SPATIALBRAIN: SPATIAL INTELLIGENCE GROWING FROM EXPERIENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial reasoning requires more than access to geometric tools: a reasoner must decide when direct observation is sufficient and when to reconstruct, measure, or change reference frames. We present SpatialBrain, a framework that retains experience about these decisions while keeping its vision-language backbone fixed. The key distinction is between evidence about a scene and experience about how to reason with that evidence. 3DFusionMemory fuses current-scene observations into queryable spatial relations. A Spatial Skill Library provides reusable tool procedures, and Spatial Cognition associates task conditions with experience about when and how to use memory and skills. Completed trajectories are abstracted into process records and consolidated between task blocks, allowing later decisions to draw on earlier execution without transferring scene coordinates or reference answers. Across five spatial-reasoning benchmarks, SpatialBrain improves on direct answers from the same backbone on four and ties on the fifth. Component comparisons, experience-coverage curves, and traced resource decisions characterize how the framework accumulates and uses these resources.

1 INTRODUCTION

The same visual scene can pose very different reasoning problems. A visible attribute may be read directly, an object-relative direction requires a reference frame, and a distance estimate may require geometric evidence from several views. Route questions additionally require keeping track of successive movements. Spatial benchmarks expose these differences through multi-image reasoning, limited-view scene construction, dynamic spatial aptitude, and degraded observations Yang et al. (2025); Wang et al. (2025); Jia et al. (2025); Ray et al. (2024); Zhou et al. (2026a). A useful reasoner must therefore choose the computation as well as produce the answer. Reconstructing every scene can add unnecessary cost and noisy evidence, while answering directly can overlook information that must be integrated across views.

Spatial representations learned through training Chen et al. (2024); Cheng et al. (2024) and tools for perception, reconstruction, and simulation Marsili et al. (2025); Chen et al. (2025); Cho et al. (2026); Zhu et al. (2026) expand the computations available to a reasoner. Agent methods also retain reflections and reusable workflows from earlier tasks Shinn et al. (2023); Wang et al. (2023); Zhao et al. (2023); Wang et al. (2024). Access to a resource, however, does not specify when it is useful. A procedure for computing relative directions can be reused across scenes; the coordinates on which it previously operated cannot. The challenge is to retain experience that informs resource use while grounding each answer in the current scene.

SpatialBrain separates scene evidence, reusable computation, and experience about resource use. It operates as an agent: a vision-language reasoner that can select and call tools. After completing tasks, it retains abstract records of the conditions encountered, resources used, and execution feedback. Later tasks consult these records to choose a reasoning route, memory-access mode, or skill. We call this cross-task adaptation *self-evolution*. The model weights remain fixed; the external experience available to the controller changes. Unlike correcting an answer within one task, self-evolution changes the state available to subsequent tasks.

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.

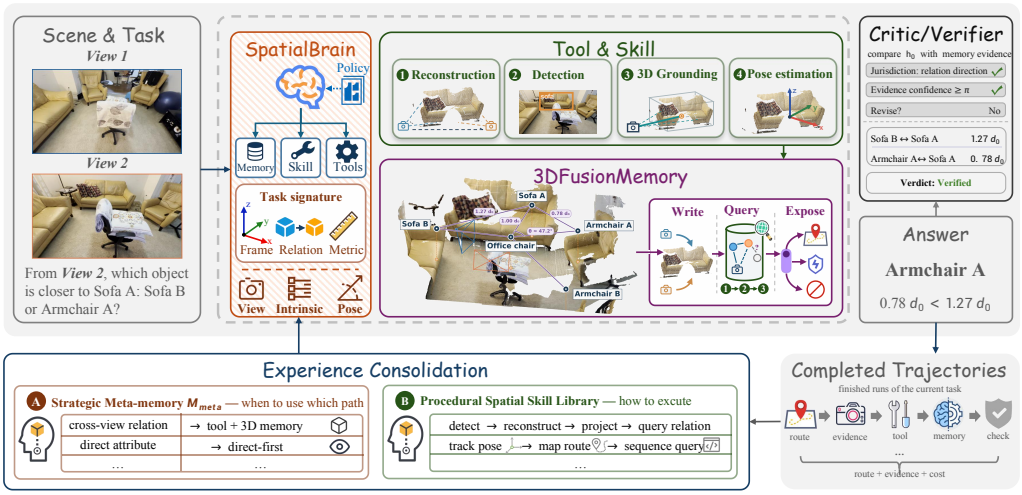


Figure 1: **SpatialBrain: spatial reasoning with experience reuse.** Spatial Cognition chooses how to use direct reasoning, spatial skills, and scene evidence. After execution, abstract process records update procedures and resource-use experience for later tasks. 3DFusionMemory is rebuilt for each scene, and model weights remain fixed. The scene is illustrative.

Our core contributions address three parts of this reasoning process (Figure 1):

- **3DFusionMemory** grounds reasoning in a shared scene state by fusing object, pose, camera, and route observations into queryable spatial evidence with supporting views and uncertainty.
- **Spatial Skill Library** makes recurring spatial computations reusable through parameterized tool procedures, separating the operation from the scene-specific inputs on which it acts.
- **Spatial Cognition** accumulates condition-specific experience about when and how to use memory and skills. Past process feedback and costs inform resource selection, connecting completed trajectories to later reasoning decisions.

Experience is updated only after a block of test tasks finishes, without reference answers. Every task within the block reads the same prior experience and maintains its own scene memory. This makes the temporal dependence explicit even when tasks are processed in parallel.

Evaluation with GPT-5.6-sol shows improvements over direct answers on four of five benchmarks and a tie on the fifth. We first examine these gains and their distribution across task types, then analyze component configurations, the accumulation of experience, and its use in later trajectories. The execution traces connect earlier experience to changes in resource decisions.

2 Related Work

Spatial intelligence. SpatialVLM and SpatialRGPT develop spatial representations through targeted supervision Chen et al. (2024); Cheng et al. (2024), while Thinking in Space studies cognitive maps for video reasoning Yang et al. (2024a). Tool-based approaches expose visual programs, geometric constraints, stateful computation, and simulation to a reasoner Hu et al. (2024b); Marsili et al. (2025); Chen et al. (2025); Cho et al. (2026); Zhu et al. (2026). Structured representations include S-Agent's hierarchical spatial memory and AlloSpatial's allocentric maps Dai et al. (2026); Ruan et al. (2026). SpatialCLI provides specialist tools, and SATURN and OmniView-Space emphasize perspective-aware reasoning Zhou et al. (2026b); Kamali et al. (2026); Li et al. (2026c).

These approaches motivate both the scene representations and computations available to SpatialBrain. Our focus is the complementary problem of using execution experience to decide when and how to invoke them.

Agent skills and memory. Executable skills and recurring workflows provide reusable computation, as in Voyager and Agent Workflow Memory Wang et al. (2023; 2024). Skill-3D distills scene-aware tool workflows and failure lessons from rollouts and uses skill-guided trajectories for agentic post-training Li et al. (2026a). SpatialBrain instead studies blockwise, reference-label-free resource adaptation with fixed model weights, separating current-scene geometry from persistent resource-use experience. Episodic reflection and linked memory notes retain broader experience Park et al. (2023); Xu et al. (2025), whereas object-centric scene graphs and exploration snapshots retain spatial evidence Gu et al. (2023); Yang et al. (2024b). In SpatialBrain, the Skill Library stores *how to execute* a procedure, strategic experience records *when and how to use* it, and 3DFusionMemory represents *what the current scene supports*.

Self-evolving agents. Self-evolving agents differ in what they update. Reflexion and ExpeL retain feedback or distilled experience for later tasks Shinn et al. (2023); Zhao et al. (2023); Agentic Context Engineering consolidates contextual playbooks through generation, reflection, and curation Zhang et al. (2025b). Automated Design of Agentic Systems searches agent programs using an expanding archive Hu et al. (2024a), while the Darwin Gödel Machine modifies and tests agent code Zhang et al. (2025a). SpatialBrain updates external spatial strategies and procedural records from completed test-time trajectories. Its backbone stays fixed, and scene evidence remains task-specific. We examine the progression from retained experience to retrieval, execution, and resource decisions.

3 SPATIALBRAIN

3.1 OVERVIEW

Given images $I_{1:K}$, a question q, and optional answer choices O, SpatialBrain solves a task $x_i = (I_{1:K}, q, O)$ using a fixed vision-language backbone and perception and geometry tools. Spatial Cognition selects the reasoning route and access to memory and skills. Tool observations populate the scene memory, and skills carry out reusable computations on the current inputs. An execution trajectory records these decisions, calls, returned evidence, and process checks. Completed trajectories supply experience for later tasks without updating model weights, as illustrated in Figure 2.

