THOUGHT–STATE CO-EVOLUTION: CLOSED-LOOP COGNITIVE MAPPING FOR SPATIAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Scene-level spatial reasoning requires multimodal large language models to integrate partial video observations into task-sufficient spatial states. Existing cognitive-map pipelines support this integration but often lack reasoning-to-state feedback: externally supplied entity inventories leave relevance assessment outside the reasoning process, while fixed state estimates prevent evidence-driven revision when inference reveals uncertainty. To close this loop, we propose Reasoning-in-the-Loop Cognitive Mapping (RiCM), a framework in which reasoning directs state expansion and verification, and updated states guide subsequent inference. A dedicated spatial tool uses instance-consistent 3D anchors to ground entity association and evidence-gated local revision, decoupling geometric estimation from high-level reasoning. Final-answer rewards alone, however, do not explicitly supervise the sufficiency of intermediate states. We therefore introduce Progressive State Sufficiency Training (PSST), which promotes early and sustained state sufficiency through evaluation by an independent reader and unbiased estimation of the fraction of a fixed interaction horizon covered by sufficient states. Its reward combines joint correctness of the policy and the terminal-state reader with an interaction penalty, enabling the policy to learn when to expand, verify, and stop without ground-truth maps or oracle entity inventories. Experiments on ReVSI and MMSI-Video-Bench show consistent gains, supporting the value of bidirectional reasoning-state feedback for spatial reasoning.

1 INTRODUCTION

Spatial intelligence requires both local spatial perception and global spatial reasoning (Li et al., 2026e). Local spatial perception identifies visible objects and their directly observable spatial relations within an individual view (Yang et al., 2019; Chen et al., 2024). Global spatial reasoning, in contrast, integrates partial observations across time and viewpoints to infer scene-level spatial structure, including relations between objects that may never appear together in a single view (Ravi et al., 2025). This requires associating object instances across views, reconciling viewpoint-dependent observations, and composing spatial relations beyond those directly observed. As illustrated in Figure 1, modern multimodal large language models (MLLMs) can exhibit local spatial awareness yet struggle to form coherent global spatial representations (Yang et al., 2025a). We therefore focus on improving the global spatial reasoning capabilities of MLLMs.

A key difficulty underlying this gap is that accurate local observations do not automatically constitute a coherent spatial state (Wu et al., 2025; Gwak et al., 2026). Video presents evidence indexed by time and viewpoint, whereas global inference depends on persistent entity identities and spatial relations expressed in a common reference frame (Gu et al., 2024). Even when individual observations are correctly interpreted, errors in cross-view entity association, reference-frame transformation, or relational composition can still lead to incorrect global conclusions (Yang et al., 2026). The challenge is therefore to consolidate fragmented evidence into a task-sufficient state that captures the entities and relations needed to answer the question. Cognitive maps offer a way to make this integration explicit by organizing entity identities and spatial relations into a persistent representation, providing a structured basis for reasoning across observations (Wang et al., 2025a).

To provide a task-sufficient basis for reasoning, cognitive maps must in turn be constructed and maintained in response to the information requirements of inference (Bajcsy, 1988). Under partial

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.

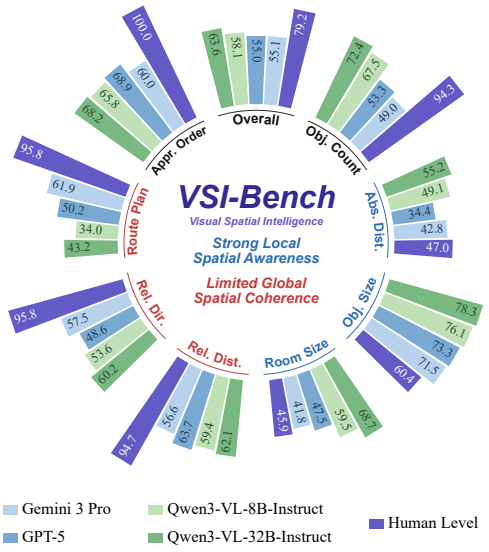


Figure 1: VSI-Bench performance highlights a gap between local spatial awareness and global spatial reasoning: MLLMs exhibit strong local perception but lag behind humans on tasks requiring global spatial understanding.

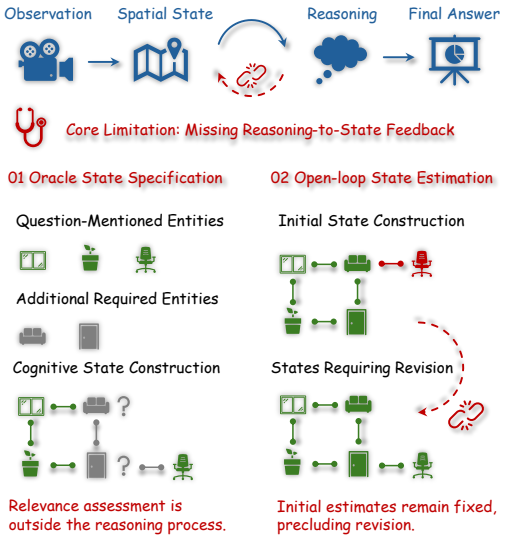


Figure 2: Missing reasoning-to-state feedback limits its cognitive mapping in two ways: oracle state specification leaves relevance assessment outside reasoning, while open-loop state estimation prevents revision of uncertain estimates during inference.

observability, these requirements may become apparent as reasoning establishes the spatial dependencies underlying an answer. Representative cognitive-map methods, however, construct a spatial representation before using it for downstream answer reasoning (Huang et al., 2025; Gao et al., 2026), without an explicit feedback mechanism through which emerging reasoning requirements can trigger targeted evidence retrieval and state revision. As illustrated in Figure 2, this absence of reasoning-to-state feedback manifests in two limitations:

- **Oracle state specification.** Supplied object inventories prescribe which entities to represent, leaving relevance assessment outside reasoning. Such assessment may require identifying entities absent from the question but necessary for cross-view association or instance disambiguation.
- **Open-loop state estimation.** Initial state estimates remain fixed even when reasoning exposes uncertainty through relational composition. Without feedback to trigger targeted evidence retrieval and state revision, errors in answer-critical variables can propagate into the final answer.

To address these limitations, we propose Reasoning-in-the-Loop Cognitive Mapping (RiCM), a closed-loop framework that integrates state specification and revision into spatial reasoning. Through *reasoning-to-state feedback*, the policy specifies task-relevant entities and relations as spatial dependencies emerge and requests targeted evidence to expand the map or verify answer-critical estimates. To support these requests, *geometry-assisted map construction* uses a dedicated spatial tool to consolidate cross-view observations into instance-consistent 3D anchors, decoupling geometric estimation from high-level reasoning. Within a shared reference frame, these anchors ground entity association and support evidence checks that guide local fusion, correction, or downweighting of estimates. The updated map and unresolved variables then guide subsequent inference and evidence requests, closing a loop in which map content adapts to task demands and estimates remain revisable.

Training this closed-loop policy requires learning to construct task-sufficient states, yet final-answer correctness alone does not establish whether the state itself supports answering. We therefore introduce Progressive State Sufficiency Training (PSST), which combines independent state evaluation with trajectory rewards for achieving sufficiency early and sustaining it as reasoning proceeds. A frozen reader assesses sufficiency by answering from the question and constructed state alone, providing a training signal without ground-truth maps or oracle entity inventories. Applying this test at a uniformly sampled checkpoint yields an unbiased estimate of the fraction of a fixed interaction horizon covered by sufficient states. Positive reward requires both the policy and the reader using

the terminal state to answer correctly, while an interaction penalty discourages unnecessary state updates. Together, these signals guide the policy in learning when to expand the state, verify uncertain estimates, and stop gathering evidence within a limited interaction budget.

2 Motivation: Cognitive Mapping Without Reasoning Feedback

Strong local spatial perception does not ensure reliable global spatial reasoning (Figure 1), which requires integrating fragmented observations into a task-sufficient spatial state. Cognitive maps externalize entity identities and spatial relations, yet existing methods largely decouple state construction from the reasoning it supports. This unidirectional interface prevents reasoning from determining what the state should represent or revising estimates as uncertainty emerges during inference. Under partial observability, however, *state support* is task-dependent and *state estimates* remain provisional. As illustrated in Figure 2, the missing reasoning-to-state feedback leads to two distinct failures.

Feedback Failure 1: Oracle state specification. Existing cognitive-map methods often assume task-relevant entities and relations are known before map construction, treating state support as an external input. Task-specific object inventories thus provide oracle relevance judgments that the model would otherwise need to infer from the spatial dependencies underlying the answer. Such dependencies may involve entities absent from the question but necessary for cross-view association or instance disambiguation. Representing all detected entities avoids explicit relevance selection but increases computational cost and introduces irrelevant content. Consequently, task-conditioned selection of compact, sufficient state support remains outside the reasoning process.

To test whether supplied inventories encode task relevance beyond entity count and question mentions, we compare four map-construction inventories while holding the video, question, and model fixed: (1) the original inventory; (2) a size-matched random subset of detected entities; (3) question-mentioned entities; and (4) all detected entities. As shown in Figure 3, random and question-only inventories reduce accuracy by 7.4 and 4.5 points, respectively, relative to the original inventory. Using all detected entities also reduces accuracy by 6.0 points despite including 4.0× as many entities. These results show that supplied inventories encode task-relevant information beyond entity count or question mentions, motivating explicit task-conditioned selection of state support.

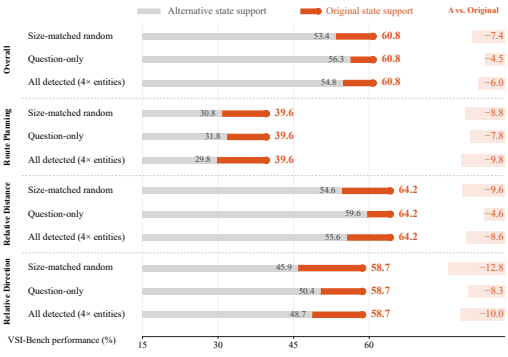


Figure 3: Supplied inventories outperform random, question-only, and exhaustive support, revealing dependence on externally supplied relevance beyond entity count and question mentions.

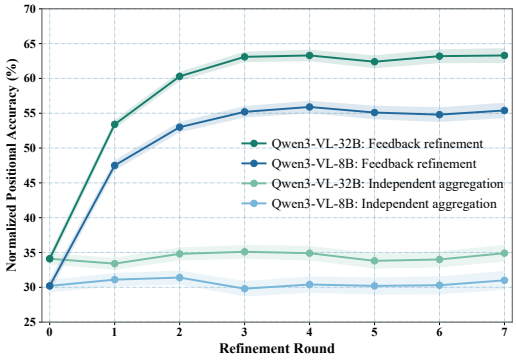


Figure 4: Reasoning-guided refinement outperforms independent aggregation under matched budgets, showing that reasoning feedback enables verification and correction of state estimates.

Feedback Failure 2: Open-loop state estimation. Existing cognitive-map methods treat map construction as the endpoint of state estimation, keeping the resulting map fixed throughout downstream reasoning. Partial and viewpoint-dependent observations, however, can yield ambiguous entity identities, incorrect cross-view correspondences, or inconsistent spatial relations. These deficiencies may become apparent when reasoning combines map entries to establish the spatial dependencies required by the answer. Resolving them may require localizing the uncertain state variables and retrieving observations that distinguish competing hypotheses about their values. An open-loop pipeline provides no mechanism for reasoning to initiate this verification or revise the corresponding

estimates. Consequently, the map remains unchanged even when reasoning exposes uncertainty in answer-critical variables, allowing initial estimation errors to propagate into the final answer.

We therefore test whether reasoning-guided verification improves state estimation beyond repeated independent map construction. Using videos with 3D scene annotations, we align predicted maps with ground-truth object locations and measure normalized positional accuracy across refinement rounds. Under matched inference budgets, we compare (1) *independent aggregation*, which combines one-pass maps constructed without feedback, and (2) *reasoning-guided refinement*, which uses reasoning-identified uncertainty to retrieve disambiguating observations and revise affected state variables. As shown in Figure 4, reasoning-guided refinement increases accuracy from 30.2 to 55.4, whereas independent aggregation yields no improvement. This contrast supports reasoning-to-state feedback for verifying and correcting the state estimates on which subsequent inference depends.

3 Closing the Loop Between Reasoning and Cognitive Mapping

The preceding analysis yields two requirements for reliable cognitive mapping: state construction must be conditioned on the information needs of reasoning, and uncertain state estimates must remain open to verification and revision. We instantiate these requirements with **Reasoning-in-the-Loop Cognitive Mapping** (RiCM), which brings cognitive-map construction inside the reasoning process. Given a video $V = \{I_t\}_{t=1}^{T}$ and a question q, a policy first diagnoses which spatial variables are missing or unreliable. A spatial tool then retrieves diagnostic evidence and uses instance-consistent 3D anchors to expand or revise the cognitive map. The updated state, together with its unresolved variables, is returned to the policy to trigger another evidence request or the final answer. As illustrated in Figure 5, RiCM makes state support task-conditioned and state revision uncertainty-driven.

