* and *Uniform-50*. #Entries denotes normalized retained memory entries.

Variant	VSI↑	OSI↑	MMSI-Video↑	#Entries↓
Full (ConsiSpace_{SFT})	71.2	45.8	52.3	1.0×
w/o GateWrite	69.8	44.5	47.2	2.3×
w/o Fusion	70.3	45.0	47.9	1.8×
w/o GeoFilter	69.5	43.9	46.0	1.0×

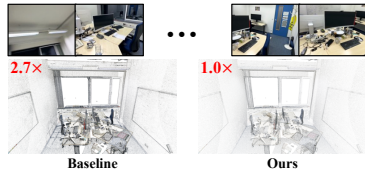


Fig. 4: Qualitative effect of geometry-gated writing and fusion. Dense updates introduce redundant points, while GCM produces cleaner and more compact point clouds. Red numbers denote normalized retained entries.

4.7 Ablation Study

Effect of geometry-consistent memory. Tab. 6 and Fig. 4 evaluate the GCM evidence lifecycle. Removing geometry-gated writing increases retained entries to $2.3\times$, while removing geometry-consistent fusion increases them to $1.8\times$; both variants reduce accuracy, showing the importance of suppressing and merging redundant observations. Removing geometry-filtered retrieval keeps the entry count unchanged but degrades OSI and MMSI-Video, indicating that retrieval quality depends on filtering geometrically inconsistent evidence under viewpoint changes. Qualitatively, GCM produces cleaner and more compact point clouds than dense updates. Overall, writing, fusion, and retrieval are complementary for compact and reliable spatial evidence organization.

Effect of the UC-SSRL. Fig. 5 evaluates UC-SSRL starting from $\text{ConsiSpace}_{\text{SFT}}$. UC-SSRL consistently improves all benchmarks, with larger gains on OSI-Bench and MMSI-Video where viewpoint changes and metric relations are more challenging. Since UC-SSRL is initialized from the SFT model, these gains indicate that consistency rewards refine viewpoint stability rather than learning task correctness from scratch.

Table 7: UC-SSRL reward ablation. We ablate answer-level, metric, and topological consistency rewards.

R_{KL}	R_{metric}	R_{topo}	VSI $\uparrow$	OSI $\uparrow$	MMSI-SC $\uparrow$	MMSI-U50 $\uparrow$
$\times$	$\times$	$\times$	71.2	45.8	48.4	48.9
$\checkmark$	$\times$	$\times$	73.5	48.2	52.1	52.6
$\times$	$\checkmark$	$\times$	72.8	49.0	51.4	51.9
$\times$	$\times$	$\checkmark$	72.9	47.1	51.7	52.2
$\checkmark$	$\checkmark$	$\times$	75.1	51.6	55.4	56.0
$\checkmark$	$\times$	$\checkmark$	75.4	50.3	55.7	56.4
$\times$	$\checkmark$	$\checkmark$	74.5	50.8	54.8	55.5
$\checkmark$	$\checkmark$	$\checkmark$	76.6	53.0	57.5	58.1

UC-SSRL reward ablation. Tab. 7 ablates the three UC-SSRL rewards. Each reward improves over SFT, pairwise combinations bring further gains, and the full objective performs best across VSI, OSI, and MMSI-Video. This shows that answer-level, metric, and topological consistency are complementary, and that UC-SSRL refines cross-view stability from an SFT-initialized model rather than learning correctness from consistency alone.

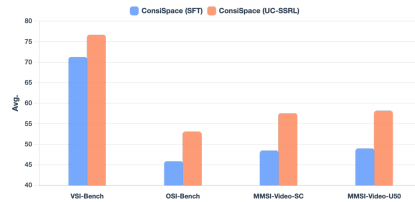


Fig. 5: Effect of UC-SSRL. UC-SSRL improves over the SFT baseline across benchmarks, with larger gains on OSI-Bench and MMSI-Video where viewpoint changes and metric relations are more challenging.

Table 8: Robustness to noisy VGGT geometry. We perturb VGGT geometry before GCM operations.

Method	Noise	VSI $\uparrow$	OSI $\uparrow$	SC $\uparrow$	U50 $\uparrow$
GeoThinker	Clean	72.6	44.5	48.2	49.1
ConsiSpace _{UC-SSRL}	Clean	76.6	53.0	57.5	58.1
GeoThinker	10%	70.8	42.1	45.3	46.5
ConsiSpace _{UC-SSRL}	10%	75.9	52.4	56.8	57.4
GeoThinker	20%	68.4	38.8	41.6	42.8
ConsiSpace _{UC-SSRL}	20%	74.9	51.3	55.6	56.1
GeoThinker	30%	65.2	35.1	38.2	39.4
ConsiSpace _{UC-SSRL}	30%	73.1	49.6	53.4	54.3

Robustness to noisy geometry. Tab. 8 evaluates robustness to perturbed VGGT geometry. We add noise to depth, translation, and rotation cues before GCM operations while keeping RGB inputs and questions unchanged. Performance degrades gradually as noise increases, but ConsiSpace remains consistently stronger than GeoThinker, suggesting that GCM benefits from geometry cues without requiring perfect geometry estimates.