This separation gives two timescales of state: *scene evidence* changes within a task, while *experience about using evidence* changes between task blocks. Let i index tasks, t a within-task execution step, and b a block. The state available to the reasoner is

$$\mathcal{S}_{i,t} = (\mathcal{G}_{i,t}, \mathcal{H}_b), \qquad \mathcal{H}_b = (\mathcal{L}_b, \mathcal{P}_b). \tag{1}$$

Here $\mathcal{G}_{i,t}$ is the current scene graph, $\mathcal{L}_b$ the skill library, and $\mathcal{P}_b$ the strategic experience used by Spatial Cognition. $\mathcal{H}_b$ groups the two persistent stores. In equation 1, observations update $\mathcal{G}_{i,t}$ during execution; all tasks in a block read the same committed snapshot $\mathcal{H}_b$ and keep separate scene graphs. A procedure for comparing object-relative directions can thus persist without carrying its earlier object coordinates into a new scene.

3.2 3DFUSIONMEMORY

Different views expose different objects and different sides of the same object. 3DFusionMemory fuses these observations into a shared scene frame, where object identities connect evidence across views and spatial relations connect objects. The memory for the current task is

$$\mathcal{G}_{i,t} = (\mathcal{V}_{i,t}, \mathcal{E}_{i,t}, \mathcal{C}_{i,t}, \mathcal{R}_{i,t}), \tag{2}$$

where $\mathcal{V}_{i,t}$ contains object nodes, $\mathcal{E}_{i,t}$ spatial relations, $\mathcal{C}_{i,t}$ camera states, and $\mathcal{R}_{i,t}$ ordered route history. Equation 2 associates view-dependent observations with object identities in a common scene frame. Nodes retain detections, 3D centers, available point sets and extents, pose, semantic

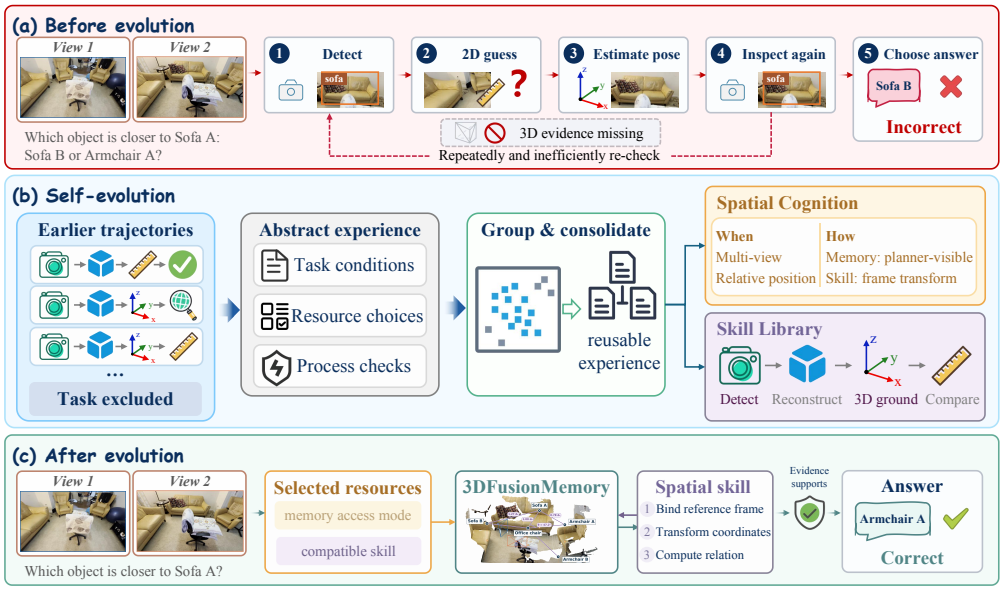


Figure 2: **Self-evolution from earlier trajectories.** (a) Reasoning with initial experience leaves the reference frame unresolved. (b) Completed earlier trajectories are abstracted and consolidated into resource-use experience and procedural records. (c) Accumulated experience guides memory access and skill selection for the same target. Experience becomes available between task blocks; scene evidence remains task-specific and model weights stay fixed. The trajectories and outcomes are illustrative.

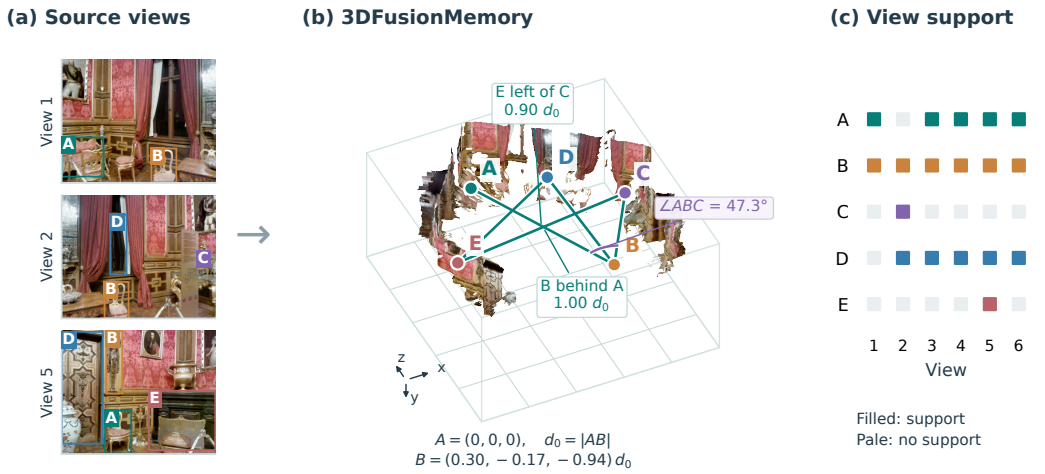


Figure 3: **Multi-view observations become spatial relations.** Three of six input views, a saved fused 3D scene, and view support for five estimated object nodes. Colors link identities. The RGB surface uses saved depth and camera poses; edges show unambiguous stored relations. Labels give stored directions and 3D measurements derived from object centers. Coordinates use A as origin and reconstruction-world axes; lengths are in units of $d_0 = |AB|$, not meters.

axes, and supporting views. Depth and camera poses place observations in the scene frame; matched observations update an object's center through confidence-weighted fusion (Appendix E). Repeated observations support the same node, while complementary views add objects that a single view does not cover. Relations are then computed between these fused object representations (Figure 3).

Queries return distances and camera- or object-frame relations together with confidence, supporting views, and ambiguity. Ordered route state supports successive translations and rotations. The planner may receive compact spatial facts, or memory may be reserved for answer checking. Skills and reasoning steps can therefore refer to the same grounded objects instead of reconstructing their relations independently. Each task starts with a fresh graph; cross-task experience concerns obtaining and using evidence, not retaining the previous scene's facts.

3.3 Spatial Skill Library

Recurring spatial operations can reuse a computation even when the objects and images change. The library represents each skill as a procedure with an explicit applicability condition and input–output interface:

$$\ell = (A_\ell, X_\ell, P_\ell, Y_\ell), \qquad \ell \in \mathcal{L}_b. \tag{3}$$

In equation 3, A_ℓ specifies applicability, X_ℓ typed input requirements, P_ℓ a parameterized tool procedure, and Y_ℓ its returned evidence. For example, a relative-direction procedure binds a reference frame before computing a displacement. Its parameters refer to the current task's objects, so the same procedure can operate on different scenes.

Seed skills initialize the library. Completed trajectories add recurring execution patterns and records of their use. Spatial Cognition can withhold a skill, expose it as a planning hint, or permit execution. The executable configuration also compiles compatible primitive calls into parameterized procedures and tests them under a limited budget; earlier hint-based configurations retain procedural records without this compilation. A change in an existing skill's availability is consequently distinct from the creation of new skill code.

A skill operates on current inputs and may build or query the current scene memory. It cannot supply an earlier answer or bypass a disabled memory interface. This keeps procedure reuse separate from both answer reuse and access to scene evidence.