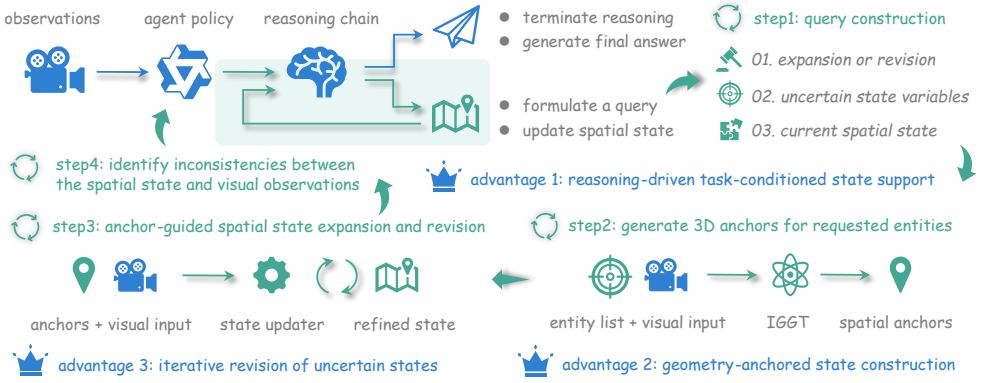


Figure 5: Overview of RiCM. Reasoning directs task-conditioned map expansion and revision, while 3D geometric anchors ground state construction and localized updates. Updated map and unresolved variables guide subsequent inference, closing the loop between reasoning and cognitive mapping.

3.1 Reasoning-Conditioned State Support

To replace oracle state specification, RiCM lets the policy request variables needed for the answer and expand the state as new dependencies emerge. At interaction step i, it maintains

$$\mathcal{B}_i \triangleq (\mathcal{S}_i, \mathcal{M}_i, \mathcal{U}_i), \qquad \mathcal{M}_i = (\mathcal{N}_i, \mathcal{E}_i). \tag{1}$$

Here, $\mathcal{S}_i$ specifies the entity and relation variables admitted to reasoning. The map $\mathcal{M}_i$ stores instance identities, 3D locations, confidence estimates, and evidence provenance in nodes $\mathcal{N}_i$, and spatial relations in edges $\mathcal{E}_i$. The unresolved set $\mathcal{U}_i$ tracks missing targets and conflicting relations that may affect the answer. This separation makes both missing state content and uncertain estimates explicit: observations support entries in $\mathcal{M}_i$, while unsettled hypotheses remain in $\mathcal{U}_i$ for verification.

The policy receives a sampled context $\widetilde{V} \subset V$, while the spatial tool retrieves additional views in response to the requests. Given q and the current belief, π_θ chooses whether to search or answer:

$$u_i \sim \pi_\theta(\cdot \mid q, \widetilde{V}, \mathcal{B}_i), \qquad u_i \in \{\textsc{Search}, \textsc{Answer}\}. \tag{2}$$

A search request $\rho_i = \langle \mathcal{T}_i, \mathcal{L}_i, m_i \rangle$ specifies entities to locate or re-examine $\mathcal{T}_i$, relations to resolve $\mathcal{L}_i$, and a mode $m_i \in \{\text{EXPAND}, \text{VERIFY}\}$. EXPAND requests an answer-critical entity or relation missing from the state; VERIFY targets a proposition in $\mathcal{U}_i$ and seeks evidence that distinguishes its competing hypotheses. Requests specify information needs, while retrieved evidence supplies candidate instances. The request can therefore change as the map exposes missing dependencies or conflicting evidence, making state specification part of the reasoning trajectory.

3.2 GEOMETRY-ANCHORED STATE CONSTRUCTION

To consolidate observations across viewpoints, the spatial tool $\mathcal{G}$ retrieves a compact view set $\mathcal{F}_i \subset V$ conditioned on ρ_i, prioritizing target coverage, co-visibility, and complementary viewing directions. Verification also retrieves supporting views of the relevant historical anchors to align new evidence with the existing state. A pretrained Instance-Grounded Geometry Transformer (IGGT) (Li et al., 2026a) jointly predicts dense depth, 3D point maps, and instance features from these views.

Cross-view predictions for instance k are aggregated into an anchor $\boldsymbol{\alpha}_{i,k} \triangleq (\mathbf{h}_{i,k}, \boldsymbol{\mu}_{i,k}, \boldsymbol{\Sigma}_{i,k}, \mathcal{V}_{i,k})$: an association feature $\mathbf{h}_{i,k}$, 3D point mean $\boldsymbol{\mu}_{i,k} \in \mathbb{R}^3$ and covariance $\boldsymbol{\Sigma}_{i,k} \in \mathbb{R}^{3\times3}$, and supporting-view indices $\mathcal{V}_{i,k}$. Anchors are registered to a persistent bird's-eye-view frame using camera estimates on overlapping views and shared instances. This shared frame lets verification compare new geometric evidence with existing locations and relations across interaction steps. Each anchor thus ties an instance identity to its geometry and source observations for subsequent verification.

From $\mathcal{A}_i = \{\boldsymbol{\alpha}_{i,k}\}_k$, the map generator selects candidates that locate requested entities, establish requested relations, or resolve ambiguities in $\mathcal{U}_i$. Repeated observations are associated with existing nodes, and relations are derived in the common 3D frame. Auxiliary entities are selected when needed for cross-view association or instance disambiguation, even if unmentioned in q. This lets state support extend beyond question mentions while avoiding exhaustive inclusion of detections: geometry supplies the candidates, and reasoning demand determines which enter the evidence check.

3.3 EVIDENCE-GATED LOCAL REVISION

Because partial observations can leave identities or relations unresolved, RiCM checks candidate updates against retrieved evidence and returns unresolved variables to the policy. Let $\widehat{\mathbf{p}}_{i,k}$ be the instance location implied by the candidate map after semantic grounding. With $\boldsymbol{\delta}_{i,k} = \widehat{\mathbf{p}}_{i,k} - \boldsymbol{\mu}_{i,k}$, agreement with the anchor is measured by the regularized Mahalanobis distance

$$d^2_{i,k} = \boldsymbol{\delta}^\top_{i,k}(\boldsymbol{\Sigma}_{i,k} + \varepsilon \mathbf{I}_3)^{-1}\boldsymbol{\delta}_{i,k}, \tag{3}$$

where $\varepsilon > 0$ ensures invertibility. Together with association confidence and relational consistency, $d^2_{i,k}$ forms an evidence statistic $\boldsymbol{\omega}_i$. Consistent estimates are fused into the map; conflicting nodes or relations are downweighted; unresolved hypotheses enter $\mathcal{U}_{i+1}$ with their conflict type and suggested retrieval target. This feedback identifies both what remains uncertain and where to seek disambiguating evidence, allowing the policy to direct the next request to specific state variables.

Both request modes use the same update operator:

$$\mathcal{B}_{i+1} = \text{Update}(\mathcal{B}_i, \mathcal{A}_i, \rho_i; \boldsymbol{\omega}_i). \tag{4}$$

EXPAND adds evidence-supported variables to $\mathcal{S}_i$ and $\mathcal{M}_i$; VERIFY re-associates, corrects, or downweights the requested nodes and relations while preserving unrelated reliable structure. The policy answers when all answer-critical variables are represented and remaining uncertainty would not change the answer; otherwise, it continues targeted retrieval up to a maximum horizon $I_{\max}$.

4 PROGRESSIVE STATE SUFFICIENCY TRAINING

The RiCM policy must learn to expand the state, verify uncertain estimates, and stop gathering evidence. GRPO (Shao et al., 2024) with rewards based solely on final-answer correctness does not require state construction itself to succeed: answering from a sufficient map and compensating for an inadequate map through video context receive the same correctness reward. *Answer correctness does not establish state sufficiency.* We introduce **Progressive State Sufficiency Training** (PSST), which augments trajectory rewards to supervise whether a state that supports correct answering is formed and how early it emerges. Figure 6 summarizes the state evaluation and reward construction in PSST.

4.1 Testing Sufficiency through State Handoff

PSST tests whether the required information has been consolidated into an explicit state. We pass the question q and state $\mathcal{B}_i$ to a frozen reader f_{read} with no access to the trajectory, and define

$$v_i = c(f_{\text{read}}(q, \mathcal{B}_i), y^*) \in \{0, 1\}, \tag{5}$$

where y^* is the ground-truth answer and c is the task-specific correctness criterion. The reader sees only q and the structured state maintained by the tool, including unresolved variables and conflicts; video, interaction history, and the policy's answer are excluded. This operational test measures task sufficiency without ground-truth maps or externally specified entity inventories.

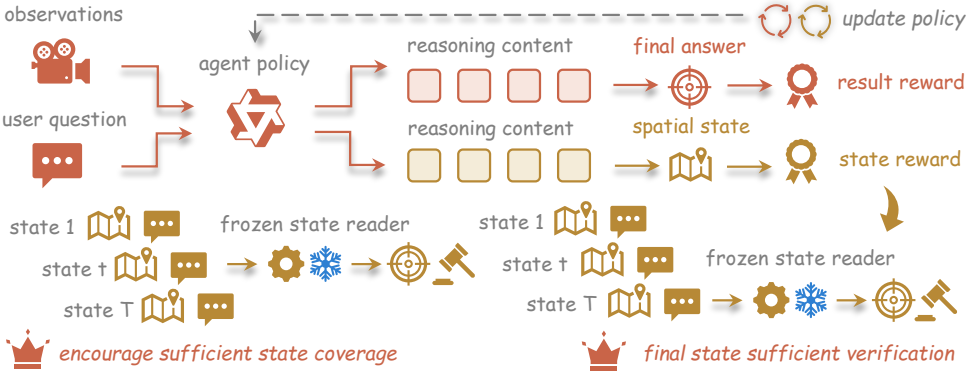


Figure 6: Overview of PSST. A frozen reader evaluates state sufficiency from the question and structured state. Terminal verification measures final sufficiency, while a sampled checkpoint estimates sufficiency coverage over the interaction horizon. Reward combines joint correctness, a success-gated sufficiency bonus, and an interaction penalty, encouraging early and sustained state sufficiency.

4.2 Sampling Sufficiency across the Interaction Horizon

A terminal check cannot distinguish early from late state construction, while evaluating only the first successful state overlooks subsequent degradation. We therefore optimize the *fraction of a fixed interaction horizon covered by sufficient states*. For a trajectory τ ending after $K \le I_{\max}$ updates, we extend its terminal state over the remaining horizon for evaluation, defining $\bar{v}_j = v_{\min(j,K)}$. After rollout, we sample one checkpoint, analogous to reasoning evaluation under sampled budgets:

$$J \sim \text{Unif}\{1, \dots, I_{\max}\}, \qquad \mathbb{E}_J[\bar{v}_J \mid \tau] = \frac{1}{I_{\max}} \sum_{j=1}^{I_{\max}} \bar{v}_j. \tag{6}$$

One check thus provides an unbiased estimate of this fraction (Qi et al., 2025). Among trajectories ending in sufficient states, later attainment or longer subsequent loss of sufficiency lowers the score, all else equal. Terminal extension prevents repetition of sufficient states from inflating their fraction in these trajectories. Each rollout requires at most two additional reader evaluations: the terminal state and one sampled state. We reuse the terminal result when $J \ge K$.

4.3 Optimizing Closed-Loop Decisions with a Unified Reward

Let $\hat{y}_\tau$ be the policy's final answer and $s_\tau = c(\hat{y}_\tau, y^*) v_K$ indicate joint success. We define

$$R(\tau, J) = s_\tau\big(1 + \lambda \bar{v}_J\big) - \beta \frac{K}{I_{\max}}, \qquad \lambda > 0, \quad 0 < \beta < 1. \tag{7}$$

Positive utility requires both the policy and the terminal-state reader to answer correctly; transient sufficiency cannot compensate for an inadequate final state. Since $\beta < 1$, jointly successful trajectories have strictly positive raw rewards, and all others have nonpositive raw rewards. The weights λ and β balance the additional reward for early, sustained sufficiency against interaction cost. Starting from a policy with basic tool-use competence, we freeze the spatial tool and reader and apply GRPO to trajectory groups sampled for the same input, optimizing only policy-generated tokens and masking tool outputs. Expand, Verify, and stopping thus share one criterion: construct a state that supports correct answering within a limited interaction budget.