Summary vs. Top- K readout.

Tab. 9 ablates the readout strategy from the Geometry-Consistent Memory. Summary-only readout lacks fine-grained local evidence, while Top- K -only readout misses global spatial context. Combining summary tokens with Top- K retrieval gives the best overall performance, showing that global summarization and sparse evidence selection are complementary.

Table 10: Sensitivity to writing thresholds. VSI and OSI averages of ConsiSpace_{SFT} under different (τ_p, τ_θ) . The default setting is $(0.15\text{ m}, 15^\circ)$.

τ_p (m)	τ_θ (deg)	VSI $\uparrow$	OSI $\uparrow$
0.10	15	71.0	45.6
0.15	15	71.2	45.8
0.25	15	70.9	45.3
0.15	10	71.1	45.7
0.15	15	71.2	45.8
0.15	30	70.8	45.4

Table 9: Readout strategy ablation.

We compare summary-only, Top- K -only, and their combination under the same backbone and memory budget.

Variant	VSI $\uparrow$	OSI $\uparrow$	MMSI-SC $\uparrow$	MMSI-U50 $\uparrow$
Summary + Top- K	71.2	45.8	51.8	52.9
Summary-only	69.8	43.9	48.7	47.1
Top- K -only	70.9	46.2	49.3	50.9

Table 11: Sensitivity to fusion thresholds.

VSI and OSI averages under different fusion neighborhoods (r, θ) . #Entries denotes retained entries normalized by the default $(0.40\text{ m}, 30^\circ)$.

r (m)	θ (deg)	VSI $\uparrow$	OSI $\uparrow$	#Entries $\downarrow$
0.20	30	71.0	45.5	1.18 $\times$
0.40	30	71.2	45.8	1.00 $\times$
0.60	30	70.9	45.4	0.92 $\times$
0.40	15	71.1	45.6	1.10 $\times$
0.40	30	71.2	45.8	1.00 $\times$
0.40	45	70.8	45.3	0.94 $\times$

Threshold sensitivity. Tab. 10 and Tab. 11 evaluate the sensitivity of writing thresholds (τ_p, τ_θ) and fusion neighborhood thresholds (r, θ) . Around the default settings, VSI and OSI vary only slightly, indicating that both geometry-gated writing and geometry-consistent fusion are robust to threshold choices. For fusion, the retained entries change as expected: smaller neighborhoods retain more entries, while larger neighborhoods enable more aggressive fusion. Overall, ConsiSpace exhibits a broad stable regime and does not rely on finely tuned geometric thresholds.

4.8 Qualitative Results

Geometry-filtered retrieval. Fig. 6 visualizes the two-stage retrieval process. After chunk-level semantic retrieval, GCM performs frame-level selection with a direction-consistency cue. Frames that match the query semantics but conflict with the inferred viewpoint are suppressed.

Consistency under revisit views. Fig. 7 shows a revisit-view example for UC-SSRL. Given the same video and query, baseline models produce different answers across nearby viewpoints, indicating unstable spatial grounding. In contrast, ConsiSpace with UC-SSRL gives consistent predictions across View A

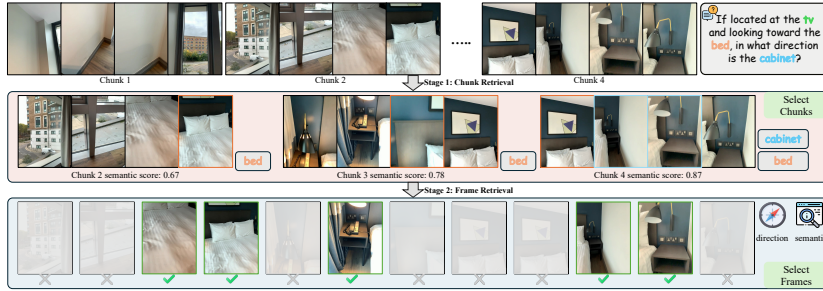


Fig. 6: Qualitative visualization of geometry-filtered retrieval. GCM first retrieves semantically relevant chunks and then selects frames using direction consistency. Gray $\times$ denotes suppressed frames, and green $\checkmark$ denotes selected evidence.