3.4 Spatial Cognition

Deciding when and how to use resources. Spatial Cognition turns execution experience into conditional resource decisions. Its strategic store associates input-visible task conditions with past resource choices, process checks, failures, and costs. A task signature σ_i describes the reference frame, relation scope, output form, temporal structure, metric demand, and view count. It uses neither reference answers nor benchmark category labels. Given this signature and the available experience, the controller selects

$$g_i = \pi(\sigma_i; \mathcal{P}_b, \mathcal{L}_b) = (r_i, m_i, \mathcal{K}_i, \nu_i). \tag{4}$$

Here r_i chooses direct-first or tool-assisted reasoning, m_i controls memory access (off, answer-checking only, or planner-visible), $\mathcal{K}_i \subseteq \mathcal{L}_b$ gives permitted skills, and ν_i specifies verification and revision. Equation 4 is resource selection, not answer prediction. A skill's *how* describes a computation; Cognition's *how* describes how to use that computation and its evidence. The controller combines initial rules with stored experience rather than training an additional model.

Resource changes require supporting experience. In the executable configuration, route and memory alternatives must have sufficient historical support, acceptable process risk, and lower estimated token cost; otherwise the default remains. At most one learned route or memory factor changes per request. Skill selection requires compatible inputs and availability, while verification and resource limits follow fixed rules. Appendix E specifies these conditions. Process feedback is not a correctness reward: it describes execution reliability and cost, not agreement with a reference answer.

The self-evolution loop. Tasks first execute using the fixed block snapshot. After task i finishes, its trajectory τ_i is projected to a label-free record $d_i = \mathrm{Abs}(\tau_i)$: task signature, resource decisions, tool sequence, process checks, failures, and available costs. This abstraction excludes answers, literal scene coordinates, and raw task content from the transferable state. For completed block $\mathcal{B}_b$, let $\mathcal{D}_b = \{d_i : i \in \mathcal{B}_b\}$. Only then do the stores update:

$$\mathcal{P}_{b+1} = U_{\mathrm{policy}}(\mathcal{P}_b, \mathcal{D}_b), \qquad \mathcal{L}_{b+1} = U_{\mathrm{skill}}(\mathcal{L}_b, \mathcal{D}_b). \tag{5}$$

Table 1: **Spatial reasoning performance across five benchmarks.** Scores are percentages; higher is better. Our SpaceDG evaluations use a 1,000-question subset. Avg is the unweighted mean across benchmarks. Bold marks the best displayed score; dashes indicate unavailable results. Evaluation settings are given in Appendix A.

Method	MMSI	MindCube	OmniSp.	SpaceDG	SAT	**Avg**
Foundation VLMs						
Gemini-2.5-Pro	36.9	57.5	55.2	47.2	79.3	55.2
Qwen3-VL-8B Bai et al. (2025)	27.4	43.4	51.5	43.1	74.7	48.0
GPT-5.5	51.0	74.3	67.8	57.5	84.7	67.1
GPT-5.6-sol	53.8	67.2	68.3	56.9	**91.3**	67.5
Training-based Spatial VLMs						
SenseNova-SI-1.1 Cai et al. (2025)	43.3	85.7	42.2	36.3	50.7	51.6
SpatialLadder-3B Li et al. (2026b)	29.2	43.4	29.8	34.9	27.3	32.9
RoboBrain2.5-8B-NV Tan et al. (2026)	26.6	37.1	46.1	36.8	74.0	44.1
ACE-Brain-0.5-8B ACE-Brain Team et al. (2026)	35.5	86.3	41.4	39.8	82.7	57.1
Agentic Frameworks						
GCA Chen et al. (2025)	47.6	64.2	63.0	50.6	90.7	63.2
AlloSpatial (GPT-5.5) Ruan et al. (2026)	43.6	63.2	63.1	52.9	89.3	62.4
SpatialClaw (Gemma-4) Cho et al. (2026)	50.0	75.7	62.2	52.4	86.0	65.3
Skill-3D Li et al. (2026a)	48.8	69.4	66.6	51.4	89.3	65.1
SpatialBrain	**57.1**	**87.6**	**69.1**	**57.9**	**91.3**	**72.6**

Here U_{policy} consolidates condition-specific strategic records and statistics; U_{skill} updates procedural records and their availability. Equation 5 is a state update, not a gradient step. The next block can retrieve these records and make different resource choices. No task in $\mathcal{B}_b$ reads these updates, future observations, or reference answers. Tasks run in parallel within a block; experience becomes available only across the block boundary. Self-evolution therefore connects completed execution to future decisions through an explicit, inspectable state change.

4 Experiments

We evaluate the complete framework against direct reasoning and existing spatial models, then examine which task types account for the differences. Section 5 analyzes the components and experience reuse.

4.1 Experimental Setup

We evaluate MMSI, MindCube, OmniSpatial, SpaceDG, and SAT on 1,000, 1,050, 1,533, 1,000, and 150 items, respectively Yang et al. (2025); Wang et al. (2025); Jia et al. (2025); Zhou et al. (2026a); Ray et al. (2024). We report accuracy, using mean relative accuracy (MRA) for numeric SpaceDG questions. SpaceDG uses a fixed 1,000-item subset, and SAT uses fixed option order. The direct-answer baseline and SpatialBrain both use GPT-5.6-sol on the same items. Predictions are fixed before scoring; experience updates use neither reference answers nor future tasks.

Table 1 includes foundation VLMs, spatially trained models, and agentic frameworks. It combines published results with our evaluations, so backbones and protocols vary across methods. These comparisons position the framework among available approaches, while the shared-backbone pair measures its gain over direct answers. Appendix A gives evaluation settings and comparison details.

4.2 Main Results

SpatialBrain improves on direct GPT-5.6-sol answers on four benchmarks and ties on SAT (Table 1), raising the unweighted mean from 67.5 to 72.6. MindCube accounts for the largest gain, at 20.4

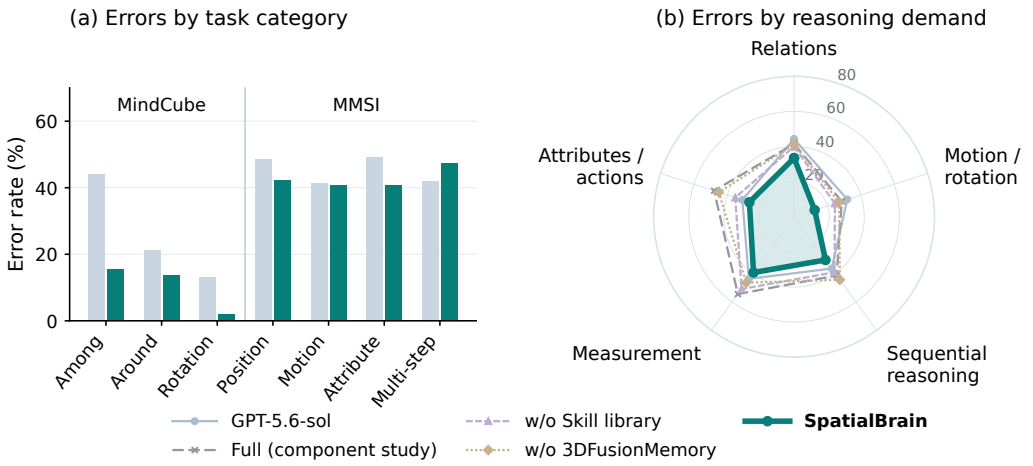


Figure 4: **Errors by task type and reasoning demand.** All 1,050 MindCube and 1,000 MMSI items. (a) Direct GPT-5.6-sol and SpatialBrain by official category. (b) Input-derived categories for the same pair and component configurations, all using GPT-5.6-sol. Radii show pooled error rates (0–80%). Appendix B.3 gives the mapping, counts, and within-study comparisons.

percentage points; the smaller changes on OmniSpatial and SpaceDG show that the benefit depends on the task setting.

These results measure the complete framework, including its tools and reasoning structure. Section 5 examines its components and the role of accumulated experience.

4.3 Error Analysis

Figure 4 breaks errors down by benchmark task type. On MindCube, errors decrease in all three categories, from 344 to 130 overall, a 62.2% reduction. The gain is therefore not confined to one category. On MMSI, SpatialBrain makes 33 fewer errors overall, but the direction of change varies: positional and attribute errors decrease, with no uniform reduction across all task types. The left panel groups MMSI into its four official families. The right panel pools the two benchmarks by input-derived reasoning demand, showing the largest reduction for motion and rotation. These groups describe what a question requires, not an inferred cause of failure.

The pooled radar summarizes shared demands but can hide benchmark-specific differences: sequential-reasoning errors decrease on MindCube and increase on MMSI. Appendix B.5 retains all eleven MMSI types, and Appendix D gives category-level scores for all five benchmarks.

4.4 Memory-grounded Spatial Reasoning

Figure 5 shows two views contributing anchors to a shared scene memory. The planner receives the atlas, and a stair-relative coordinate check resolves the food-storage location.