5 EXPERIMENTS

In this section, we evaluate RiCM's overall performance and investigate the contributions of reasoning-to-state feedback, geometry-assisted map construction, and Progressive State Sufficiency Training (PSST). We structure our experiments around the following four research questions:

- **RQ1:** How does RiCM compare with existing methods on video spatial reasoning tasks, particularly those requiring cross-view integration and relational composition?
- **RQ2:** How does reasoning-to-state feedback affect the construction and revision of task-relevant states? What are the respective contributions of dynamic state expansion and targeted verification?
- **RQ3:** Can geometric anchors and evidence gating improve the reliability of state construction and revision while preserving previously correct estimates?
- **RQ4:** Can PSST help the policy achieve state sufficiency earlier, maintain it throughout reasoning, and learn when to stop gathering evidence under a limited interaction budget?

Table 1: Spatial reasoning performance of models and agent-based methods on ReVSI. Red and blue indicate the best result in each column among proprietary and non-proprietary methods, respectively.

Methods	Obj. Cnt	A. Dist.	Obj. Size	R. Size	Rel.Dist.	Rel.Dir.	R. Plan	Overall
Proprietary Models (API)								
GPT-5.2	56.2	41.5	73.9	63.0	48.4	34.9	38.2	50.9
Gemini 3 Flash	65.7	53.1	77.6	52.8	64.6	47.9	41.8	57.6
Gemini 3 Pro	60.1	54.7	79.3	51.9	68.1	56.0	56.4	60.9
Open-source Models								
Qwen3-VL-8B-Instruct	40.4	52.3	69.0	45.1	57.1	39.5	40.5	49.1
Qwen3-VL-32B-Instruct	46.9	65.0	70.4	55.8	53.8	34.0	47.3	53.3
InternVL3.5-8B	43.3	54.6	64.2	47.6	45.0	36.3	44.4	47.9
InternVL3.5-38B	43.8	60.6	70.2	58.4	57.4	45.9	42.7	54.1
Spatial-centric Models								
Cambrian-S-7B	48.4	60.5	65.5	46.7	37.1	48.5	37.0	49.1
VLM-3R-7B	41.6	61.6	64.8	52.5	46.5	49.5	34.1	50.1
GeoThinker-8B	35.1	56.9	63.5	37.9	50.7	45.3	40.7	47.1
SpaceMind-8B	38.3	54.8	68.4	56.8	51.9	47.1	40.2	51.1
Spatial Reasoning Agents								
Think3D-4B	25.1	35.0	38.2	32.6	19.4	29.4	23.7	29.1
OmniAgent-7B	21.7	44.6	38.8	33.5	28.6	32.4	34.0	33.4
Skill-3D-8B	24.1	47.8	41.6	35.3	23.9	30.0	34.9	33.9
RiCM-8B (Ours)	55.4	49.3	64.1	54.6	61.9	73.5	64.2	60.2

5.1 EXPERIMENTAL SETUP

Module Implementation. We use separate copies of Qwen3-VL-8B-Instruct (Bai et al., 2025a) for the policy and cognitive-map generator, and construct 3D anchors from pretrained IGGT (Li et al., 2026a) predictions. During training, a text-only Qwen3-8B (Qwen Team, 2025) reader answers using only the question and serialized state; its answer correctness provides the sufficiency signal. We optimize only the policy, keeping the spatial tool and reader frozen.

Training Settings and Data. We train the policy with PSST, using AdamW at a learning rate of 10^{-6}. Each batch contains 32 questions, with 16 rollouts per question and at most $I_{\max} = 7$ state updates per rollout. The sufficiency and interaction-cost weights are set to $\lambda = 1.0$ and $\beta = 0.2$, respectively. Training videos are drawn from S3DIS (Armeni et al., 2016), ScanNet (Dai et al., 2017), ScanNet++ V2 (Yeshwanth et al., 2023), ARKitScenes (Baruch et al., 2021), and Aria Digital Twin (Pan et al., 2023), with question–answer annotations from VSI-590K (Yang et al., 2025b). Rejection sampling with Qwen3-VL-8B-Instruct yields 80K training examples.

Evaluation Benchmarks. We evaluate RiCM on ReVSI and MMSI-Video-Bench using their official metrics. For the initial visual context, we uniformly sample 64 frames per ReVSI video and allocate a total budget of 50 frames across all clips in each MMSI-Video-Bench question. On both benchmarks,

the policy can then request additional views through the spatial tool to resolve missing or uncertain spatial information.

Table 2: Quantitative comparison of spatial reasoning performance on MMSI-Video-Bench. Red and blue indicate the best result in each column among proprietary and open-source methods, respectively.

Methods	Attribute	Inst.-Inst.	Inst.-Scen.	Scen.-Scen.	Planning	Prediction	Overall
Proprietary Models (API)							
GPT-5	44.6	38.2	44.2	39.1	34.7	28.1	36.8
Gemini 3 Pro	44.6	39.5	44.2	33.3	38.7	35.4	38.0
Doubao-Seed-1.6-vision	43.1	26.7	16.7	38.6	37.3	25.9	34.9
Spatial-centric Models							
Cambrian-S-7B	34.8	38.5	38.3	20.2	36.9	28.6	32.5
VLM-3R-7B	44.1	33.4	36.3	24.1	28.0	25.7	34.3
GeoThinker-8B	37.7	40.2	30.5	22.8	30.3	31.1	31.0
SpaceMind-8B	33.4	20.1	38.9	30.8	28.5	31.3	34.2
Spatial Reasoning Agents							
Think3D-4B	31.9	35.7	35.2	18.3	33.1	25.4	30.6
OmniAgent-7B	26.4	28.1	25.3	25.8	30.3	39.1	29.2
Skill-3D-8B	19.2	36.1	29.6	28.2	25.5	25.4	28.3
RiCM-8B (Ours)	44.9	37.1	41.8	35.6	40.8	39.6	38.5

5.2 RQ1: Performance on Spatial Reasoning Tasks

RiCM achieves overall scores of 60.2 on ReVSI and 38.5 on MMSI-Video-Bench (Tables 1 and 2), outperforming its Qwen3-VL-8B-Instruct backbone by 11.1 and 10.9 points, respectively. It also surpasses the strongest listed open-source baseline on each benchmark, exceeding InternVL3.5-38B by 6.1 points on ReVSI and VLM-3R-7B by 4.2 points on MMSI-Video-Bench. On ReVSI, RiCM leads all compared methods in relative direction (73.5) and route planning (64.2), with gains of 34.0 and 23.7 percentage points over the backbone. On MMSI-Bench, it scores 40.8 in planning and 39.6 in prediction, improving on the backbone by 15.8 and 10.3 percentage points, respectively.Together, these results demonstrate RiCM's effectiveness across spatial reasoning tasks. The gains in relative direction and route planning, which require cross-view integration and relational composition, suggest progress toward addressing the global spatial reasoning bottleneck motivating this work.

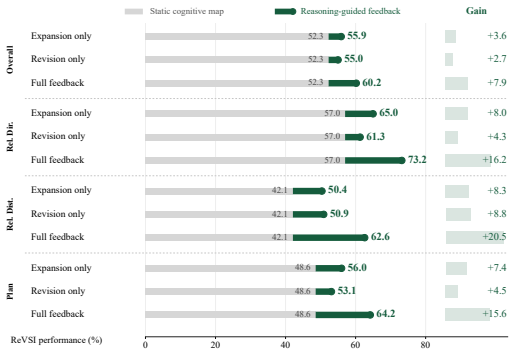


Figure 7: Ablation of reasoning-to-state feedback on ReVSI. State expansion and targeted revision provide complementary gains, with their combination achieving the strongest overall performance.

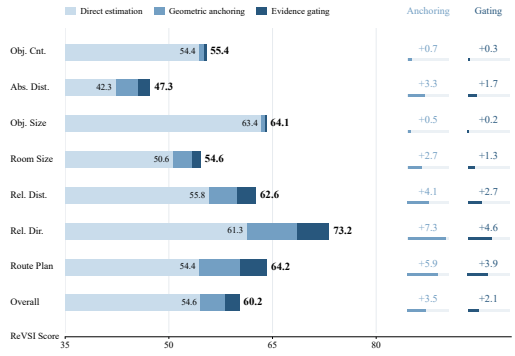


Figure 8: Ablation of geometric anchoring and evidence gating on ReVSI. Geometric anchoring provides the primary gain, with evidence gating offering further improvement.

5.3 RQ2: Effect of Reasoning-to-State Feedback

We evaluate reasoning-to-state feedback through a 2×2 ablation on ReVSI: static map, expansion only, revision only, and full feedback. All variants use the same initial map, constructed from

the question and video context without an external entity inventory, and share frozen spatial tools, evidence access, and interaction limits. Expansion adds entities or relations while preserving existing estimates; revision updates estimates within a fixed variable set. Final answers are scored using the official ReVSI metric. Figure 7 shows that expansion and revision alone improve the score from 52.3 to 55.9 and 55.0, respectively. Full feedback reaches 60.2, outperforming the two variants by 4.3 and 5.2 points. Freezing either state support or state estimates reduces performance, demonstrating the complementary value of reasoning-guided expansion and revision.

5.4 RQ3: Effect of Geometry-Assisted Mapping

To assess geometric anchoring and evidence gating, we compare RiCM with two ablated variants on ReVSI. The first removes both components and estimates entity locations and spatial relations directly from retrieved images. The second retains IGGT-derived anchors but disables the evidence checks used to accept or downweight candidate updates. All variants use the same policy and map-generator weights, initial video input, evidence access permissions, and interaction limit. Each constructs its own initial map, with subsequent revisions restricted to the requested state variables. Figure 8 reports final-answer scores under the official ReVSI metric: restoring geometric anchoring raises the score from 54.6 to 58.1, and restoring evidence gating raises it further to 60.2. The additional gain from gating supports validating candidate updates against the retrieved geometric evidence.

5.5 RQ4: Effect of PSST on State Sufficiency and Stopping Behavior

We compare four reward configurations: answer correctness only, terminal sufficiency supervision, PSST without the interaction penalty, and full PSST. All variants are trained on the same 10K examples with identical initialization and optimization settings, allowing up to seven state updates per rollout. Each policy is evaluated on ReVSI without retraining under inference budgets of one to seven state updates. The policy may answer early but must answer at the budget limit.

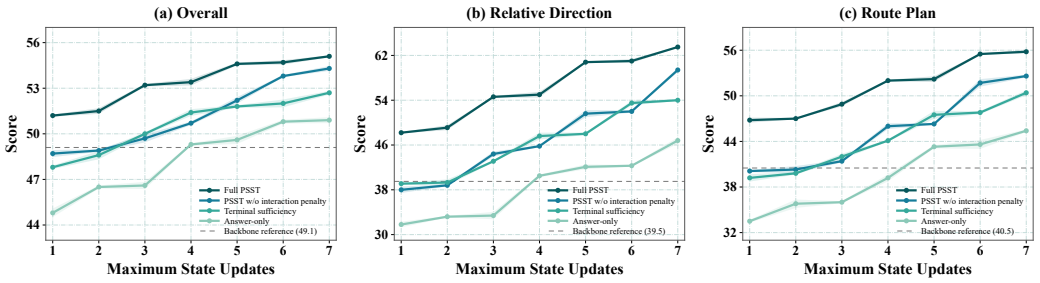


Figure 9: PSST ablations on ReVSI across interaction budgets. Progressive sufficiency supervision and interaction regularization provide complementary gains across budgets.

Figure 9 shows that full PSST consistently outperforms all ablations across interaction budgets. With a single update, it achieves an overall score of 51.2, surpassing answer-only supervision even with seven updates (50.9). Removing either the progressive reward or interaction penalty degrades performance, particularly on relative direction and route planning, while both variants remain stronger than answer-only supervision. Full PSST maintains its advantage at the maximum budget, demonstrating the complementary roles of progressive sufficiency supervision and interaction regularization in achieving strong performance under limited interaction while preserving final task performance.

6 Conclusion

We presented Reasoning-in-the-Loop Cognitive Mapping (RiCM), a framework that couples spatial reasoning with task-conditioned map expansion and targeted verification. Instance-consistent 3D anchors ground map construction and local revision, allowing spatial states to evolve alongside reasoning as evidence accumulates. To train the RiCM policy, we introduced Progressive State Sufficiency Training (PSST), which promotes early and sustained state sufficiency while regularizing interaction cost, without requiring ground-truth maps or oracle entity inventories. Together, RiCM and PSST provide a concrete approach to global spatial reasoning in which the model learns not only to reason from a spatial state, but also to construct, verify, and refine the state its reasoning requires.