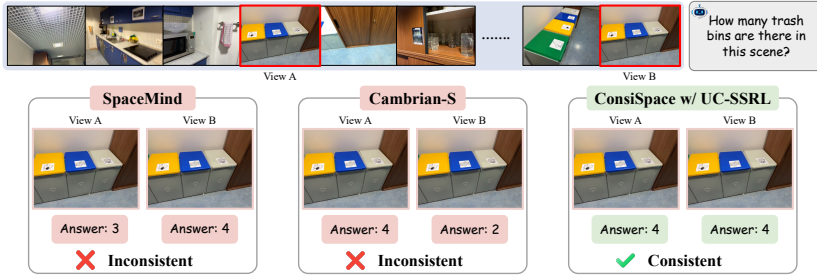


Fig. 7: Consistency under revisit views. Given the same query, baseline models produce inconsistent counting answers across two revisited viewpoints, while ConsiSpace with UC-SSRL gives consistent predictions.

and View B. This suggests that UC-SSRL refines viewpoint stability after SFT, rather than learning task correctness from consistency alone.

Additional visualizations. We provide more qualitative examples on VSI-Bench, MMSI-Video-Bench, and OSI-Bench in the *Supp. Mat. C*.

5 Conclusion

In this paper, we propose **ConsiSpace**, a geometry-consistent framework for video spatial reasoning. ConsiSpace builds a Geometry-Consistent Memory on top of generic video encoders and VLMs, and manages evidence via geometry-gated writing, geometry-consistent fusion, and geometry-filtered retrieval to reduce redundancy and retrieve query-relevant spatial cues. We further introduce a Unified Consistency SSRL objective that encourages cross-view spatial agreement after supervised fine-tuning, improving viewpoint stability without additional human annotations. Experiments on three challenging video spatial reasoning benchmarks show that ConsiSpace consistently outperforms strong prior methods. We hope our results highlight geometric consistency as a simple and effective principle for robust video spatial reasoning.

Acknowledgements. This work was supported by the NSFC under Grant No. 62376121, Basic Research Program of Jiangsu under Grant No. BK20251999, Gusu Innovation Leading Talent Program under Grant No. ZXL2025319, and Jiangsu Provincial Science & Technology Major Project under Grant No. BG2024042.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
2. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nusenes: A multimodal dataset for autonomous driving. arXiv preprint arXiv:1903.11027 (2019)
3. Chen, G., Liu, Y., Huang, Y., Pei, B., Xu, J., He, Y., Lu, T., Wang, Y., Wang, L.: CG-bench: Clue-grounded question answering benchmark for long video understanding. In: The Thirteenth International Conference on Learning Representations (2025)
4. Chen, Y., Qi, Z., Zhang, W., Jin, X., Zhang, L., Liu, P.: Reasoning in space via grounding in the world. In: The Fourteenth International Conference on Learning Representations (2026)
5. Chen, Y., Xue, F., Li, D., Hu, Q., Zhu, L., Li, X., Fang, Y., Tang, H., Yang, S., Liu, Z., et al.: Longvila: Scaling long-context visual language models for long videos. In: International Conference on Learning Representations. vol. 2025, pp. 18227–18246 (2025)
6. Dongfang, Z., Zheng, X., Weng, Z., Lyu, Y., Paudel, D.P., Van Gool, L., Yang, K., Hu, X.: Are multimodal large language models ready for omnidirectional spatial reasoning? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9759–9769 (2026)
7. Fan, Z., Zhang, J., Li, R., Zhang, J., Chen, R., Hu, H., Wang, K., Qu, H., Wang, D., Yan, Z., et al.: Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. arXiv preprint arXiv:2505.20279 (2025)
8. Gao, X., Zhang, Z., Chen, D.Z., Xu, S., Quan, L., Pérez-Pellitero, E., Jang, Y.: Map2thought: Explicit 3d spatial reasoning via metric cognitive maps. arXiv preprint arXiv:2601.11442 (2026)
9. Hu, W., Lin, J., Long, Y., Ran, Y., Jiang, L., Wang, Y., Zhu, C., Xu, R., Wang, T., Pang, J.: G²vlm: Geometry grounded vision language model with unified 3d reconstruction and spatial reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9535–9546 (2026)
10. Huang, T., Li, D., Yang, R., Zhang, Z., Yang, Z., Tang, H.: Mobilevla-r1: Reinforcing vision-language-action for mobile robots. arXiv preprint arXiv:2511.17889 (2025)
11. Huang, T., Zhang, Z., Tang, H.: 3d-r1: Enhancing reasoning in 3d vlms for unified scene understanding. arXiv preprint arXiv:2507.23478 (2025)
12. Huang, T., Zhang, Z., Wang, Y., Tang, H.: 3d coca: Contrastive learners are 3d captioners. In: Thirteenth International Conference on 3D Vision (2026)
13. Huang, T., Zhang, Z., Zhang, R., Zhao, Y.: Dc-scene: Data-centric learning for 3d scene understanding. arXiv preprint arXiv:2505.15232 (2025)
14. Huang, X., Wu, J., Xie, Q., Han, K.: 3drs: Mllms need 3d-aware representation supervision for scene understanding. arXiv preprint arXiv:2506.01946 (2025)
15. Ji, Y., Tan, H., Shi, J., Hao, X., Zhang, Y., Zhang, H., Wang, P., Zhao, M., Mu, Y., An, P., et al.: Robobrain: A unified brain model for robotic manipulation from abstract to concrete. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1724–1734 (2025)
16. Krantz, J., Wijmans, E., Majumdar, A., Batra, D., Lee, S.: Beyond the nav-graph: Vision-and-language navigation in continuous environments. In: European Conference on Computer Vision. pp. 104–120. Springer (2020)

17. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: European conference on computer vision. pp. 71–91. Springer (2024)
18. Li, D., Li, H., Wang, Z., Yan, Y., Zhang, H., Chen, S., Hou, G., Jiang, S., Zhang, W., Shen, Y., et al.: Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. arXiv preprint arXiv:2505.21500 (2025)
19. Li, F., Song, W., Zhao, H., Wang, J., Ding, P., Wang, D., Zeng, L., Li, H.: Spatial forcing: Implicit spatial representation alignment for vision-language-action model. arXiv preprint arXiv:2510.12276 (2025)
20. Li, H., Cao, Q., Tang, T., Xiang, K., Guo, Z., Han, J., Xu, H., Liang, X.: Thinking with geometry: Active geometry integration for spatial reasoning. arXiv preprint arXiv:2602.06037 (2026)
21. Li, H., Li, D., Wang, Z., Yan, Y., Wu, H., Zhang, W., Shen, Y., Lu, W., Xiao, J., Zhuang, Y.: Spatialladder: Progressive training for spatial reasoning in vision-language models. arXiv preprint arXiv:2510.08531 (2025)
22. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
23. Lin, J., Xu, R., Zhu, S., Yang, S., Cao, P., Ran, Y., Hu, M., Zhu, C., Xie, Y., Long, Y., et al.: Mmsi-video-bench: A holistic benchmark for video-based spatial intelligence. arXiv preprint arXiv:2512.10863 (2025)
24. Liu, Q., Huang, T., Zhang, Z., Tang, H.: Nav-r1: Reasoning and navigation in embodied scenes. arXiv preprint arXiv:2509.10884 (2025)
25. Liu, Y., Zhang, B., Zang, Y., Cao, Y., Xing, L., Dong, X., Duan, H., Lin, D., Wang, J.: Spatial-ssrl: Enhancing spatial understanding via self-supervised reinforcement learning. arXiv preprint arXiv:2510.27606 (2025)
26. Ouyang, K., Liu, Y., Wu, H., Liu, Y., Zhou, H., Zhou, J., Meng, F., Sun, X.: Spacer: Reinforcing mllms in video spatial reasoning. arXiv preprint arXiv:2504.01805 (2025)
27. Ramakrishnan, S.K., Gokaslan, A., Wijmans, E., Maksymets, O., Clegg, A., Turner, J., Undersander, E., Galuba, W., Westbury, A., Chang, A.X., et al.: Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. arXiv preprint arXiv:2109.08238 (2021)
28. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., et al.: Habitat: A platform for embodied ai research. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9339–9347 (2019)
29. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)
30. Tang, H., Huang, T., Zhang, Z.: 3d coca v2: Contrastive learners with test-time search for generalizable spatial intelligence. arXiv preprint arXiv:2601.06496 (2026)
31. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
32. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
33. Wang, C., Huang, T., Sun, C., Ning, X., Wang, D., Tang, H.: Let geometry guide: Layer-wise unrolling of geometric priors in multimodal llms. arXiv preprint arXiv:2604.05695 (2026)