5 Analysis and Discussion

We examine the component comparisons, the scene evidence represented by memory, and the experience that becomes available to later tasks.

5.1 Component Analysis

We compare SpatialBrain with three component-removal configurations. **w/o Skill Library** removes seed and accumulated skills, retrieval, hints, macros, and induction while retaining primitive tools. **w/o 3DFusionMemory** disables scene-graph construction and access, including queries in-

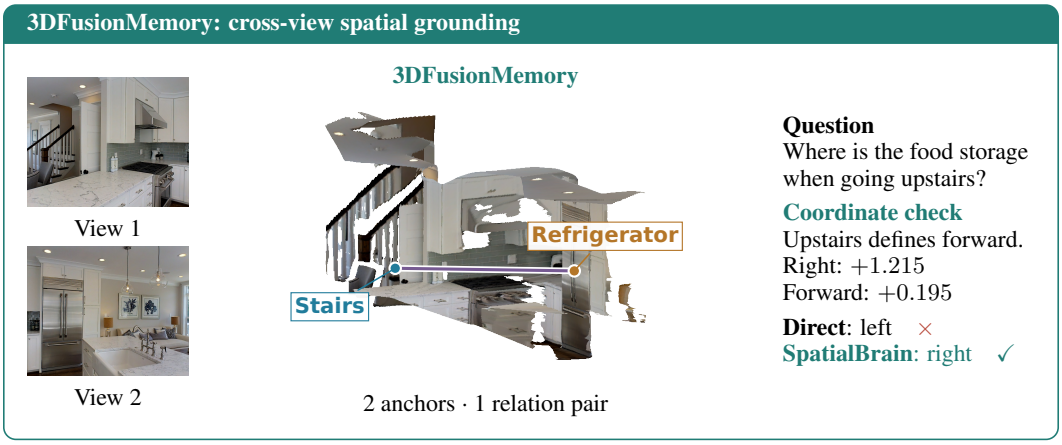


Figure 5: **Memory-grounded reference-frame reasoning.** Saved 3D anchors are shown on an RGB cutaway of the reconstructed scene. The planner sees the memory atlas with single-view support flags; a tool verifies the stair-relative direction. Coordinates use reconstruction units. Full trajectory: Appendix C.1.

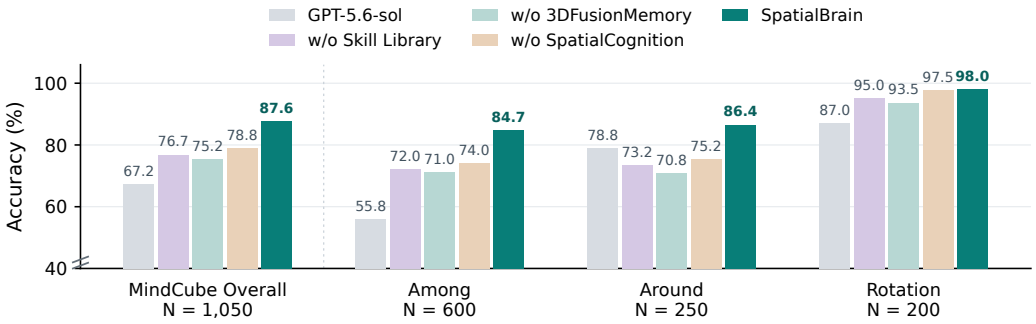


Figure 6: **MindCube component comparison.** Accuracy overall and by task category on 1,050 items. The y-axis starts at 40%. Evaluation settings are given in Appendix A.

side skills and memory-based answer checks. **w/o SpatialCognition** replaces task-specific analysis and adaptive routing with a generic controller.

Figure 6 summarizes the MindCube configurations on all 1,050 items. SpatialBrain scores 87.6, compared with 76.7 without the Skill Library, 75.2 without 3DFusionMemory, 78.8 without SpatialCognition, and 67.2 for direct answers. The gaps are larger on Among and Around than on Rotation. Missing and refused answers count as wrong throughout. Appendices A and B.3 report the settings and answer availability. Table A4 provides the category scores. Appendix B.4 reports the earlier MMSI component study (Figure A2).

5.2 What Does 3DFusionMemory Represent?

3DFusionMemory represents spatial relations supported by complementary observations. In Figure 3, one view supports nodes A and B, another adds C and D, and a third includes E. Estimated identities connect these observations in a common scene frame; three of the five 3D objects have support from multiple views. The resulting memory gives later computations a shared set of objects and relations to reference. Across five GPT-5.6-sol evaluation streams, saved 3D objects occur in 50.6% of MMSI tasks and 37.5% of SpaceDG tasks, compared with 21.0%, 29.2%, and 18.0% for MindCube, OmniSpatial, and SAT (Appendix B.1). This variation reflects which inputs produce spatial state. Stored coverage does not establish that the reasoner queried the state or that its geometry was correct; query traces and matched removals address those separate questions.

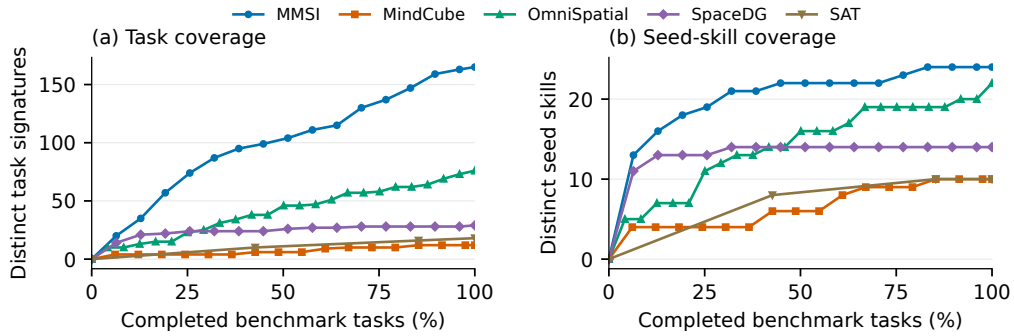


Figure 7: **Test-time experience and seed-skill coverage.** (a) Distinct task signatures with stored experience. (b) Distinct seed skills referenced by that experience. Each curve spans one benchmark's evaluation set: MMSI/MindCube/OmniSpatial/SpaceDG/SAT contain 1,000/1,050/1,533/1,000/150 tasks, respectively.

5.3 Experience and Skill Coverage

Figure 7 tracks the breadth of stored experience across five independent GPT-5.6-sol streams. Task coverage counts distinct task signatures with experience records; skill coverage counts existing seed skill IDs referenced by those records, not newly generated programs. MMSI reaches 165 task signatures and 24 seed skills, while MindCube reaches 12 and 10. On MMSI, task coverage continues to grow after skill coverage begins to plateau: further experience extends the recorded task conditions without a comparable increase in distinct referenced skills. The curves show a growing range of recorded conditions and resources, rather than their execution success or causal effect on accuracy.

Appendix B.2 separately measures how often earlier experience reaches subsequent tasks in the executable configuration. Individual trajectories show how experience changes a route or skill recommendation (Appendix C.3). An earlier GPT-5.5 hint-retrieval experiment also logged partial sequence alignment, but did not improve aggregate accuracy or cost (Appendix B.6). The relevant contrast for update benefits is an initial-frozen control with the same resources and starting experience.

Limitations. Published baselines differ in protocols and serving precision, and our SAT evaluation fixes option order. Removing 3DFusionMemory also removes memory-based answer checks, so it does not isolate storage alone. The experience curves and selected trajectories establish accumulation and use, not a general improvement from self-evolution. Perception and geometry errors also remain imperfectly calibrated. Matched evaluations should measure both answer quality and the full cost of acquiring and using experience.

6 Conclusion

SpatialBrain separates evidence about the current scene from reusable spatial computation and experience about when and how to use it. 3DFusionMemory, the Spatial Skill Library, and Spatial Cognition implement these roles around a fixed vision-language backbone. Completed trajectories update external experience between task blocks, making self-evolution an inspectable change in the resources and decisions available to later tasks. The complete framework improves on direct answers on four of five benchmarks and ties on the fifth. Coverage curves and execution traces show which task conditions acquire experience and how that experience enters later resource decisions.

AI use statement

Generative AI tools assisted development of the conceptual framework, experimental design, method implementation and debugging, data-processing and evaluation scripts, interpretation of results, and qualitative analysis of execution traces. Additional assistance covered literature search,

mathematical notation, figure preparation, manuscript drafting and revision, and LaTeX formatting. Checks included automated regression tests, re-tallying selected scores from prediction records, and inspection of rendered figures and text. This research assistance is distinct from the vision-language models used within the evaluated systems, which are described in Sections 3 and 4. The authors are responsible for the final text, claims, results, and artifacts, including those developed with AI assistance.