AI USE STATEMENT

We used generative AI tools to assist with code implementation, identify relevant literature, and improve the language and readability of the manuscript. The authors reviewed and verified all AI-assisted code for correctness and consistency with the described methods. We checked AI-suggested references against the original publications and assessed their relevance to this work. We also reviewed all AI-assisted text for factual accuracy, consistency with the reported methods and results, and preservation of the intended scientific meaning. We take full responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work focuses on improving video-based spatial reasoning in multimodal large language models. Our experiments use existing research datasets and pretrained models, with data sources and evaluation protocols described in Section 5.1. We acknowledge that our approach may inherit biases from the underlying models and data. We encourage responsible use of the proposed method, including respect for data privacy and the usage terms of the source datasets and models, as well as careful evaluation before deployment in real-world applications.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide our training and evaluation code, together with instructions for reproducing the experiments, in the supplementary material. The model configurations, data sources, principal training hyperparameters, and evaluation protocols are reported in Section 5.1. The proposed framework and training objective are described in Sections 3 and 4, respectively.

REFERENCES

Iro Armeni, Ozan Sener, Amir R. Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3D semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1534–1543, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Armeni_3D_Semantic_Parsing_CVPR_2016_paper.html.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b. URL https://arxiv.org/abs/2502.13923.

Ruzena Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):996–1005, 1988. doi: 10.1109/5.5968. URL https://doi.org/10.1109/5.5968.

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitScenes - a diverse real-world dataset for 3D indoor scene understanding using mobile RGB-D data. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL https://openreview.net/forum?id=tjZjv_qh_CE.

Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. SpatialBot: Precise spatial understanding with vision language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9490–9498, 2025. URL https://arxiv.org/abs/2406.13642.

Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural SLAM. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HklXn1BKDH.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14455–14465, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Chen_SpatialVLM_Endowing_Vision-Language_Models_with_Spatial_Reasoning_Capabilities_CVPR_2024_paper.html.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. SpatialRGPT: Grounded spatial reasoning in vision-language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f38cb4cf9a5eaa92b3cfa481832719c6-Abstract-Conference.html.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5828–5839, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Dai_ScanNet_Richly-Annotated_3D_CVPR_2017_paper.html.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Peihao Wang, Huaizhi Qu, Shijie Zhou, Dilin Wang, Zhicheng Yan, Hongyu Xu, Justin Theiss, Tianlong Chen, Jiachen Li, Zhengzhong Tu, Zhangyang Wang, and Rakesh Ranjan. 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. URL https://openaccess.thecvf.com/content/CVPR2026/html/Fan_VLM-3R_Vision-Language_Models_Augmented_with_Instruction-Aligned_3D_Reconstruction_CVPR_2026_paper.html.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision (ECCV)*, 2024. URL https://zeyofu.github.io/blink/.

Xiangjun Gao, Zhensong Zhang, Dave Zhenyu Chen, Songcen Xu, Long Quan, Eduardo Pérez-Pellitero, and Youngkyoon Jang. Map2Thought: Explicit 3D spatial reasoning via metric cognitive maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 7154–7164, 2026. URL https://openaccess.thecvf.com/content/CVPR2026F/html/Gao_Map2Thought_Explicit_3D_Spatial_Reasoning_via_Metric_Cognitive_Maps_CVPRF_2026_paper.html.

Google DeepMind. Gemini 2.5 Pro model card. Technical report, Google DeepMind, June 2025a. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Pro-Model-Card.pdf.

Google DeepMind. Gemini 3 Flash model card. Technical report, Google DeepMind, December 2025b. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf.

Google DeepMind. Gemini 3 Pro model card. Technical report, Google DeepMind, May 2026. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf. Updated May 2026; model released November 2025.

Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. ConceptGraphs: Open-vocabulary 3D scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5021–5028, 2024. URL https://arxiv.org/abs/2309.16650.

Saurabh Gupta, James Davidson, Sergey Levine, Rahul Sukthankar, and Jitendra Malik. Cognitive mapping and planning for visual navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2616–2625, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Gupta_Cognitive_Mapping_and_CVPR_2017_paper.html.

Chanyoung Gwak, Yoonwoo Jeong, Byungwoo Jeon, Hyunseok Lee, Jinwoo Shin, and Minsu Cho. Cog3DMap: Multi-view vision-language reasoning with 3D cognitive maps. *arXiv preprint arXiv:2603.23023*, 2026. URL https://arxiv.org/abs/2603.23023.

Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, London, UK, 2023. URL https://arxiv.org/abs/2210.05714.

Yibin Huang, Wang Xu, Wanyue Zhang, Helu Zhi, Jingjing Huang, Yangbin Xu, Yangang Sun, Conghui Zhu, and Tiejun Zhao. Video2Layout: Recall and reconstruct metric-grounded cognitive map for spatial reasoning. *arXiv preprint arXiv:2511.16160*, 2025. URL https://arxiv.org/abs/2511.16160.

Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Ganesh Iyer, Soroush Saryazdi, Tao Chen, Alaa Maalouf, Shuang Li, Nikhil Varma Keetha, Ayush Tewari, Joshua Tenenbaum, Celso de Melo, Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. ConceptFusion: Open-set multimodal 3D mapping. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023. doi: 10.15607/RSS.2023.XIX.066. URL https://roboticsproceedings.org/rss19/p066.html.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=zKv8qULV6n.

Hao Li, Zhengyu Zou, Fangfu Liu, Xuanyang Zhang, Fangzhou Hong, Yukang Cao, Yushi Lan, Manyuan Zhang, Gang Yu, Dingwen Zhang, and Ziwei Liu. IGGT: Instance-grounded geometry transformer for semantic 3D reconstruction. In *International Conference on Learning Representations*, 2026a. URL https://arxiv.org/abs/2510.22706.

Haoyuan Li, Qihang Cao, Tao Tang, Kun Xiang, Zihan Guo, Jianhua Han, JiaWang Bian, Hang Xu, and Xiaodan Liang. Thinking with geometry: Active geometry integration for spatial reasoning. *arXiv preprint arXiv:2602.06037*, 2026b. doi: 10.48550/arXiv.2602.06037. URL https://arxiv.org/abs/2602.06037.

Haoyuan Li, Zhengdong Hu, Jun Wang, Hehe Fan, and Yi Yang. Skill-3D: Evolving scene-aware skills for agentic 3D spatial reasoning. *arXiv preprint arXiv:2606.07436*, 2026c. URL https://arxiv.org/abs/2606.07436.

Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. SpatialLadder: Progressive training for spatial reasoning in vision-language models. In *International Conference on Learning Representations*, 2026d. URL https://openreview.net/forum?id=KtrFXlvgrK.

Yifei Li, Pengyiang Liu, Yuhang Zang, Zhongyue Shi, Qi Fu, Hongye Hao, and Jiwen Lu. OVO-S-Bench: A hierarchical benchmark for streaming spatial intelligence in multimodal LLMs. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, 2026e. URL https://arxiv.org/abs/2606.03890.

Jingli Lin, Runsen Xu, Shaohao Zhu, Sihan Yang, Peizhou Cao, Yunlong Ran, Miao Hu, Chenming Zhu, Yiman Xie, Yilin Long, Wenbo Hu, Dahua Lin, Tai Wang, and Jiangmiao Pang. MMSI-Video-Bench: A holistic benchmark for video-based spatial intelligence. *arXiv preprint arXiv:2512.10863*, 2025. URL https://arxiv.org/abs/2512.10863.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan Yuille. 3DSRBench: A comprehensive 3D spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6924–6934, 2025a. URL https://openaccess.thecvf.com/content/ICCV2025/html/Ma_3DSRBench_A_Comprehensive_3D_Spatial_Reasoning_Benchmark_ICCV_2025_paper.html.

Wufei Ma, Luoxin Ye, Celso M de Melo, Alan Yuille, and Jieneng Chen. SpatialLLM: 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, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ma_SpatialLLM_A_Compound_3D-Informed_Design_towards_Spatially-Intelligent_Large_Multimodal_Models_CVPR_2025_paper.html.

OpenAI. Introducing GPT-5, 2025a. URL https://openai.com/index/introducing-gpt-5/.

OpenAI. Introducing GPT-5.2, 2025b. URL https://openai.com/index/introducing-gpt-5-2/.

Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. SpaceR: Reinforcing MLLMs in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025. doi: 10.48550/arXiv.2504.01805. URL https://arxiv.org/abs/2504.01805.

Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng (Carl) Ren. Aria Digital Twin: A new benchmark dataset for egocentric 3D machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20133–20143, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Pan_Aria_Digital_Twin_A_New_Benchmark_Dataset_for_Egocentric_3D_ICCV_2023_paper.html.

Emilio Parisotto and Ruslan Salakhutdinov. Neural map: Structured memory for deep reinforcement learning. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Bk9zbyZCZ.

Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. OpenScene: 3D scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–824, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Peng_OpenScene_3D_Scene_Understanding_With_Open_Vocabularies_CVPR_2023_paper.html.

Penghui Qi, Zichen Liu, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Optimizing anytime reasoning via budget relative policy optimization. *arXiv preprint arXiv:2505.13438*, 2025. doi: 10.48550/arXiv.2505.13438. URL https://arxiv.org/abs/2505.13438.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Sahithya Ravi, Gabriel Herbert Sarch, Vibhav Vineet, Andrew D. Wilson, and Balasaravanan Thoravi Kumaravel. Out of sight, not out of context? egocentric spatial reasoning in VLMs across disjoint frames. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.816. URL https://aclanthology.org/2025.emnlp-main.816/.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. doi: 10.48550/arXiv.2402.03300. URL https://arxiv.org/abs/2402.03300.

Volcengine. Doubao-Seed-1.6-vision. Model documentation, 2025. URL https://www.volcengine.com/docs/82379/1799865.

Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Jiajun Wu, Fei-Fei Li, and Manling Li. MindCube: Spatial mental modeling from limited views, 2025a. URL https://arxiv.org/abs/2506.21458.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b. URL https://arxiv.org/abs/2508.18265.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-MLLM: Boosting MLLM capabilities in visual-based spatial intelligence. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2505.23747.

Zhenghao Xing, Ruiyang Xu, Yuxuan Wang, Jinzheng He, Ziyang Ma, Qize Yang, Yunfei Chu, Jin Xu, Junyang Lin, Chi-Wing Fu, and Pheng-Ann Heng. Native active perception as reasoning for omni-modal understanding. *arXiv preprint arXiv:2606.19341*, 2026. URL https://arxiv.org/abs/2606.19341.

Kashu Yamazaki, Taisei Hanyu, Khoa Vo, Thang Pham, Minh Tran, Gianfranco Doretto, Anh Nguyen, and Ngan Le. Open-Fusion: Real-time open-vocabulary 3D mapping and queryable scene representation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9411–9417, 2024. URL https://arxiv.org/abs/2310.03923.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643, 2025a.

Kaiyu Yang, Olga Russakovsky, and Jia Deng. SpatialSense: An adversarially crowdsourced benchmark for spatial relation recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2051–2060, 2019. doi: 10.1109/ICCV.2019.00214. URL https://arxiv.org/abs/1908.02660.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Daohan Lu, Rob Fergus, Yann LeCun, Fei-Fei Li, and Saining Xie. Cambrian-S: Towards spatial supersensing in video. *arXiv preprint arXiv:2511.04670*, 2025b. URL https://arxiv.org/abs/2511.04670.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. MMSI-Bench: A benchmark for multi-image spatial intelligence. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=gHRoX4vXm3.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A high-fidelity dataset of 3D indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12–22, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Yeshwanth_ScanNet_A_High-Fidelity_Dataset_of_3D_Indoor_Scenes_ICCV_2023_paper.html.

Haoyu Zhang, Meng Liu, Zaijing Li, Haokun Wen, Weili Guan, Yaowei Wang, and Liqiang Nie. Spatial understanding from videos: Structured prompts meet simulation data. In *Advances in Neural Information Processing Systems*, volume 38, 2025a. doi: 10.52202/085713-3449. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/9541101fbc2f24bce5f2462b95db88c4-Abstract-Conference.html.

Yiming Zhang, Jiacheng Chen, Jiaqi Tan, Yongsen Mao, Wenhu Chen, and Angel X. Chang. ReVSI: Rebuilding visual spatial intelligence evaluation for accurate assessment of VLM 3D reasoning. *arXiv preprint arXiv:2604.24300*, 2026a. URL https://arxiv.org/abs/2604.24300.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025b. URL https://openreview.net/forum?id=EElFGvt39K.

Zaibin Zhang, Yuhan Wu, Lianjie Jia, Yifan Wang, Zhongbo Zhang, Yijiang Li, Binghao Ran, Fuxi Zhang, Zhuohan Sun, Yizhuang Peng, Zhenfei Yin, Lijun Wang, and Huchuan Lu. Think3D: Thinking with space for spatial reasoning. *arXiv preprint arXiv:2601.13029*, 2026b. URL https://arxiv.org/abs/2601.13029.

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, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhao_SpaceMind_Camera-Guided_Modality_Fusion_for_Spatial_Reasoning_in_Vision-Language_Models_CVPR_2026_paper.html.