34. Wang, H., Zhao, Y., Wang, T., Fan, H., Zhang, X., Zhang, Z.: Ross3d: Reconstructive visual instruction tuning with 3d-awareness. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9275–9286 (2025)
35. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025)
36. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20697–20709 (2024)
37. Wang, W., He, Z., Hong, W., Cheng, Y., Zhang, X., Qi, J., Ding, M., Gu, X., Huang, S., Xu, B., et al.: Lvbench: An extreme long video understanding benchmark. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22958–22967 (2025)
38. Wang, X., Zhang, Y., Zohar, O., Yeung-Levy, S.: Videoagent: Long-form video understanding with large language model as agent. In: European Conference on Computer Vision. pp. 58–76. Springer (2024)
39. Wang, X., Wang, C., Xu, Y., Ye, M., Zhang, F.C., Tian, J., Zhan, X., Zhu, L., Lu, C., Yang, L.: Lamp: Learning vision-language-action policies with 3d scene flow as latent motion prior. arXiv preprint arXiv:2603.25399 (2026)
40. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3 : Permutation-equivariant visual geometry learning. In: The Fourteenth International Conference on Learning Representations (2026)
41. Wang, Y., Ke, L., Zhang, B., Qu, T., Yu, H., Huang, Z., Yu, M., Xu, D., Yu, D.: N3d-vlm: Native 3d grounding enables accurate spatial reasoning in vision-language models. arXiv preprint arXiv:2512.16561 (2025)
42. Wu, D., Liu, F., Hung, Y.H., Duan, Y.: Spatial-mlm: Boosting mllm capabilities in visual-based spatial intelligence. arXiv preprint arXiv:2505.23747 (2025)
43. Wu, J., Guan, J., Feng, K., Liu, Q., Wu, S., Wang, L., Wu, W., Tan, T.: Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. arXiv preprint arXiv:2506.09965 (2025)
44. Wu, M., Wang, Z., Wang, F., Yang, J., Pollefeys, M., Zhang, T.: From indoor to open world: Revealing the spatial reasoning gap in mllms. arXiv preprint arXiv:2512.19683 (2025)
45. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in space: How multimodal large language models see, remember, and recall spaces. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10632–10643 (2025)
46. Yang, S., Yang, J., Huang, P., Brown, E., Yang, Z., Yu, Y., Tong, S., Zheng, Z., Xu, Y., Wang, M., et al.: Cambrian-s: Towards spatial supersensing in video. arXiv preprint arXiv:2511.04670 (2025)
47. Zhang, J., Chen, Y., Xu, Y., Huang, Z., Mei, J., Chen, C., Zhou, Y., Yuan, Y.J., Cai, X., Huang, G., et al.: From flatland to space: Teaching vision-language models to perceive and reason in 3d. *Advances in Neural Information Processing Systems* **38** (2026)
48. Zhang, L., Wan, W., Huang, T., Fang, Z., Zhang, H.: Multigranularity-3dqa: Dynamic multi-granularity perception and gated dual-stage encoder-decoder for 3d question answering. *Expert Systems with Applications* p. 131676 (2026)
49. Zhang, Z., Wu, Y., Jia, L., Wang, Y., Zhang, Z., Li, Y., Ran, B., Zhang, F., Sun, Z., Yin, Z., et al.: Think3d: Thinking with space for spatial reasoning. arXiv preprint arXiv:2601.13029 (2026)

50. Zhao, R., Zhang, Z., Xu, J., Chang, J., Chen, D., Li, L., Sun, W., Wei, Z.: Space-mind: Camera-guided modality fusion for spatial reasoning in vision-language models. arXiv preprint arXiv:2511.23075 (2025)
51. Zheng, D., Huang, S., Li, Y., Wang, L.: Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. arXiv preprint arXiv:2505.24625 (2025)
52. Zheng, D., Huang, S., Wang, L.: Video-3d llm: Learning position-aware video representation for 3d scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8995–9006 (2025)
53. Zhi, H., Chen, P., Li, J., Ma, S., Sun, X., Xiang, T., Lei, Y., Tan, M., Gan, C.: Lscenellm: Enhancing large 3d scene understanding using adaptive visual preferences. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3761–3771 (2025)

A Outdoor Data Synthesis

This section describes our automated pipeline for synthesizing **nuScenes-10K**, an outdoor spatial-reasoning instruction set from nuScenes [2]. nuScenes-10K complements the indoor VSI-590K training data with outdoor geometry patterns, such as long-range depth, large-scale scenes, and dynamic agents, while keeping the supervision compatible with our video spatial reasoning formulation.

A.1 Data Sources and Representations

We build on nuScenes sequences with calibrated multi-camera imagery, ego poses, 3D object annotations, and optional LiDAR-based geometric verification. For each driving clip, we sample short temporal windows and form a multi-view video input $V = I_{t=1}^T$ using the same frame-budget setting as training. We reuse the same geometry extraction stack as in the main pipeline, including pose and depth cues, so that outdoor samples follow the same input interface as ConsiSpace.

A.2 Automated Instruction and Answer Generation

For each clip, we generate question-answer pairs covering three families:

- **Topological and relational:** relative direction, relative ordering, and ego-heading-conditioned spatial relations.
- **Metric:** approximate distance, size comparison, and nearest or farthest object queries, with numeric, binned, or multiple-choice answers.
- **Motion and interaction:** whether an agent approaches, recedes, crosses, or changes its relative state within the temporal window.

Questions are generated by templates with controlled lexical diversity. Answers are computed from calibrated geometry: we transform 3D object centers into the ego frame, derive relative directions by azimuth with respect to the ego forward axis, and compute metric quantities by Euclidean distance with optional depth verification. For multiple-choice questions, we construct fixed option sets and keep only samples with unambiguous labels.

A.3 Quality Control and Verification

We generate 22,518 candidate QA pairs and apply multi-stage filtering to improve label reliability. Specifically, we filter candidates by geometry validity, answer parseability, consistency checking, MLLM verification, and de-duplication, yielding 10,000 final QA pairs. Tab. 12 summarizes the filtering statistics.

We further check potential overlap between nuScenes-10K and OSI-Bench. nuScenes-10K uses autonomous-driving ego-pose and 3D-box annotations, while OSI-Bench uses pedestrian-view videos with independent 3D ground truth and QA protocols. We find no shared videos or annotations between the two datasets, with only generic object-category overlap.