Ethics statement

This work studies spatial reasoning on existing image and video benchmarks. Such media can contain people and private interiors, and derived spatial representations can expose scene layouts. Reuse of benchmark media, reconstructions, and execution traces should respect the original data terms and applicable privacy restrictions. Perception errors and incorrect spatial relations can also propagate through memory and reused procedures. Our benchmark evaluations do not establish safety for autonomous physical actions; deployment in robotics or other safety-critical settings would require independent validation, appropriate safeguards, and human oversight.

Reproducibility statement

Section 3 specifies the framework and its state updates; Appendix E details scene fusion and strategy admission. Section 4.1 and Appendix A describe evaluation scopes, metrics, subset selection, answer handling, and configuration differences. Appendix D provides category-level results, while Appendices B and C describe the supporting analyses and selected execution traces. These references distinguish main-table evaluations, component comparisons, and experience analyses rather than treating them as one configuration. Exact reproduction of hosted-model outputs also depends on service versions and inference variability.

References

ACE-Brain Team, Ziyang Gong, Haoming Gu, Zehang Luo, Tianyi Zhang, Tao Tao, Yixiao Chi, Zhe Liu, Lingsi Zhu, Jingyuan Liu, Anke Tang, Songze Li, Yilun Kong, Ningjing Liu, Tianyu Zhu, Yunpeng Qing, Shuang Luo, Xiang Liu, Shi Fu, Dawei Nie, Sixiang Liu, Zhexi Wen, Feng Pan, Xiaofeng Wang, Zhi Hou, Chunxiao Liu, Xue Yang, Junchi Yan, Hengshuang Zhao, Dacheng Tao, and Xiaogang Wang. ACE-Brain-0.5: A unified embodied foundational model for physical agentic AI. arXiv preprint arXiv:2607.04426, 2026. URL https://arxiv.org/abs/2607.04426.

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, 2025. URL https://arxiv.org/abs/2511.21631.

Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Qingping Sun, Tongxi Zhou, Jiaqi Li, Hui En Pang, Oscar Qian, Yukun Wei, Zhiqian Lin, Xuanke Shi, Kewang Deng, Xiaoyang Han, Zukai Chen, Xiangyu Fan, Hanming Deng, Lewei Lu, Liang Pan, Bo Li, Ziwei Liu, Quan Wang, Dahua Lin, and Lei Yang. Scaling spatial intelligence with multimodal foundation models. arXiv preprint arXiv:2511.13719v4, 2025. URL https://arxiv.org/abs/2511.13719v4. Version 4, March 2026; SenseNova-SI v1.1.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial rea-

soning capabilities. arXiv preprint arXiv:2401.12168, 2024. URL https://arxiv.org/abs/2401.12168.

Zeren Chen, Xiaoya Lu, Zhijie Zheng, Pengrui Li, Lehan He, Yijin Zhou, Jing Shao, Bohan Zhuang, and Lu Sheng. Geometrically-constrained agent for spatial reasoning. arXiv preprint arXiv:2511.22659, 2025. URL https://arxiv.org/abs/2511.22659.

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. arXiv preprint arXiv:2406.01584, 2024. URL https://arxiv.org/abs/2406.01584.

Seokju Cho, Ryo Hachiuma, Abhishek Badki, Hang Su, Byung-Kwan Lee, Chan Hee Song, Sifei Liu, Subhashree Radhakrishnan, Seungryong Kim, Yu-Chiang Frank Wang, and Min-Hung Chen. SpatialClaw: Rethinking action interface for agentic spatial reasoning. arXiv preprint arXiv:2606.13673, 2026. URL https://arxiv.org/abs/2606.13673.

Yalun Dai, Hao Li, Shulin Tian, Runmao Yao, Yuhao Dong, Fangzhou Hong, Zhaoxi Chen, Fangfu Liu, Tao Wang, Kim-Hui Yap, and Ziwei Liu. S-Agent: Spatial tool-use elicits reasoning for spatial intelligence. arXiv preprint arXiv:2606.20515, 2026. URL https://arxiv.org/abs/2606.20515.

Qiao Gu, Alihusein 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. arXiv preprint arXiv:2309.16650, 2023. URL https://arxiv.org/abs/2309.16650.

Shengran Hu, Cong Lu, and Jeff Clune. Automated design of agentic systems. arXiv preprint arXiv:2408.08435, 2024a. URL https://arxiv.org/abs/2408.08435.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. arXiv preprint arXiv:2406.09403, 2024b. URL https://arxiv.org/abs/2406.09403.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. OmniSpatial: Towards comprehensive spatial reasoning benchmark for vision language models. arXiv preprint arXiv:2506.03135, 2025. URL https://arxiv.org/abs/2506.03135.

Danial Kamali, Tanawan Premsri, Shreya Rajpal, Amir Zadeh, Chuan Li, and Parisa Kordjamshidi. SATURN: Symbolic spatial reasoning for multi-perspective grounding. arXiv preprint arXiv:2606.22694, 2026. URL https://arxiv.org/abs/2606.22694.

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, 2026a. 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*, 2026b. URL https://openreview.net/forum?id=KtrFXlvgrK.

Xudong Li, Mengdan Zhang, Peixian Chen, Jiaxi Tan, Zihao Huang, Jingyuan Zheng, Yan Zhang, Xiawu Zheng, Xing Sun, and Rongrong Ji. OmniView-Space: Reinforcing spatial reasoning via multi-perspective spatial mapping. arXiv preprint arXiv:2607.00881, 2026c. URL https://arxiv.org/abs/2607.00881.

Damiano Marsili, Rohun Agrawal, Yisong Yue, and Georgia Gkioxari. Visual agentic AI for spatial reasoning with a dynamic API. arXiv preprint arXiv:2502.06787, 2025. URL https://arxiv.org/abs/2502.06787.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. arXiv preprint arXiv:2304.03442, 2023. URL https://arxiv.org/abs/2304.03442.

Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, and Kate Saenko. SAT: Dynamic spatial aptitude training for multimodal language models. arXiv preprint arXiv:2412.07755, 2024. URL https://arxiv.org/abs/2412.07755.

Shouwei Ruan, Bin Wang, Zhenyu Wu, Qihui Zhu, Yuxiang Zhang, Jingzhi Li, Yubin Wang, and Xingxing Wei. AlloSpatial: Agentic harness framework for spatial reasoning in foundation models. arXiv preprint arXiv:2606.08952, 2026. URL https://arxiv.org/abs/2606.08952.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. arXiv preprint arXiv:2303.11366, 2023. URL https://arxiv.org/abs/2303.11366.

Huajie Tan, Enshen Zhou, Zhiyu Li, Yijie Xu, Yuheng Ji, Xiansheng Chen, Cheng Chi, Pengwei Wang, Huizhu Jia, Yulong Ao, Mingyu Cao, Sixiang Chen, Zhe Li, Mengzhen Liu, Zixiao Wang, Shanyu Rong, Yaoxu Lyu, Zhongxia Zhao, Peterson Co, Yibo Li, Yi Han, Shaoxuan Xie, Guocai Yao, Songjing Wang, Leiduo Zhang, Xi Yang, Yance Jiao, Donghai Shi, Kunchang Xie, Shaokai Nie, Chunlei Men, Yonghua Lin, Zhongyuan Wang, Tiejun Huang, and Shanghang Zhang. RoboBrain 2.5: Depth in sight, time in mind. arXiv preprint arXiv:2601.14352, 2026. URL https://arxiv.org/abs/2601.14352.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. arXiv preprint arXiv:2305.16291, 2023. URL https://arxiv.org/abs/2305.16291.

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

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. arXiv preprint arXiv:2409.07429, 2024. URL https://arxiv.org/abs/2409.07429.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-MEM: Agentic memory for LLM agents. arXiv preprint arXiv:2502.12110, 2025. URL https://arxiv.org/abs/2502.12110.

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. arXiv preprint arXiv:2412.14171, 2024a. URL https://arxiv.org/abs/2412.14171.

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. arXiv preprint arXiv:2505.23764, 2025. URL https://arxiv.org/abs/2505.23764.

Yuncong Yang, Han Yang, Jiachen Zhou, Peihao Chen, Hongxin Zhang, Yilun Du, and Chuang Gan. 3D-Mem: 3d scene memory for embodied exploration and reasoning. arXiv preprint arXiv:2411.17735, 2024b. URL https://arxiv.org/abs/2411.17735.

Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune. Darwin Gödel machine: Open-ended evolution of self-improving agents. arXiv preprint arXiv:2505.22954, 2025a. URL https://arxiv.org/abs/2505.22954.

Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, Urmish Thakker, James Zou, and Kunle Olukotun. Agentic context engineering: Evolving contexts for self-improving language models. arXiv preprint arXiv:2510.04618, 2025b. URL https://arxiv.org/abs/2510.04618.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. ExpeL: LLM agents are experiential learners. arXiv preprint arXiv:2308.10144, 2023. URL https://arxiv.org/abs/2308.10144.

Xiaolong Zhou, Yifei Liu, Ziyang Gong, Jiarui Li, Qiyue Zhao, Muyao Niu, Yuanyuan Gao, Le Ma, Xue Yang, Hongjie Zhang, and Zhihang Zhong. SpaceDG: Benchmarking spatial intelligence under visual degradation. arXiv preprint arXiv:2605.22536, 2026a. URL https://arxiv.org/abs/2605.22536.

Yang Zhou, Zixuan Huang, Sunzhu Li, Zhuo Yang, Chen Zhang, Shunian Chen, Caijun Yan, Jianyao Xu, Shunyu Liu, Weijie Fu, Peiliang Li, Xiaozhi Chen, and Yuxiang Cai. SpatialCLI: Learning to reason with spatial tools, then without them. arXiv preprint arXiv:2607.27703, 2026b. URL https://arxiv.org/abs/2607.27703.

Chenming Zhu, Jingli Lin, Yilin Long, Peizhou Cao, Tai Wang, Jiangmiao Pang, and Xihui Liu. Thinking with imagination: Agentic visual spatial reasoning with world simulators. arXiv preprint arXiv:2606.06476, 2026. URL https://arxiv.org/abs/2606.06476.

APPENDIX CONTENTS

Page

A Evaluation Settings 15

B Supplementary Analyses 16

B.1 3DFusionMemory Coverage 16

B.2 Experience Coverage and Subsequent Exposure 16

B.3 Input-Derived Reasoning Categories 17

B.4 MMSI Component Comparison 18

B.5 MMSI Errors by Task Type 19

B.6 Earlier Retrieval and Sequence Alignment 19

C Qualitative Analysis and Visualizations 20

C.1 3DFusionMemory 21

C.2 Spatial Skill Library 30

C.3 SpatialCognition 32

D Detailed Benchmark Results 42

D.1 MindCube 42

D.2 MMSI 43

D.3 OmniSpatial 44

D.4 SAT 44

D.5 SpaceDG 45

E Experience Update Details 46

A Evaluation Settings

Table A1 summarizes the settings used in our evaluations. The SpaceDG subset is sampled with seed 0 from the 9,918-item release.

Table A1: **Evaluation settings.** The final column specifies which comparisons share each setting. Published baselines follow their original protocols.

Factor	Setting	Applies to
Evaluation size	MMSI 1,000; MindCube 1,050; OmniSpatial 1,533; SpaceDG 1,000; SAT 150.	Our main-table evaluations
Metric	Accuracy (%); MRA for numeric SpaceDG questions.	Our scored comparisons
Option order	Fixed question-option order, without circular evaluation.	Our SAT evaluations
Backbone	GPT-5.6-sol.	Direct / SpatialBrain pair
Reference answers	Excluded from agent inputs and experience updates; predictions fixed before scoring.	Our agent evaluations
Unanswered items	Retained in the denominator and counted as wrong, including refusals.	Component comparisons

Backbones, decoding settings, and tool access differ across methods in Table 1. SenseNova-SI-1.1 denotes the InternVL3-8B variant. AlloSpatial uses the public World2Mind tool and reasoning harness with GPT-5.5. Qwen3-VL-8B uses a collaborator-provided score summary; raw predictions were not available for independent verification. The World2Mind OmniSpatial result includes three failed tool observations.

MindCube component comparisons. MindCube keeps all 600 Among, 250 Around, and 200 Rotation items. The main table, category breakdown, and error analysis use the same 920/1,050 SpatialBrain run as the Full result in Table A4. The five-stream memory and experience coverage analyses use a separate test-time batch. Full/no-Skill/no-Memory/no-Cognition in Table A4 have 0/142/135/85 blank answers, all counted wrong. The matched Full and no-Memory comparison scores 75.9 and 75.2, respectively. Their difference is +0.67 points overall and +2.83/-5.20/+1.50 by category; the larger gap in Table A4 cannot be attributed to memory removal alone. On 908 jointly answered Full/no-Skill items from Table A4, correct counts are 799 and 805. Answer availability therefore affects the overall gaps alongside other configuration changes. The reported scores still include every item.

MMSI component comparisons. Figure A2 includes all four MMSI families (Position/Motion/Attribute/multi-step: 522/150/130/198 items). Full/no-Skill/no-Memory/no-Cognition blank counts are 68/99/116/101. Full-minus-removal gains of 18/32/35 correct answers comprise $-6/+12/+24$ on jointly answered pairs and $+24/+20/+11$ on pairs with a blank. These earlier runs were not verified as complete component removals, so their differences cannot establish isolated component effects. Grouping results by task family preserves the original experience stream and retains zero and negative differences.

B SUPPLEMENTARY ANALYSES

The Direct/SpatialBrain error comparisons use the reported main-table predictions. The five-stream memory and experience coverage analyses use the earlier GPT-5.6-sol test-time batch (Appendix A). The multi-view illustration in Figure 3 comes from a separate saved MMSI scene in the strict-component run. The experience-exposure counts in Appendix B.2 use the later executable configuration on full MMSI and MindCube. These analyses answer different questions; their counts are not pooled across versions.

B.1 3DFUSIONMEMORY COVERAGE

Figure A1 counts tasks with nonempty saved 3D objects and spatial relations. Object and relation coverage overlap, so the left panel plots them separately. The right panel partitions each evaluation set into snapshots with 3D objects, snapshots without 3D objects, and missing snapshots. The last category is unknown rather than an observed empty memory. Every percentage uses the full evaluation denominator, including 1,000 sampled SpaceDG items.

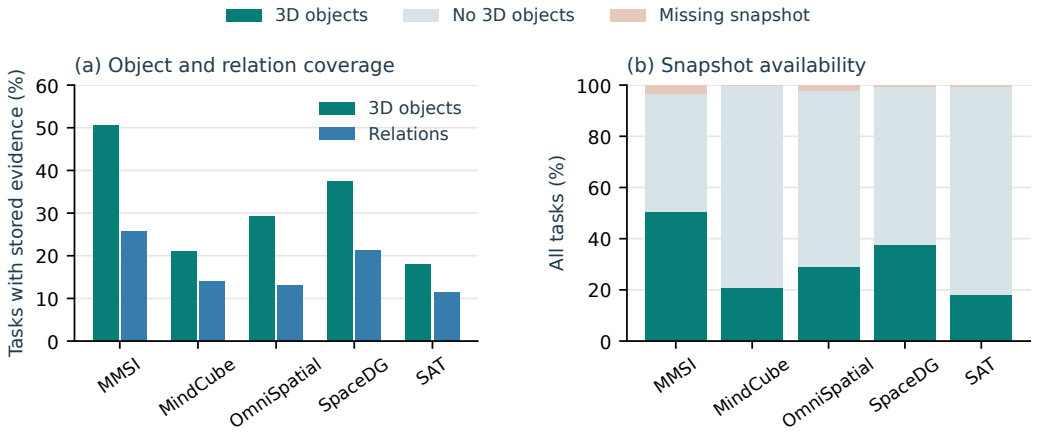


Figure A1: **Stored spatial evidence across benchmarks.** (a) Tasks with 3D objects or relations. (b) Snapshot availability and 3D-object presence. The two panels distinguish overlapping evidence types from a complete, non-overlapping partition of all tasks. Coverage measures stored state, not retrieval or answer correctness.

There are 506/220/448/375/27 tasks with 3D objects and 257/146/201/213/17 with relations on MMSI/MindCube/OmniSpatial/SpaceDG/SAT. Missing snapshots number 33/0/30/7/1. Multiple camera states occur in 710/496/0/574/0 tasks, respectively. MMSI and SpaceDG show higher object and relation coverage than the other streams, but coverage does not follow a universal 3D-versus-2D split. The example in Figure 3 illustrates complementary view support: the displayed views support two, three, and four of its five 3D objects, and their union covers all five. This describes observed support, not object recall against ground truth.