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*, volume 38, 2025. URL https://arxiv.org/abs/2505.24625.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. LLaVA-3D: A simple yet effective pathway to empowering LMMs with 3D capabilities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4295–4305, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhu_LLaVA-3D_A_Simple_yet_Effective_Pathway_to_Empowering_LMMs_with_ICCV_2025_paper.html.

A Related Work

We review research on multimodal spatial reasoning and cognitive maps, focusing on how spatial evidence is acquired, represented, and used for inference.

A.1 Spatial Reasoning in Multimodal Models

Spatial reasoning in MLLMs has advanced through spatial supervision, geometric representations, and selective evidence acquisition. BLINK and 3DSRBench assess capabilities spanning depth, orientation, object relations, and multi-view perception (Fu et al., 2024; Ma et al., 2025a). SpatialVLM, SpatialBot, and SpatialRGPT develop metric question–answer data, RGB–depth training, and region-level geometric grounding, respectively (Chen et al., 2024; Cai et al., 2025; Cheng et al., 2024), while SpatialLLM and Cambrian-S further strengthen spatial capabilities through model design and spatial training data (Ma et al., 2025b; Yang et al., 2025b). Architectural approaches incorporate 3D positional information, geometric encoders, or reconstruction features into visual representations (Zhu et al., 2025; Wu et al., 2025; Fan et al., 2026). These methods differ in their geometric inputs: LLaVA-3D uses depth and camera information, whereas VLM-3R derives scene structure from monocular video. SpaceMind conditions geometric fusion on camera information, whereas GeoThinker selectively queries geometry according to internal reasoning demands (Zhao et al., 2026; Li et al., 2026b). Interactive 3D exploration, skill-guided tool use, and selective video inspection provide complementary inference-time mechanisms (Zhang et al., 2026b; Li et al., 2026c; Xing et al., 2026). These advances improve spatial encoding and evidence acquisition; we focus on organizing this evidence into an explicit state for scene-level reasoning.

A.2 Cognitive Maps and Explicit Spatial Representations

Persistent spatial representations support reasoning across partial observations. Early systems established joint mapping and planning, spatially addressed memory, and map-based exploration (Gupta et al., 2017; Parisotto & Salakhutdinov, 2018; Chaplot et al., 2020). VLMaps, ConceptFusion, and OpenScene integrate vision–language semantics with geometric representations for language-based navigation and open-vocabulary scene queries (Huang et al., 2023; Jatavallabhula et al., 2023; Peng et al., 2023). Open-Fusion updates posed RGB-D maps online, while ConceptGraphs represents objects and their spatial relations as 3D graphs (Yamazaki et al., 2024; Gu et al., 2024). These semantic maps rely on depth, camera poses, or reconstructed geometry, making their input assumptions distinct from RGB-only video reasoning. For visual question answering, SpatialMind aligns local coordinates into global representations, MindCube learns map-then-reason strategies, and Video2Layout reconstructs metric bird's-eye-view layouts (Zhang et al., 2025a; Wang et al., 2025a; Huang et al., 2025). Map2Thought combines grid and metric representations with question-conditioned evidence selection; Cog3DMap recurrently integrates observations into 3D memory (Gao et al., 2026; Gwak et al., 2026). Building on this line of work, RiCM makes state construction and revision responsive to information needs that emerge during answer reasoning.

B Annotation-Provided Object Inventories in Cognitive Mapping

Supplied object inventories provide a concrete example of the oracle state specification discussed in the main text. We examine the Plain-CGMap-Out prompt in MindCube (Wang et al., 2025a), using the released sample `among_group693_q1_5_2`. Four images show a black sneaker and its surroundings from different viewpoints; the question asks which object lies to the sneaker's right in the second image. The model is instructed to construct a cognitive map on a 10×10 grid before selecting an answer. The complete text prompt is reproduced below, with the four images supplied separately in the original sample order. Bold blue text emphasizes the two instructions that specify what the map should represent; this emphasis is added for presentation.

The object inventory is supplied by dataset annotations before the model constructs the map. Rule 1 explicitly restricts construction to five categories: black sneaker, light purple sofa, brown curtains and windows, TV, and wooden dining table. In the released code, these categories are read from the sample's `meta_info` field and inserted into the instruction. Rule 4 additionally selects the sneaker and sofa as the objects whose facing directions should be estimated. These instructions specify

the categories and orientation attributes to represent, while leaving instance identification, position estimation, orientation estimation, and spatial inference to the model.[1]

Prompt

An instantiated Plain-CGMap-Out prompt from MindCube **[Task]**

Your task is to analyze the spatial arrangement of objects in the scene by examining the provided images, which show the scene from different viewpoints. You will then create a detailed cognitive map representing the scene using a **10x10 grid coordinate system**.

[Rules]

1. Focus ONLY on these categories of objects in the scene: {black sneaker, light purple sofa, brown curtains and windows, TV, wooden dining table}

2. Create a cognitive map with the following structure in the bird's view:

- A 10x10 grid where [0, 0] is at the top-left corner and [9, 9] is at the bottom-right corner
- up = towards the top of the grid (decreasing y)
- right = towards the right of the grid (increasing x)
- down = towards the bottom of the grid (increasing y)
- left = towards the left of the grid (decreasing x)
- Include positions of all objects from the specified categories
- Estimate the center location (coordinates [x, y]) of each instance within provided categories
- If a category contains multiple instances, include all of them
- Object positions must maintain accurate relative spatial relationships
- Combine and merge information from the images since they are pointing to the same scene, calibrating the object locations with grid coordinates accordingly

3. Carefully integrate information from all views to create a single coherent spatial representation.

4. For objects [black sneaker light purple sofa], determine their facing direction as up, right, down, or left. For other objects, omit the facing direction.

[Answer Instruction]

1. Given the provided views and main objects mentioned in the above rules, you **MUST** present your cognitive map in the following JSON format **before your answer**:

```
```json
{
    "object_category_1": {"position": [x, y]},
    "object_category_2": {"position": [x, y], "facing": "direction"},
     # if the object is asked for orientation
    ...
}
```
```

2. Next, provide *ONE* correct answer selecting from the options. Your answer field must be in the format like "A. Above"

3. In general, your response's format should be like "Based on my observation, the answer is:

<cogmap>(Replace with your cogmap here)</cogmap><answer>(Replace with your answer here)</answer>". Your option must be from the available options.

[Question]

Based on these four images (image 1, 2, 3, and 4) showing the black sneaker from different viewpoints (front, left, back, and right), with each camera aligned with room walls and partially capturing the surroundings: From the viewpoint presented in image 2, what is to the right of the black sneaker? A. TV B. Wooden dining table C. Light purple sofa D. Brown curtains and windows

This example illustrates an externally specified representation scope, which is the sense in which we use *oracle relevance judgments* here. The annotation-provided inventory determines the categories the model is asked to cover; it does not establish that each category is necessary for the answer or that the inventory is a minimal sufficient set. In this particular sample, all five categories also appear

[1]We use the released sample, the Plain-CGMap-Out template, and the annotation-to-instruction code, at the linked repository revisions.

in the question or answer options, so the example alone does not demonstrate relevance information unavailable from the question text. It documents the protocol's reliance on a supplied inventory, rather than measuring its benefit. RiCM instead allows the reasoning policy to request entities and attributes and expand the state as inference reveals additional dependencies, making representation scope a decision within the reasoning process.

C Experimental Settings for the Motivation Studies

This section details the settings of the two experiments motivating reasoning-to-state feedback. The first examines how the supplied object inventory affects spatial question answering, while the second compares independent map aggregation with reasoning-guided refinement of object locations. We describe the data, models, comparison conditions, and evaluation metrics for both experiments.

C.1 Object Inventory Comparison

We use VSI-Bench and Qwen3-VL-8B-Instruct to examine how the supplied object inventory affects spatial question answering. For each video–question pair, the model first constructs a cognitive map from the video under the specified inventory and then generates an answer. All conditions share the same examples, video frames, model weights, and inference settings, changing only the inventory supplied to map construction. Inventory entries denote object categories; the model must still distinguish individual instances within each category and infer their spatial relationships from the video.

We compare four inventories. Original inventory uses the categories provided by the dataset annotations. Size-matched random samples the same number of categories from a candidate pool generated by asking Qwen3-VL-8B-Instruct to identify objects in the video, controlling inventory size while changing its content. Question-only contains categories explicitly mentioned in the question, testing whether these mentions provide sufficient support for answering. All detected includes every category in the same model-generated pool, testing whether broader inclusion can substitute for relevance selection. Each inventory remains fixed throughout map construction and answering, so subsequent reasoning cannot compensate for omitted categories by adding them.

Final answers are evaluated using VSI-Bench's task-specific scoring rules, and the overall score is the unweighted mean across its eight tasks. Each task thus contributes equally regardless of its number of questions. We also report the mean number of unique inventory categories per example and divide it by the corresponding mean for the original inventory. This statistic describes the extent of the supplied inventory, rather than directly measuring computational cost. Because the detected pool may omit annotated categories, differences between conditions can reflect candidate coverage as well as relevance selection.

C.2 State Refinement Comparison

We use 200 examples from ScanNet++, with annotated object identities and 3D locations as evaluation references. The comparison is conducted separately for Qwen3-VL-8B-Instruct and Qwen3-VL-32B-Instruct. Within each model, both conditions start from the same initial map and observations. The represented and evaluated objects remain fixed, allowing the comparison to focus on revising existing estimates rather than expanding the map with additional objects. We evaluate the initial map and seven additional rounds of construction or refinement, giving eight evaluation points for each condition.

Independent aggregation constructs a new map at each round without access to previous maps or reasoning feedback. Each construction uses the same question and inventory, but does not receive requests identifying which estimates need verification. After coordinate registration and instance matching, location estimates for the same object are averaged across the initial map and all independently constructed maps available at that round. This baseline tests whether additional observations and repeated construction improve estimates without reasoning-directed revision.

Reasoning-guided refinement uses the current map to identify uncertainties relevant to the question, such as ambiguous object identities or inconsistent spatial relations. It retrieves views that help distinguish competing interpretations and updates the affected estimates while preserving unrelated

map entries. Both conditions access the same video pool and share cumulative limits on visual observations and input/output tokens within each model. Reasoning, retrieval planning, and map updates count toward these limits, so the feedback condition does not receive an additional unaccounted inference budget. Ground-truth annotations are used only for evaluation and do not guide view retrieval or map updates.

C.3 Normalized Positional Accuracy

We define normalized positional accuracy (NPA) to measure the quality of object-location estimates after each round, independently of final-answer correctness. Predicted maps are aligned to the ground-truth coordinate frame using a transform fixed at initialization and shared by both conditions. Keeping this transform unchanged prevents repeated alignment from absorbing positional errors introduced during refinement. Registration anchors are excluded from the evaluated objects.

At each evaluation round, NPA averages the normalized location scores over objects within each example and then equally over all examples:

$$\mathrm{NPA} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_i}\sum_{j=1}^{n_i}\max\left\{0,\, 1-\frac{d_{ij}}{0.2D_i}\right\} \tag{8}$$

Here, $N = 200$ is the number of examples, n_i is the number of evaluated objects in example i, d_{ij} is the Euclidean distance between the aligned prediction and annotated 3D location of object j, and D_i is the diagonal length of the ground-truth scene's bounding box. Each object contributes one at zero error, 0.5 at an error of 10% of the scene diagonal, and zero at an error of 20% or more. Missing objects and objects without a unique match contribute zero and remain included in n_i. Averaging within examples before averaging across them gives every example equal weight. NPA lies between zero and one, with higher values indicating more accurate locations; the evaluated objects and normalization scales remain fixed across conditions and rounds.

D Detailed Experimental Setup

This section details the evaluation setup for the motivation study and the assessment of RiCM. We describe the tasks, evaluation metrics, and visual input protocols of VSI-Bench, ReVSI, and MMSI-Video-Bench, followed by the baseline methods, their model sizes, and pretrained backbones.

D.1 Benchmarks

VSI-Bench (Yang et al., 2025a) evaluates spatial understanding from egocentric videos of indoor environments. It contains over 5,000 question–answer pairs from 288 videos in ScanNet, ScanNet++, and ARKitScenes. Its eight tasks assess spatial configuration, metric estimation, and temporal order, including object counting, distance and direction reasoning, object and room size estimation, route planning, and appearance order. Numerical answers are evaluated using mean relative accuracy (MRA), which averages correctness across relative-error thresholds; multiple-choice answers use accuracy.

ReVSI (Zhang et al., 2026a) improves evaluation validity by correcting scene annotations and aligning questions with the visual evidence available at a given frame budget. It covers 381 scenes from ScanNetv2, ScanNet++, ARKitScenes, 3RScan, and MultiScan. The benchmark retains seven tasks from VSI-Bench, excluding appearance order, and provides question sets tailored to different frame budgets. It follows the MRA and accuracy metrics of VSI-Bench. For RiCM, the policy initially receives 64 uniformly sampled frames and can request additional views through the spatial tool as needed during reasoning.