Table 12: nuScenes-10K filtering statistics. Multi-stage filtering reduces 22,518 generated candidates to 10,000 final QA pairs.

Stage	#QA	Ret.
Raw candidates	22,518	100%
Geometry-valid	19,806	88.0%
Parseable	17,932	79.6%
Consistency checked	14,516	64.5%
MLLM pass	11,842	52.6%
De-duplicated	10,436	46.3%
Final	10,000	44.4%

Table 13: Human verification of nuScenes-10K. Manual verification on 1,000 random QA pairs shows 91.6% correctness.

Category	Ratio	Main reason
Correct	91.6%	Valid object/relation/answer
Ambiguous reference	3.4%	Multiple plausible objects
Occluded / distant	2.1%	Hard to verify visually
Annotation sparsity	1.7%	Missing/incomplete 3D box
Incorrect	1.2%	Wrong relation/metric answer

To estimate label quality, we manually verify 1,000 randomly sampled QA pairs. As shown in Tab. 13, 91.6% of the verified samples are correct. The remaining errors mainly come from ambiguous references, occlusion or distance ambiguity, sparse annotations, and incorrect metric or relation answers. We keep the final data format identical to indoor instruction tuning, enabling mixed SFT training with VSI-590K.

B More Training and Implementation Details

B.1 Backbone and Adaptation

We build ConsiSpace on **Qwen3-VL-8B-Instruct** and apply LoRA to the language-model linear layers. The visual encoder, aligner, and geometry encoder are frozen throughout both SFT and UC-SSRL. Unless otherwise stated, all training uses bfloat16 precision with DeepSpeed ZeRO-3. This design keeps training lightweight and makes UC-SSRL a post-SFT refinement stage rather than full-model retraining. LoRA uses rank 16 and scaling factor $\alpha=32$. The trainable-parameter ratio is small because only LoRA parameters are updated.

B.2 SFT Settings

We instruction-tune the model on a mixture of indoor VSI-590K and outdoor nuScenes-10K using next-token prediction. The default configuration is:

- **Optimization:** AdamW optimizer, learning rate 1×10^{-4} , per-device batch size 1, gradient accumulation 2, and 200 training steps.
- **LoRA:** rank 16 and scaling factor $\alpha=32$.
- **Sequence length:** maximum sequence length 5,120 with right truncation.
- **Efficiency settings:** FlashAttention, padding-free batching, gradient checkpointing, and multi-process data loading.

B.3 GCM Configuration

We enable the Geometry-Consistent Memory during both SFT and UC-SSRL. The memory uses $M=64$ memory tokens with dimension $d_{\text{mem}}=512$ and $n_{\text{chunks}}=4$ chunks. Memory tokens are injected every 4 transformer layers starting from

layer 8. Retrieval uses fused hierarchical selection with top- K chunk retrieval and top- K frame retrieval, both set to $K=8$. Each chunk contains 8 value tokens, and each selected frame contributes 4 value tokens. We use FAISS acceleration and cache retrieval keys when possible to reduce repeated retrieval overhead.

B.4 UC-SSRL Settings

UC-SSRL is initialized from the SFT checkpoint and updates only LoRA parameters. We keep the same batch size, step count, and optimizer setting as SFT for controlled comparison. The overall UC-SSRL weight is set to 1×10^{-2} . The two-view sampler constructs paired contexts by perturbing temporal windows and retrieval contexts while keeping the underlying scene and question unchanged. This setting optimizes cross-view consistency without requiring additional human annotations.

B.5 Inference Protocol

We follow the official evaluation protocol of each benchmark. For VSI-Bench and OSI-Bench, we use their official MCA and NA evaluation metrics. For MMSI-Video-Bench, we report both *Sufficient-Coverage* and *Uniform-50* settings. Unless otherwise stated, all methods use the same decoding parameters and frame-budget settings for fair comparison. Efficiency and scalability experiments are measured on the same A100 80GB GPU setting.

C Visualization

Figs. 8, 9, and 10 provide additional qualitative examples on VSI-Bench, MMSI-Video-Bench, and OSI-Bench. The examples cover metric estimation, relative-direction reasoning, object counting, navigation, and motion-related questions. Compared with the baseline, ConsiSpace produces answers closer to the ground truth across different spatial reasoning settings, supporting the effectiveness of geometry-consistent evidence organization.

D Limitation

While ConsiSpace improves efficiency and robustness through geometric consistency, several limitations remain:

- **Geometry estimation errors.** ConsiSpace uses predicted pose and depth cues for memory control. Errors under motion blur, reflective surfaces, or textureless regions may affect writing, fusion, and retrieval quality.
- **Threshold-based memory control.** Although our sensitivity analysis shows stable performance across a range of thresholds, adaptive gating and fusion may further improve robustness across diverse scenes.

- **Long-horizon scaling.** Extremely long videos may still require stronger memory compression or learned eviction strategies to maintain bounded compute under strict budgets.