B.2 EXPERIENCE COVERAGE AND SUBSEQUENT EXPOSURE

Figure 7 tracks five independent GPT-5.6-sol streams. Task coverage counts distinct task signatures with stored experience. Skill coverage counts existing seed skill IDs referenced in those records, not newly generated programs. Both describe the range of experience available as evaluation proceeds. Normalized progress allows different-length streams to be compared; coverage is not successful execution or accuracy.

The later executable configuration is evaluated separately on complete MMSI and MindCube streams. MMSI forms 18 experience records, six of which become available; MindCube forms five, with three available. These are condition-specific resource-use records, not new algorithms. Nonempty experience from an earlier committed state reaches 231 of 1,000 MMSI tasks and 212

of 1,050 MindCube tasks. Matching the injected record IDs to each target's input state verifies that their supporting events precede the target block. These exposure counts establish delivery to the reasoner, not successful execution or an improvement in answers.

B.3 Input-Derived Reasoning Categories

The radar in Figure 4 uses input-derived task signatures recorded by the label-blind router, with the same label assigned to every method's response for each item. We group `relation` as Relations; `motion` and `rotation` as Motion/rotation; `route_sequence` as Sequential reasoning; `visual_measurement`, `count`, `diagram_geometry`, and `depth` as Measurement; and `attribute`, `appearance`, `affordance`, and `spatial_physics` as Other attributes/actions. This mapping was fixed before aggregating outcomes and retains every item. It describes question demands, not a diagnosis inferred from a wrong answer. Direct responses often contain no reasoning trace sufficient to establish such a diagnosis. Table A2 reports the category counts and pooled error rates used in the radar.

Table A2: **Errors within input-derived categories.** Top: Direct/SpatialBrain wrong-answer counts, with category size in parentheses. Bottom: pooled error rates (%) for all radar curves. Each method uses the same 2,050 items and fixed category assignment.

Category	MindCube	MMSI	Pooled
Relations	178/98 (459)	279/247 (574)	457/345 (1,033)
Motion/rotation	117/7 (405)	64/63 (163)	181/70 (568)
Sequential reasoning	46/25 (153)	51/56 (113)	97/81 (266)
Measurement	—	49/44 (112)	49/44 (112)
Other attributes/actions	3/0 (33)	19/19 (38)	22/19 (71)

Method	Relations	Motion	Sequential	Measure.	Other
GPT-5.6-sol	44.2	31.9	36.5	43.8	31.0
Full (component study)	41.5	28.5	41.4	54.5	47.9
w/o Skill library	40.0	24.5	39.1	50.9	35.2
w/o 3DFusionMemory	41.2	26.9	44.4	46.4	45.1
SpatialBrain	33.4	12.3	30.5	39.3	26.8

The pooled error rate is lower for all five groups, while the contribution of each benchmark differs. Motion/rotation improves mainly on MindCube; sequential reasoning regresses on MMSI, and its Other group is unchanged. Measurement has no MindCube items under this signature taxonomy. Thus, the pooled polygon is not evidence of uniform gains on both benchmarks.

The component curves come from the later executable implementation and include its own Full reference. On MindCube/MMSI, this Full makes 253/543 errors, compared with 238/500 without the Skill library and 260/521 without 3DFusionMemory. The gap to main-table SpatialBrain also includes implementation changes and cannot isolate a component's contribution. MindCube's component runs passed their runtime checks; the MMSI curves are independently scored complete predictions whose runtime audits remain unverified. Blank answers count as errors throughout.

B.4 MMSI Component Comparison

This earlier GPT-5.6-sol study compares Full with three removal configurations on the same 1,000 MMSI items (Figure A2). It is separate from the main-table comparison. Full makes 487 errors, versus 505/519/522 for Skill/Memory/Cognition removal. Errors differ most on positional questions; multi-step errors are unchanged by memory removal. The configurations were not verified as complete removals, and unanswered items account for part of the gaps (Appendix A).

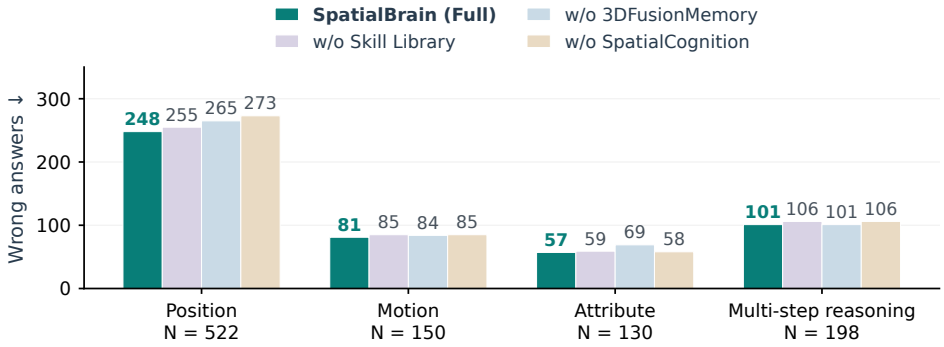


Figure A2: **MMSI component errors by task family.** Earlier component configurations on all 1,000 items, separate from the main-table runs. Bars count wrong answers, including unanswered questions. The four families partition the evaluation set; configuration and answer-availability differences are detailed in Appendix A.

B.5 MMSI Errors by Task Type

Figure A3 compares Direct and SpatialBrain on all 1,000 MMSI items and eleven task types. Direct makes 462 errors and SpatialBrain makes 429, with 199 corrections and 166 reversals. Camera–object errors decrease from 53 to 24 and appearance errors from 39 to 25, whereas multi-step errors increase from 83 to 94. The plot retains every type, including regressions and the unchanged region–region category. These categories identify where errors occur, not their underlying causes.

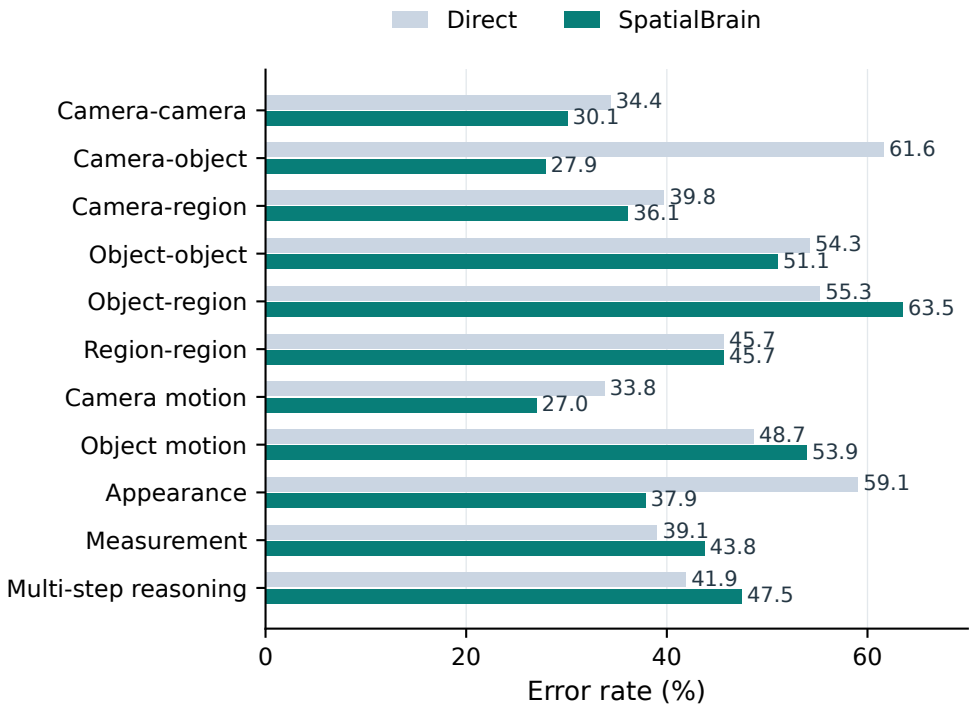


Figure A3: **MMSI errors by task type.** Direct GPT-5.6-sol answers and SpatialBrain on all 1,000 items. Each rate divides wrong answers by the number of items in that type. Lower is better.

B.6 Earlier Retrieval and Sequence Alignment

In an earlier GPT-5.5 experiment, 500 MMSI source tasks produce 81 procedural records across 36 task–path families and 16 capability skeletons. Among 500 disjoint target tasks, 332 are eligible for retrieval: 122 retrieve a record (36.7%) and 59 execute a partially matching tool sequence (17.8%). Partial alignment requires ordered overlap covering at least half of a retrieved sequence; it occurs in 48.4% of tasks with retrieved records. This analysis does not count newly generated skills or measure answer improvements. The experiment did not improve aggregate accuracy or cost. It is separate from the full executable-cohort exposure counts in Appendix B.2.