MMSI-Video-Bench (Lin et al., 2025) extends spatial evaluation to dynamic scenes and reasoning across videos. It contains 1,106 manually authored multiple-choice questions grounded in 1,278 clips from 25 public datasets and in-house recordings. Its 13 subtypes span five categories: spatial construction, motion understanding, planning, prediction, and cross-video reasoning. Cross-video tasks require integrating multiple viewpoints or updating scene memory across clips. Evaluation uses exact-match accuracy under Uniform-50 or Sufficient-Coverage, with the latter providing all frames

used during annotation. For RiCM, we allocate an initial budget of 50 frames across all clips in each question; the policy can then request additional views through the spatial tool as needed during reasoning.

D.2 BASELINES

We compare RiCM with four categories of baselines: proprietary multimodal models, general-purpose open-source models, spatial-centric models with specialized training or architectures, and spatial reasoning agents. Table 3 summarizes one representative configuration for each method, including its nominal model size, pretrained language or vision-language backbone, and backbone size.

Table 3: Baseline methods and their pretrained backbones.

Methods	Nominal Size	Pretrained Backbone	Backbone Size
Proprietary Models (API)			
Doubao-Seed-1.6-vision (Volcengine, 2025)	—	—	—
GPT-5 (OpenAI, 2025a)	—	—	—
GPT-5.2 (OpenAI, 2025b)	—	—	—
Gemini 2.5 Pro (Google DeepMind, 2025a)	—	—	—
Gemini 3 Flash (Google DeepMind, 2025b)	—	—	—
Gemini 3 Pro (Google DeepMind, 2026)	—	—	—
Open-source Models			
Qwen2.5-VL-Instruct (Bai et al., 2025b)	7B	Qwen2.5	7B
Qwen3-VL-Instruct (Bai et al., 2025a)	8B	Qwen3	8B
InternVL3.5 (Wang et al., 2025b)	8B	Qwen3	8B
LLaVA-OneVision (Li et al., 2025)	7B	Qwen2-Instruct	7B
LLaVA-Video (Zhang et al., 2025b)	7B	Qwen2-Instruct	7B
Spatial-centric Models			
SpatialLadder (Li et al., 2026d)	3B	Qwen2.5-VL-Instruct	3B
Spatial-MLLM (Wu et al., 2025)	4B	Qwen2.5-VL-Instruct	3B
VG-LLM (Zheng et al., 2025)	8B	Qwen2.5-VL-Instruct	7B
SpaceR (Ouyang et al., 2025)	7B	Qwen2.5-VL-Instruct	7B
Cambrian-S (Yang et al., 2025b)	7B	Qwen2.5-Instruct	7B
VLM-3R (Fan et al., 2026)	7B	LLaVA-Video-Qwen2	7B
GeoThinker (Li et al., 2026b)	8B	Qwen3-VL-Instruct	8B
SpaceMind (Zhao et al., 2026)	8B	InternVL3	8B
Spatial Reasoning Agents			
Think3D (Zhang et al., 2026b)	4B	Qwen3-VL-Instruct	4B
OmniAgent (Xing et al., 2026)	7B	Qwen2.5-Omni	7B
Skill-3D (Li et al., 2026c)	8B	Qwen3-VL-Instruct	8B

E ADDITIONAL EXPERIMENTAL RESULTS

This section supplements the main experiments with broader benchmark comparisons, an empirical hyperparameter sensitivity analysis, and an evaluation of inference cost. We first expand the MMSI-Video-Bench comparison to a broader set of models (Section E.1), followed by VSI-Bench results and an examination of task-level gains across VSI-Bench and ReVSI (Section E.2). We then evaluate the sensitivity of PSST to its sufficiency and interaction-cost weights (Section E.3). Finally, we compare RiCM's performance and runtime with the backbone and mandatory single-pass mapping to assess the computational cost of adaptive cognitive mapping (Section E.4).

E.1 EXTENDED EVALUATION ON MMSI-VIDEO-BENCH

Due to space constraints, the main text presents only a subset of the MMSI-Video-Bench comparisons. We provide results for the full set of 25 methods in Table 4, spanning proprietary models, general-purpose open-source models, spatial-centric models, and spatial reasoning agents. The table follows

the main-text reporting format, with six task categories and an overall score. Overall scores retain the original aggregation and are not recomputed from the six displayed categories.

RiCM achieves the highest Overall score among the listed methods (38.5), exceeding its Qwen3-VL-8B-Instruct backbone by 10.9 points and the strongest listed open-source baseline, VLM-3R-7B, by 4.2 points. Among open-source methods, RiCM leads on four of the six reported task categories: attribute, instance–scene, scene–scene, and planning. Two task-level gaps remain: RiCM scores 37.1 on instance–instance reasoning, compared with 40.2 for GeoThinker-8B, and 39.6 on prediction, compared with 40.2 for InternVL2.5-38B.

Table 4: Comparison on MMSI-Video-Bench across six task categories and overall performance. Red and blue highlight column-wise best results among proprietary and open-source methods, respectively. Higher scores are better.

Methods	Attribute	Inst.-Inst.	Inst.-Scen.	Scen.-Scen.	Planning	Prediction	Overall
Proprietary Models (API)							
GPT-5	44.6	38.2	44.2	39.1	34.7	28.1	36.8
Gemini 3 Pro	44.6	39.5	44.2	33.3	38.7	35.4	38.0
Gemini 2.5 Flash	37.4	38.2	40.3	30.4	30.6	24.4	35.4
Doubao-Seed-1.6-vision	43.1	26.7	16.7	38.6	37.3	25.9	34.9
Open-source Models							
InternVL3-8B	33.7	32.9	29.9	27.5	33.1	26.8	30.4
InternVL2.5-8B	21.7	36.8	31.2	30.4	26.6	28.1	29.1
LLaVA-Video-7B-Qwen2	27.7	28.9	28.6	30.4	21.0	35.4	28.5
Qwen2.5-VL-7B-Instruct	26.5	25.0	29.9	34.8	34.7	28.1	29.7
Qwen3-VL-8B-Instruct	33.7	25.0	24.7	26.1	25.0	29.3	27.6
InternVL2_5-38B	27.7	30.3	26.0	27.5	33.1	40.2	31.0
InternVL3-38B	33.7	36.8	31.2	24.6	24.2	28.1	28.8
Qwen2.5-VL-32B-Instruct	25.3	26.3	31.2	20.3	27.4	31.7	28.6
Qwen3-VL-30B-A3B	31.3	28.9	26.0	24.6	30.6	35.4	29.1
InternVL2_5-78B	33.7	36.8	37.7	18.8	35.5	26.8	31.4
InternVL3-78B	43.4	31.6	35.1	23.2	27.4	24.4	32.5
LLaVA-Video-72B-Qwen2	36.1	39.5	28.6	21.7	29.0	29.3	30.4
Qwen2.5-VL-72B-Instruct	32.5	18.4	37.7	29.0	26.6	30.5	32.7
Spatial-centric Models							
Cambrian-S-7B	34.8	38.5	38.3	20.2	36.9	28.6	32.5
VLM-3R-7B	44.1	33.4	36.3	24.1	28.0	25.7	34.3
GeoThinker-8B	37.7	40.2	30.5	22.8	30.3	31.1	31.0
SpaceMind-8B	33.4	20.1	38.9	30.8	28.5	31.3	34.2
Spatial Reasoning Agents							
Think3D-4B	31.9	35.7	35.2	18.3	33.1	25.4	30.6
OmniAgent-7B	26.4	28.1	25.3	25.8	30.3	39.1	29.2
Skill-3D-8B	19.2	36.1	29.6	28.2	25.5	25.4	28.3
RiCM-8B (Ours)	44.9	37.1	41.8	35.6	40.8	39.6	38.5

E.2 VSI-Bench Performance and Cross-Benchmark Generalization

RiCM is among the strongest methods on VSI-Bench (Yang et al., 2025a), achieving an overall score of 67.2, only 0.3 points below the highest listed score of 67.5 (Table 5). This result complements the ReVSI evaluation in the main text and allows us to examine whether competitive aggregate scores reflect improvements in the global spatial reasoning capabilities targeted by our method.

Similar overall scores can conceal different strengths across local perception and global reasoning. Cambrian-S and SpaceMind outperform RiCM on object-size and room-size estimation, tasks on which modern MLLMs already exceed the reported human reference in Figure 1. RiCM instead delivers substantial gains on relative direction and route planning, which require cross-view integration and relational composition and remain central bottlenecks in global spatial reasoning. Relative to its Qwen3-VL-8B backbone, RiCM improves these two tasks by 32.5 and 24.9 points on VSI-Bench, respectively, while increasing the overall score by 11.3 points.

The ReVSI results further distinguish benchmark-specific strengths from improvements that persist across evaluations (Table 1). For example, Cambrian-S exceeds RiCM by 13.5 points on room-size estimation in VSI-Bench, but trails it by 7.9 points on ReVSI; its object-size advantage also narrows from 8.7 to 1.4 points. These changes suggest that some local-perception gains on VSI-Bench transfer poorly, consistent with benchmark-specific fitting rather than uniformly generalizable spatial competence. They do not, however, establish overfitting independently of changes in the evaluation protocol. RiCM retains gains of 34.0 points in relative direction and 23.7 points in route planning over its backbone on ReVSI, together with an overall improvement of 11.1 points. This consistent improvement across both benchmarks supports the effectiveness of reasoning-guided cognitive mapping for the global reasoning bottleneck identified in the Introduction.

Table 5: Spatial reasoning performance of models and agent-based methods on VSI-Bench. Red and blue indicate the best result among proprietary and non-proprietary methods, respectively.

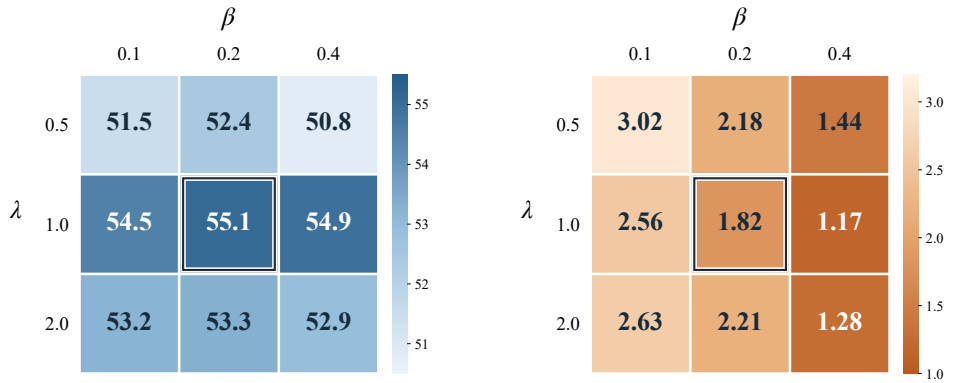
Methods	Obj. Cnt	A. Dist.	Obj. Size	R. Size	Rel.Dist.	Rel.Dir.	R. Plan	A. Order	Overall
Proprietary Models (API)									
Seed-1.6	43.5	34.3	66.1	52.8	55.0	35.7	44.3	67.9	49.9
GPT-5.2	57.1	33.4	64.6	59.0	48.0	33.3	36.7	68.9	49.2
Gemini-2.5-Pro	46.0	37.3	68.7	54.3	61.9	43.9	47.4	68.7	53.5
Gemini-3-Pro	45.3	38.3	73.0	47.4	70.0	60.8	65.3	69.8	60.5
Open-source Models									
Qwen2.5-VL-7B	40.9	14.8	43.4	10.7	38.6	38.5	33.0	29.8	33.0
Qwen3-VL-8B	63.5	47.6	74.8	59.0	57.8	50.9	29.9	63.9	55.9
LLaVA-OneVision-7B	47.7	20.2	47.4	12.3	42.5	35.2	29.4	24.4	32.4
LLaVA-NeXT-Video-7B	48.5	14.0	47.8	24.2	43.5	42.4	34.0	30.6	35.6
Qwen2.5-VL-72B	25.1	29.3	54.5	38.8	38.2	37.0	34.0	28.9	37.0
Qwen3-VL-32B	74.0	57.0	76.6	70.8	55.6	59.1	39.7	65.2	61.8
LLaVA-OneVision-72B	43.5	23.9	57.6	37.5	42.5	39.9	32.5	44.6	40.2
LLaVA-NeXT-Video-72B	48.9	22.8	57.4	35.3	42.4	36.7	35.0	48.6	40.9
Spatial-centric Models									
Cambrian-S-7B	73.2	50.5	74.9	72.2	71.1	76.2	41.8	80.1	67.5
VLM-3R-7B	70.2	49.4	69.2	67.1	65.4	80.5	45.4	40.1	60.9
GeoThinker-8B	74.0	57.0	76.6	60.8	65.6	79.1	49.7	65.2	64.1
SpaceMind-8B	71.1	59.1	74.8	71.8	62.5	83.6	39.4	68.2	66.3
Spatial Reasoning Agents									
Think3D-4B	41.5	29.4	44.2	48.7	29.6	44.1	30.8	52.2	40.1
OmniAgent-7B	49.9	33.6	58.5	44.3	47.1	40.7	34.0	59.7	46.0
Skill-3D-8B	56.5	48.6	59.8	67.9	52.0	60.1	58.4	66.8	58.8
RiCM-8B (Ours)	72.8	59.2	66.2	58.7	72.1	83.4	54.8	70.4	67.2

E.3 Hyperparameter Sensitivity of PSST

We evaluate the sensitivity of PSST to the sufficiency weight λ and interaction-cost weight β by jointly varying $\lambda \in \{0.5, 1.0, 2.0\}$ and $\beta \in \{0.1, 0.2, 0.4\}$ around the default configuration $(\lambda, \beta) = (1.0, 0.2)$. Each configuration follows the RQ4 protocol with 10K training examples and at most seven state updates per trajectory, allowing early stopping. We report the official ReVSI Overall score and the average number of executed EXPAND and VERIFY actions, excluding initial state construction and reader evaluations.