Future work may improve geometry estimation, develop adaptive memory control, and extend consistency learning to broader video spatial reasoning settings.

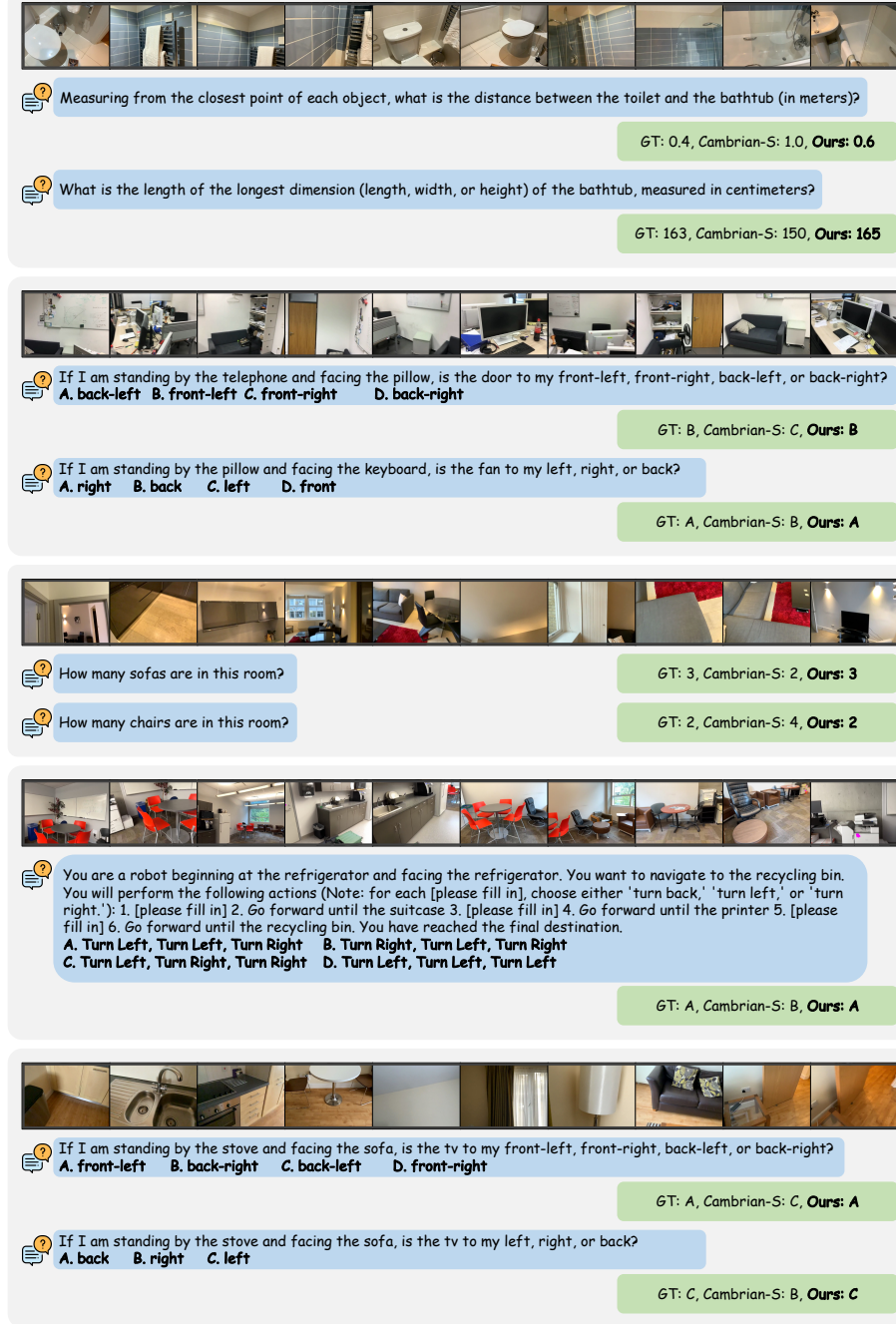
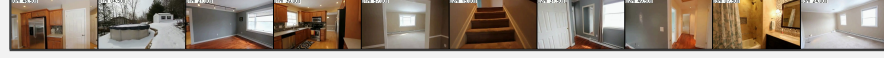


Fig. 8: Qualitative results on VSI-Bench.



- 1 This video is non-sequentially filmed in a building, and it is known that there are only two toilets and one staircase inside this building. Currently, I am facing a large gray refrigerator. I want to go to the toilet and need to plan a route to reach the toilet as quickly as possible. What are the correct steps to take next?
- A. Turn right, go straight into the new room, then turn right and walk out of the room, open the door opposite the staircase and go in.
 - B. Turn right, go straight into the new room, then turn right and walk out of the room to reach the staircase and go upstairs.
 - C. Turn left, go straight into the new room, then turn right and walk out of the room to reach the staircase and go upstairs.
 - D. Turn left, go straight into the new room, then turn right and walk out of the room, open the door on the left side of the staircase and go in.

GT: C, Cambrian-S: B, Ours: C



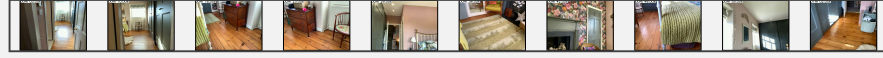
- 2 For the black car shown in this image, to leave this square parking area, how many accessible routes are there in total?
- A. 3
 - B. 4
 - C. 5
 - D. 2

GT: B, Cambrian-S: A, Ours: B



- 2 This is a video of a car driving. Assuming the distance between the same ends of two solid white lines on the road in the video is 3 meters, how many meters did the car travel in this video segment?
- A. 21m
 - B. 18m
 - C. 19.5m
 - D. 18.5m

GT: C, Cambrian-S: A, Ours: C



- 1 This is a video of indoor scene roaming. Define the front of the chair as the side a person faces when sitting on it in normal use. Take the front of the first chair seen when entering the door as the 0-degree forward direction. Where is the floor mirror located in the scene?
- A. Clockwise 90 degrees
 - B. Counterclockwise 90 degrees
 - C. Clockwise 45 degrees
 - D. Counterclockwise 45 degrees

GT: C, Cambrian-S: B, Ours: C



- 2 Assuming you are currently facing the green chopping board, what will you see if you turn clockwise in the order of first seeing this type of object?
- A. Door, chair, table, refrigerator
 - B. Chopping board, refrigerator, red sofa, chair
 - C. Chopping board, refrigerator, chair, door
 - D. Refrigerator, kettle, plate, door

GT: A, Cambrian-S: B, Ours: A

- 2 Observe this first-person cooking video. At 26.50 seconds, what will you see when you turn around?
- A. Dining room table
 - B. Chair in the dining room
 - C. Large wooden door
 - D. A door leading to another room

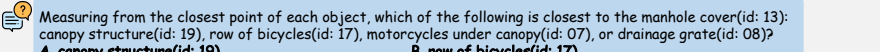
GT: D, Cambrian-S: C, Ours: D

Fig. 9: Qualitative results on MMSI-Video-Bench.



① Measuring from the closest point of each object, which of the following is closest to the canopy with blue roof(id: 20): row of bicycles under canopy(id: 10), canopy structure(id: 19), drainage grate(id: 08), or motorcycles(id: 21)?
A. row of bicycles under canopy(id: 10) **B. canopy structure(id: 19)**
C. drainage grate(id: 08) **D. motorcycles(id: 21)**

GT: B, Cambrian-S: C, **Ours: B**



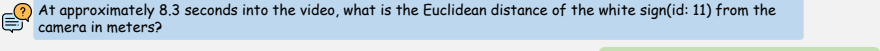
① Measuring from the closest point of each object, which of the following is closest to the manhole cover(id: 13): canopy structure(id: 19), row of bicycles(id: 17), motorcycles under canopy(id: 07), or drainage grate(id: 08)?
A. canopy structure(id: 19) **B. row of bicycles(id: 17)**
C. motorcycles under canopy(id: 07) **D. drainage grate(id: 08)**

GT: C, Cambrian-S: A, **Ours: C**



② What is the displacement distance of the person wearing gray shirt(id: 16) between 6.2s and 18.7s in meters?

GT: 2.68, Cambrian-S: 1.4, **Ours: 2.51**



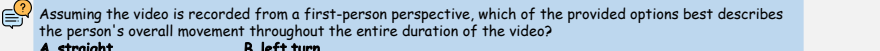
② At approximately 8.3 seconds into the video, what is the Euclidean distance of the white sign(id: 11) from the camera in meters?

GT: 4.25, Cambrian-S: 3.4, **Ours: 4.41**



② If I am standing by the stroller(id: 06) and facing the stroller(id: 07), is the Chanel logo(id: 11) to my front-left, front-right, back-left, or back-right?
A. front-left **B. front-right**
C. back-left **D. back-right**

GT: A, Cambrian-S: C, **Ours: A**



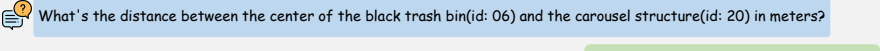
② Assuming the video is recorded from a first-person perspective, which of the provided options best describes the person's overall movement throughout the entire duration of the video?
A. straight **B. left turn**
C. right turn **D. U turn**

GT: B, Cambrian-S: A, **Ours: B**



② How long has the camera travelled throughout the entire duration of the video in meters?

GT: 3.9, Cambrian-S: 2.4, **Ours: 3.5**



② What's the distance between the center of the black trash bin(id: 06) and the carousel structure(id: 20) in meters?

GT: 11.10, Cambrian-S: 9.67, **Ours: 10.53**

Fig. 10: Qualitative results on OSI-Bench.