Figure 10 shows that answer performance is sensitive to the sufficiency reward weight. At $\beta = 0.2$, increasing λ from 0.5 to 1.0 raises the Overall score from 52.4 to 55.1, whereas further increasing it to 2.0 lowers the score to 53.3. The setting $\lambda = 1.0$ also achieves the highest score at both other penalty weights, supporting a moderate sufficiency reward within the tested range.

Increasing the interaction penalty consistently reduces the average number of state updates across all tested values of λ. At $\lambda = 1.0$, increasing β from 0.2 to 0.4 reduces updates from 1.82 to 1.17 (35.7%), while the Overall score decreases from 55.1 to 54.9 by 0.2 points. The default configuration achieves the highest Overall score among the nine tested configurations, while the larger penalty offers fewer state updates with a small reduction in answer performance. These results support the default weights and demonstrate a tunable trade-off between answer performance and interaction cost.



(a) ReVSI Overall ↑ (b) Average Updates ↓

Figure 10: Joint sensitivity of PSST to the sufficiency weight λ and interaction-cost weight β under the RQ4 training protocol. Rows vary λ and columns vary β. Cells report (a) ReVSI Overall and (b) average executed state updates; darker colors indicate higher scores and fewer updates, respectively. Outlined cells mark the default configuration.

E.4 Evaluation Speed of RiCM

RiCM controls the cost of cognitive mapping through adaptive tool invocation. For some local-perception questions, the policy answers directly from visual observations without calling the spatial tool to generate a cognitive map. When reasoning requires additional spatial information, the policy can instead request map construction and subsequent updates. This selective invocation avoids imposing map-generation costs on every question. To assess the resulting performance and runtime, we evaluate Qwen3-VL-8B-Instruct, the same backbone with a mandatory single map-generation pass per question, and RiCM-8B on the full ReVSI evaluation set using eight NVIDIA H200 GPUs.

Table 6: Performance and inference efficiency on ReVSI. Time denotes total runtime on the full evaluation set, normalized to the backbone in Rel. time. Blue indicates the best value in each metric.

Methods	**Mapping**	**GPUs**	**Overall ↑**	**Time (h) ↓**	**Rel. time ↓**
Qwen3-VL-8B-Instruct	None	8× NVIDIA H200	49.1	1.81	1.00×
Qwen3-VL-8B-Instruct	Single-pass	8× NVIDIA H200	48.4	2.55	1.41×
RiCM-8B (Ours)	Iterative	8× NVIDIA H200	60.2	3.36	1.86×

With this adaptive invocation strategy, RiCM achieves substantial performance gains at a total runtime of 1.86× that of the backbone (Table 6). The overall ReVSI score increases from 49.1 to 60.2, while evaluation time rises from 1.81 to 3.36 hours. Mandatory single-pass mapping, by comparison, takes 2.55 hours but scores 48.4, providing no improvement over the backbone. RiCM therefore gains 11.8 points over this fixed mapping configuration with an additional 0.81 hours (48.6 minutes) of evaluation time.

This performance–runtime trade-off is consistent with the selective use of cognitive mapping described above. Direct answers to some local-perception questions avoid unnecessary map-generation costs, while questions requiring additional spatial evidence can trigger map construction and revision. The aggregate runtime thus reflects a mixture of direct answering and tool-assisted reasoning, rather than repeated mapping for every question. This flexibility helps contain the cost of iterative inference while enabling the substantial gains observed on ReVSI.

F Prompts for the Policy and Map Generator

This section presents two general prompt templates that specify the interaction between reasoning and cognitive mapping in RiCM. The policy prompt in Section F.1 directs the search for missing or

uncertain spatial information and determines when to answer. The cognitive-map generator prompt in Section F.2 translates the retrieved visual and geometric evidence into a candidate state for verification. Together, these templates specify how information requests lead to map updates and how unresolved state variables guide subsequent reasoning. Each module reuses its template across interaction steps with updated inputs.

For readability, the serialized fields `support`, `map`, and `unresolved` denote $\mathcal{S}_i$, $\mathcal{M}_i$, and $\mathcal{U}_i$, respectively; `targets`, `relations`, and `mode` denote the components of ρ_i. Angle-bracketed input slots are populated by the calling code before each model call, and visual inputs are attached with their view identifiers. Both models return textual JSON. The spatial tool retains visual evidence and its view/frame references; these are not part of the generated responses.

F.1 Policy Prompt

The policy receives the question q, sampled visual context $\widetilde{V}$, and current belief $\mathcal{B}_i$, together with the remaining interaction budget. Its prompt distinguishes missing state content from uncertain estimates: Expand requests absent variables, whereas Verify seeks evidence that distinguishes unresolved hypotheses. Both are modes of Search; the policy selects Answer when the state supports answering and the remaining uncertainty would not change the result. The following template defines this decision process and the corresponding request format.

General prompt for the policy

[Task]

Guide the construction of a spatial map sufficient to answer the question. At each step, choose Search to request evidence from the spatial tool or Answer to finish. The tool retrieves additional views and updates the belief.

[Inputs]

Question: `<question and any answer choices>`

Answer format: `<required response format and units>`

Visual context: `<uniformly sampled frames with view IDs>`

Current belief: `<serialized support, map, and unresolved>`

Coordinate convention: `<frame and viewpoint metadata>`

Remaining updates: `<remaining interaction budget>`

[Decision rules]

Read support as the variables selected for reasoning, map as instance and relation estimates with confidence and provenance, and unresolved as missing targets or conflicting hypotheses. A selected variable or unresolved hypothesis is not an established fact.

1. **Expand**: If an answer-critical entity or relation is missing, request it. Include auxiliary entities when needed for cross-view association or instance disambiguation; avoid an exhaustive scene inventory.

2. **Verify**: If an unresolved proposition could change the answer, identify its entry ID, competing hypotheses, and the observation that would distinguish them. Use its conflict type and suggested retrieval target to direct the request.

3. **Answer**: Stop when all answer-critical variables are represented and remaining uncertainty would not change the answer. Otherwise, continue Search while updates remain. At zero budget, return the best-supported answer without treating unresolved hypotheses as verified.

Reassess these choices after each update; Expand and Verify have no fixed order. Use existing IDs for known objects and descriptive selectors for unseen targets. Requests must not presume their conclusions. Sampled views may be incomplete, and answer choices do not establish object presence. Interpret directions using the specified viewpoint and coordinate frame. Leave belief updates to the tool.

[Output]

Return exactly one textual JSON object without visual data, view/frame references, Markdown fences, or extra text:

Search: keys `action`, `mode`, `targets`, `relations`. Set action to `"Search"` and mode to `"Expand"` or `"Verify"`. Targets and relations are arrays of descriptive strings; at least one must be

nonempty. For Verify, include the entry ID, hypotheses, and distinguishing observation in the relevant description.

Answer: keys `action`, `answer`. Set action to `"Answer"` and generate answer in the supplied type, format, and units. Do not combine Search and Answer.

F.2 COGNITIVE-MAP GENERATOR PROMPT

The cognitive-map generator receives the search request ρ_i, current belief $\mathcal{B}_i$, retrieved views $\mathcal{F}_i$, and registered anchors $\mathcal{A}_i$. Its output is a structured candidate map: coordinate metadata defines the shared frame, nodes store entity positions and supported intrinsic orientations, and edges store numerical distances and referenced spatial relations. The spatial tool performs retrieval, IGGT inference, and coordinate registration before generation, then applies the evidence checks and update operator described in Section 3.3. The following template specifies the candidate format; the tool returns the checked state and unresolved feedback to the policy.

General prompt for the cognitive-map generator

[Task]

Construct a structured spatial map addressing the search request. Represent entities with numerical coordinates and supported orientations, and represent their spatial relations with explicit values and references. Ground changes in evidence and preserve uncertainty. Do not answer the original question.

[Inputs]

Request: `<targets, relations, mode>`

Belief: `<serialized support, map, and unresolved>`

Views: `<retrieved and relevant historical images with IDs>`

Anchors: `<registered anchor records and image-region links>`

Coordinates: `<persistent frame, origin, axes, units, scale, and viewpoint references>`

Anchors provide 3D means, covariances, supporting-view IDs, and association evidence; relevant historical anchors are included for verification.

[Construction rules]

1. Select entities and relations needed by the request or unresolved hypotheses. Include auxiliary entities when needed for association or disambiguation, recording their purpose. Reuse supported instance IDs and assign unique IDs to new instances.

2. Express all geometry in the supplied persistent frame. Preserve its origin, axes, and scale across calls; apply only supplied registration transforms. Declare the length unit and scale to meters. If metric scale is unavailable, retain the supplied scene units and set the meter conversion to null. Do not silently recenter, rotate, or normalize the map.

3. Store each entity's 3D position as a numerical array. Store its intrinsic facing direction only when supported by evidence; a bounding-box axis alone does not establish semantic front. Keep intrinsic orientation separate from the observer's viewing heading. For a viewpoint-relative relation, record the observer origin, horizontal unit heading, yaw, and the requested toward/away convention.

4. Derive requested geometric edges from the represented positions and references. Include center-to-center distances for the entity pairs needed by the request. Use `distance_3d` for $\|\mathbf{p}_b - \mathbf{p}_a\|_2$ and `distance_horizontal` for the norm of the displacement projected onto the declared horizontal plane. Use `bearing` for the signed horizontal angle from the viewing heading to the target displacement. Distances use the declared length unit; yaw and bearing use degrees and the declared angular convention. Retain numerical geometry when also reporting qualitative relations.

5. **Expand** proposes missing entities or relations. **Verify** compares targeted estimates with new and historical evidence. After changing a position, association, orientation, or reference, recompute every dependent edge and preserve unrelated reliable records. Explain additions, revisions, and proposed resolutions in update notes.

6. If evidence cannot select a unique value, return null for that field and record the competing candidates in unresolved. Include candidate coordinates, directions, or distances and their units when available. Return null for a derived value if a required input is unavailable or degenerate, such as a zero-length

horizontal heading. A missed detection does not establish absence. Do not choose a hypothesis solely because it matches an answer option.

7. Cite only geometric anchor IDs present in the supplied evidence. Preserve confidence and its source when available; do not invent confidence scores or supporting evidence. The tool checks geometric residuals, association confidence, and relational consistency before accepting, fusing, or downweighting proposed updates.

[Output structure]

Return exactly one valid JSON object, indented with two spaces, using the following structure. Populate the arrays with the records defined below. Omit Markdown fences, explanatory text, images, image-region data, and view/frame references.

```
{
  "support": [],
  "map": {
    "frame": {
      "id": null,
      "origin": null,
      "axes": {"x": null, "y": null, "z": null},
      "horizontal_axes": null,
      "position_unit": null,
      "meters_per_unit": null,
      "angle_unit": "deg",
      "yaw_zero": null,
      "yaw_positive": null
    },
    "nodes": [],
    "edges": []
  },
  "unresolved": []
}
```

support. A string array of selected entity and relation variables. Inclusion selects a variable for reasoning; it does not establish its value.

map.frame. Copy the supplied frame ID, origin description, and axis definitions. `horizontal_axes` is the ordered pair of horizontal axis names. `position_unit` names the length unit; `meters_per_unit` is a positive conversion factor, or null when unknown. Set `angle_unit` to `"deg"`; `yaw_zero` identifies the zero-angle axis and `yaw_positive` specifies the rotation direction about the up axis. Bearing uses the same rotation sign. Unknown metadata remains null, and any quantity requiring it also remains null.

map.nodes. Each record must contain `id`, `description`, `position`, `orientation`, and the evidence fields below. IDs and descriptions are strings. `position` is $[x, y, z]$ as three JSON numbers, or null. `orientation` is null when intrinsic facing is unknown; otherwise it is an object with `forward`, a 3D unit vector in the map frame, and `yaw_deg`, its horizontal yaw. Set yaw to null when the horizontal projection is zero or the angular convention is unavailable. Do not fill intrinsic orientation with a direction toward another entity.

map.edges. Each record must contain `subject`, `object`, `relation`, `value`, `unit`, `reference`, and the evidence fields below. Endpoints are returned node IDs. For `distance_3d`, `distance_horizontal`, and `bearing`, `value` is a JSON number or null, never a sentence; `unit` is the declared length unit or `"deg"`. Distance values must be nonnegative. Other numerical relations use a number or null with the applicable unit. Qualitative relations may use strings or Booleans with unit null.

Edge reference. Every `reference` is an object with `frame_id`, `origin_node`, `heading`, `yaw_deg`, and `definition`. For a distance, use the map frame ID and set the viewpoint-specific fields to null. For a bearing, subject is the observer-origin node, object is the target node, and `origin_node` equals subject. `heading` is a horizontal 3D unit vector in the map frame, `yaw_deg` is its yaw, and `definition` states how the heading is obtained, including any toward/away instruction. Facing away reverses the toward-target heading. For other viewpoint-relative relations, supply the same reference fields. If the requested heading cannot be established, retain its definition but set its numerical fields and dependent values to null.

Evidence fields. Each node and edge must contain `confidence`, `confidence_source`, `anchor_ids`, and `update_note`. Retain numerical confidence and its source only when both are available and apply to the estimate; otherwise set both to null. Anchor references are arrays of supplied

anchor-ID strings, or empty arrays when unavailable. Update notes are strings describing changes and auxiliary dependencies. For a derived edge, identify the input nodes and reference used in the calculation; preserve their supporting anchor IDs.

unresolved. Each record contains `id`, `targets`, `hypotheses`, `conflict_type`, `anchor_ids`, and `retrieval_target`. Preserve existing IDs. Targets and hypotheses are string arrays; numerical alternatives in a hypothesis must identify their target field, frame, and unit. The retrieval target describes the evidence needed to distinguish them. Missing targets may lack nodes and anchor references. Record proposed resolutions by unresolved-entry ID in the corresponding update notes.

[Final checks]

Keep all required keys, using JSON null for unavailable values and empty arrays for no entries. Preserve unchanged spatial records. Check unique node IDs, valid edge endpoints, common coordinate units, normalized direction vectors, and consistency between positions, headings, distances, and bearings. Do not serialize coordinate arrays as strings, replace the map with a narrative summary, or substitute qualitative direction labels for numerical spatial fields.

G Case Study

We illustrate map expansion, local verification, and an incorrect reference heading using three relative-direction questions from ReVSI (Zhang et al., 2026a). Videos, questions, and answers come from the released benchmark; coordinates derive from official 3D bounding-box centers. The state transitions and verification outcomes are scripted, not recorded model runs. The traces use one, two, and two interactions, yielding two correct answers and one failure, and serve to explain the workflow rather than estimate performance.

Each panel includes 16 chronological preview frames, distinct from retrieval results and the policy's 64-frame initial context. Green shading marks decisive updates and correct predictions; red marks the erroneous heading and prediction. Maps use a fixed scene frame with horizontal axes x, y and vertical axis z. Positions and 3D distances are in meters. Yaw ψ is measured from $+x$, and bearing β from the observer heading $\mathbf{h}$, both counterclockwise in degrees. Values are rounded; unknowns use `null`. Unavailable runtime metadata uses `confidence=null`, `confidence_source=null`, and `anchor_ids=[]`.

G.1 Case 1: Completing missing state support

The initial map contains the floor lamp and wall picture but lacks the fan's location. One expansion adds the target to their shared frame, making its direction computable.

Case 1: Relative direction: map expansion (ReVSI 4808) **Success**

[Visual input]



[Question]

If I am standing by the floor lamp and facing the wall picture, is the standing fan to my left, right, or back?

An object is to my back if I would have to turn at least 135 degrees in order to face it.

Options A. back B. right C. left

[Ground truth]

B. right

Case 1: Relative direction: map expansion (ReVSI 4808, continued) **Success**

[Environment return 1: EXPAND]

Support $\mathcal{S}_1$ Floor lamp, wall picture, standing fan; viewing heading, pairwise distances, and fan bearing.

Map $\mathcal{M}_1$ Nodes in the fixed scene frame:

ID / Entity	Position (x, y, z) [m]	Status
19 / Floor lamp	$(-0.331, \quad 3.361, \quad 0.167)$	Retained
27 / Wall picture	$(\quad 0.355, \quad 0.938, \quad 0.787)$	Retained
26 / Standing fan	$(-2.954, \quad 2.916, -0.073)$	Added

Heading Origin: node 19; face node 27.
$\mathbf{h} = (0.272, -0.962, 0), \quad \psi = -74.19^\circ.$

Distance edges [m] $d(19, 27) = 2.594, \quad d(19, 26) = 2.670, \quad d(27, 26) = 3.950.$

Bearing edge $\beta(19, 26 \mid \mathbf{h}) = -96.18^\circ.$
Unresolved $\mathcal{U}_1$ $\varnothing$. The target location and its geometric relations are available.

[Final prediction]

B. right

The fan's bearing is -96.18°, corresponding to a clockwise turn below the question's 135° threshold for back. The correct answer is therefore B, right. Adding the fan supplies the missing information while preserving the reference objects and heading. With the requested relation determined and no recorded uncertainty remaining, the scripted policy stops after one update.

G.2 CASE 2: RESOLVING A CONFLICT THROUGH LOCAL VERIFICATION

The refrigerator has two candidate locations mirrored across the table-to-microwave axis. This constructed ambiguity leads to different answer options and motivates verification using the oven as an auxiliary association reference.

Case 2: Relative direction: local verification (ReVSI 5511) **Success**

[Visual input]



[Question]

If I am standing by the table and facing the microwave, is the refrigerator to my front-left, front-right, back-left, or back-right?

Options A. front-left B. back-left C. front-right D. back-right

[Ground truth]

A. front-left

[Environment return 1: EXPAND]

Support $\mathcal{S}_1$ Table, microwave, refrigerator; viewing heading, pairwise distances, and refrigerator bearing.

Case 2: Relative direction: local verification (ReVSI 5511, continued) **Success**

Map $\mathcal{M}_1$ Nodes in the fixed scene frame:

ID / Entity	Position (x, y, z) [m]	Status
14 / Table	$(-3.361, \quad 1.651, -0.661)$	Retained
19 / Microwave	$(-5.010, -0.918, -0.010)$	Retained
0 / Refrigerator	`null`	Unresolved

Heading Origin: node 14; face node 19.
$\mathbf{h} = (-0.540, -0.842, 0), \quad \psi = -122.69°$.

Distance edges [m] $d(14, 19) = 3.122$; $\quad d(14, 0) = d(19, 0) = \texttt{null}$.

Bearing edge $\beta(14, 0 \mid \mathbf{h}) = \texttt{null}$.

Unresolved $\mathcal{U}_1$ U1: conflicting cross-view locations for node 0. The two hypotheses are:

Hypothesis	Candidate position [m]	Bearing
H_{L}: front-left	$(-2.070, -1.531, -0.095)$	$+54.77°$
H_{R}: front-right	$(-6.792, \quad 1.499, -0.095)$	$-54.77°$

Retrieval target: Complementary views and historical supporting views with shared anchors linking the refrigerator, table, and microwave. Neither candidate is committed to the map.

[Environment return 2: VERIFY]

Support $\mathcal{S}_2$ Previous variables, plus the oven as an auxiliary association anchor.

Map $\mathcal{M}_2$ The refrigerator position is resolved; the oven is added:

ID / Entity	Position (x, y, z) [m]	Status
14 / Table	$(-3.361, \quad 1.651, -0.661)$	Retained
19 / Microwave	$(-5.010, -0.918, -0.010)$	Retained
0 / Refrigerator	$(-2.070, -1.531, -0.095)$	Resolved
2 / Oven	$(-3.575, -1.245, -0.446)$	Auxiliary

Heading Origin: node 14; face node 19 (unchanged).
$\mathbf{h} = (-0.540, -0.842, 0), \quad \psi = -122.69°$.

Distance edges [m] $d(14, 19) = 3.122, \quad d(14, 0) = 3.481, \quad d(19, 0) = 3.004$;
$d(2, 14) = 2.912, \quad d(2, 19) = 1.535, \quad d(2, 0) = 1.572$.

Bearing edge $\beta(14, 0 \mid \mathbf{h}) = +54.77°$.
Unresolved $\mathcal{U}_2$ $\varnothing$. U1 is resolved in favor of H_{L}; H_{R} is discarded.

[Final prediction]

A. front-left

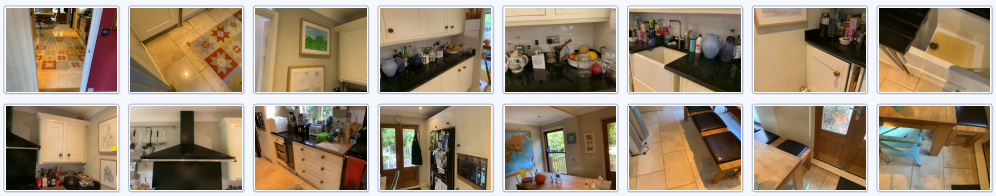
The candidates have bearings of $\pm 54.77°$ but identical distances to the table and microwave, so those distances alone cannot resolve left versus right. The first return retains both hypotheses without committing a location. The scripted verification then resolves the refrigerator position through cross-view association with the oven, yielding A, front-left. The reference objects and heading remain unchanged, illustrating a local update supported by an auxiliary entity absent from the question.

G.3 CASE 3: VERIFICATION UNDER AN INCORRECT REFERENCE HEADING

The policy uses a heading toward the telephone, although the question requires the opposite direction. Verification checks the wall-clock association but preserves this incorrect heading.

Case 3: Relative direction: opposite heading (ReVSI 5872) **Failure**

[Visual input]



[Question]

If I am standing by the room heater and facing in the opposite direction of the telephone, is the wall clock to my front-left, front-right, back-left, or back-right?

Options A. back-right B. back-left C. front-left D. front-right

[Ground truth]

D. front-right

[Environment return 1: EXPAND]

Support $\mathcal{S}_1$ Room heater, telephone, wall clock; viewing heading, pairwise distances, and clock bearing.

Map $\mathcal{M}_1$ The clock has a provisional association:

ID / Entity	Position (x, y, z) [m]	Status
32 / Room heater	$(4.307, \quad 1.144, -1.004)$	Retained
25 / Telephone	$(1.191, -0.644, -0.161)$	Retained
26 / Wall clock	$(7.262, -0.144, \quad 0.859)$	Provisional

Heading Origin: node 32; face node 25.
$\mathbf{h} = (-0.867, -0.498, 0), \quad \psi = -150.15^\circ.$

Distance edges [m] $d(32, 25) = 3.689, \quad d(32, 26) = 3.723, \quad d(25, 26) = 6.176.$

Bearing edge $\beta(32, 26 \mid \mathbf{h}) = +126.60^\circ$ (provisional).

Unresolved $\mathcal{U}_1$ U1: wall-clock association. Hypotheses: the views observe the same clock, or an anchor is misassociated. The clock position and its incident edges remain provisional.

Retrieval target: Overlapping views with shared room references to verify the clock identity.

[Environment return 2: VERIFY]

Support $\mathcal{S}_2$ Unchanged.

Map $\mathcal{M}_2$ The clock association is confirmed:

ID / Entity	Position (x, y, z) [m]	Status
32 / Room heater	$(4.307, \quad 1.144, -1.004)$	Retained
25 / Telephone	$(1.191, -0.644, -0.161)$	Retained
26 / Wall clock	$(7.262, -0.144, \quad 0.859)$	Verified

Heading Origin: node 32; face node 25 (unchanged).
$\mathbf{h} = (-0.867, -0.498, 0), \quad \psi = -150.15^\circ.$

Distance edges [m] $d(32, 25) = 3.689, \quad d(32, 26) = 3.723, \quad d(25, 26) = 6.176.$

Bearing edge $\beta(32, 26 \mid \mathbf{h}) = +126.60^\circ.$

Unresolved $\mathcal{U}_2$ $\varnothing$. U1 is resolved: the observations correspond to the same wall clock.

Case 3: Relative direction: opposite heading (ReVSI 5872, continued) **Failure**

[Final prediction]

B. back-left

The retained heading gives a bearing of $+126.60°$ and the incorrect answer B, back-left. Reversing the heading as required changes the bearing to $-53.40°$, yielding D, front-right without altering any coordinates. Thus, clearing the recorded association uncertainty does not ensure that the question's directional constraint has been applied. This failure requires correcting the heading, not further association checks under the same reference.