C Qualitative Analysis and Visualizations

We group the examples by their role in spatial reasoning: **3DFusionMemory** for grounding, the **Skill Library** for reusable procedures, and **SpatialCognition** for experience-guided resource choices. Teal, purple, and amber boxes identify these three groups; blue-gray boxes show additional tool-based trajectories.

Component / box color	Examples (MMSI IDs)
3DFusionMemory / teal	Atlas exposure and spatial verification: 8, 9, 206.
Skill Library / purple	Successful skill returns and answer correction: 32, 42.
SpatialCognition / amber	Experience-guided routing: 748; adaptive skill selection: 348.

Case selection. The 15 trajectories come from the 1,000-item MMSI evaluations with GPT-5.6-sol. Ten show an incorrect Direct answer and a correct SpatialBrain answer. We selected these examples after scoring to explain successful reasoning paths; they are not representative of the full set, which has 199 corrections and 166 reverse outcomes. Overall scores are 53.8% for Direct, 57.4% for the fixed-policy control, and 57.1% for SpatialBrain.

The boxes identify observed operations, not isolated component effects. Making memory available does not prove that memory alone caused a correction, and executing an existing skill is different from learning a new one. The six additional memory cases illustrate saved spatial representations. For the SpatialCognition cases, we checked that the cited experience precedes the target task and that the later resource choice follows that experience.

Compute comparisons. Unless stated otherwise, compute comparisons use the fixed-policy control. Of 485 both-correct fixed-policy/SpatialBrain pairs, 480 have complete token records, 221 use fewer SpatialBrain tokens, and 71 also use fewer non-final top-level calls and less observed time. Of 372 both-correct Direct/SpatialBrain pairs, 371 have complete tokens and only MMSI 748 uses fewer SpatialBrain tokens; both paths use zero tools. Call counts include failed attempts and exclude final-answer generation; primitive calls inside a skill are counted separately from top-level calls. Elapsed times are observational: concurrency and caches were not controlled. The evaluations used the same task IDs, model, and protocol, but exact image-byte identity could not be checked because per-run hashes were not recorded.

C.1 3DFusionMemory

3DFusionMemory: Reference-frame Verification MMSI 8

3DFusionMemory · Answer Correction

Question

Assuming the picture display area is on the south wall, where is the corridor passage area located in this bedroom?

A. Northeast corner
B. Southeast corner
C. Southwest corner
D. Northwest corner



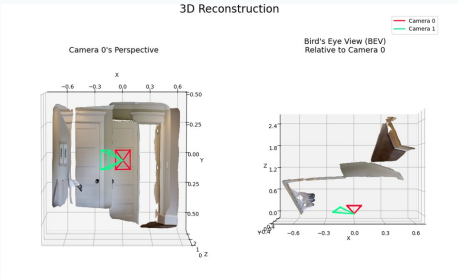
Image 1



Image 2

Direct A. Northeast corner × Incorrect

SpatialBrain



Reconstructed bedroom and camera poses

1. Locate the corridor and the south-wall pictures; reconstruct the room and project both objects into 3D.
2. 3DFusionMemory stores two anchors and one relation. Its atlas is attached to the planner, flagging single-view support and an ambiguous relation.
3. Use the picture wall as the compass reference. After a failed Python call, the retry gives north = +0.087, east = -1.897. Return D: northwest.

Answer D. Northwest corner ✓ Correct

3DFusionMemory: evidence with uncertainty
The planner receives the two-anchor memory atlas and its risk flags before the recorded coordinate check; the memory critic does not override the answer.

The north component is weak. This is observed memory exposure plus tool-based verification, not proof that memory alone corrected the answer.

3DFusionMemory: Cross-view Spatial Grounding — MMSI 9

3DFusionMemory · Answer Correction

Question

When you are going upstairs in the stairwell area, where is the food storage area located relative to you?

A. In front B. Behind C. On the left D. On the right



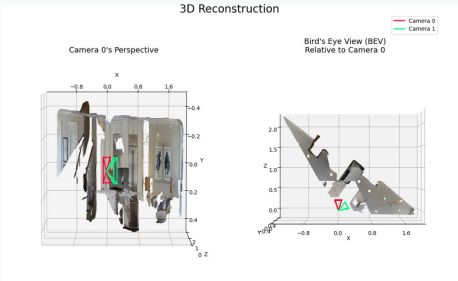
Image 1



Image 2

Direct C. On the left × Incorrect

SpatialBrain



Reconstruction and camera poses

Memory available to the planner
2 anchors; 1 relation pair.
Atlas includes relation uncertainty.

1. Reconstruct both views; detect and project the stairs and refrigerator into the shared frame.
2. The saved 3DFusionMemory atlas contains two 3D anchors and one relation and is exposed to the planner.
3. The coordinate check uses the upstairs direction as forward: right = +1.215, forward = +0.195. Return D: on the right.

Answer D. On the right ✓ Correct

The stair-relative check reports a clear directional margin. Atlas exposure is recorded; the numerical answer is verified by a tool, not by a standalone MemoryQuery call.

3DFusionMemory: Landmark-based Grounding — MMSI 206

3DFusionMemory · Answer Correction

Question

The stove is on the east side of the kitchen. In which direction is the mural in the kitchen?

A. Northeast B. Southeast C. Southwest D. Northwest



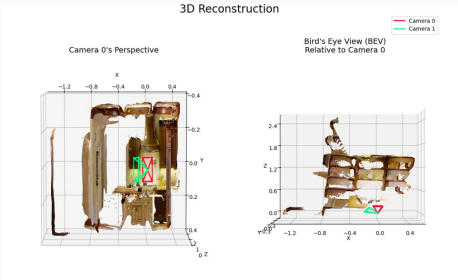
Image 1



Image 2

Direct C. Southwest × Incorrect

SpatialBrain



Reconstruction and camera poses

Memory available to the planner
4 anchors; 6 relation pairs.
Atlas includes relation uncertainty.

1. Project the stove, mural, and two island observations into the reconstructed kitchen. Four anchors and six relation pairs are retained in 3DFusionMemory.
2. Expose the atlas to the planner. The coordinate check combines the two island observations after a consistency check and uses the stove to define east.
3. The mural displacement has east = -0.360 and north = +0.432. Return D: northwest.

Answer D. Northwest ✓ Correct

The memory retains four anchors. The coordinate check combines the two island observations, but the memory itself keeps them separate and does not override the answer.

Additional 3DFusionMemory Examples

These six examples connect image detections to 3D anchors, camera poses, and supporting views. We selected them for visible geometry, without using answer correctness. The overhead inset shows the same state from a second viewpoint, not a later fusion step. Colored matrix cells indicate recorded view support, not independently verified object matches. All coordinates are unchanged; distance scales are not calibrated in meters. Dashed lines mark uncertain relations, and each diagram shows a subset of the stored relations.

3DFusionMemory: Spatial Representation Memory 01

Representation Example · MMSI 297

Objects supported across views

6 camera views · 5 3D anchors · 10 stored relation pairs



The sofa, statue, and door retain support from multiple views. The clock and fireplace have single-view support. Shared letter IDs connect detections, saved 3D anchors, and the support matrix.

3/5 anchors have support in at least two views.

3DFusionMemory: Spatial Representation Memory 02

Representation Example · MMSI 861

Mixed support in an indoor scene

6 camera views · 6 3D anchors · 15 stored relation pairs

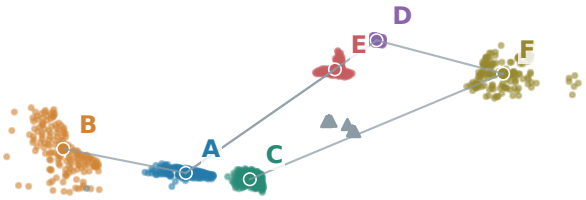
Source view 0



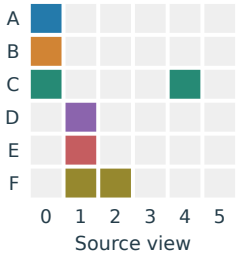
Source view 1



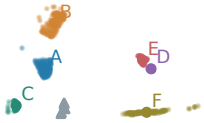
Saved 3DFusionMemory



View support



Overhead



A Bed
B Lounge chair
C Top
D Switch
E Sink
F Stairs

Triangles: cameras | Dashed: uncertain relation | Same geometry in both views

The top and stairs retain two-view support; the other four anchors are supported by one view each. The additional overhead view separates the stored layout from the display camera perspective.

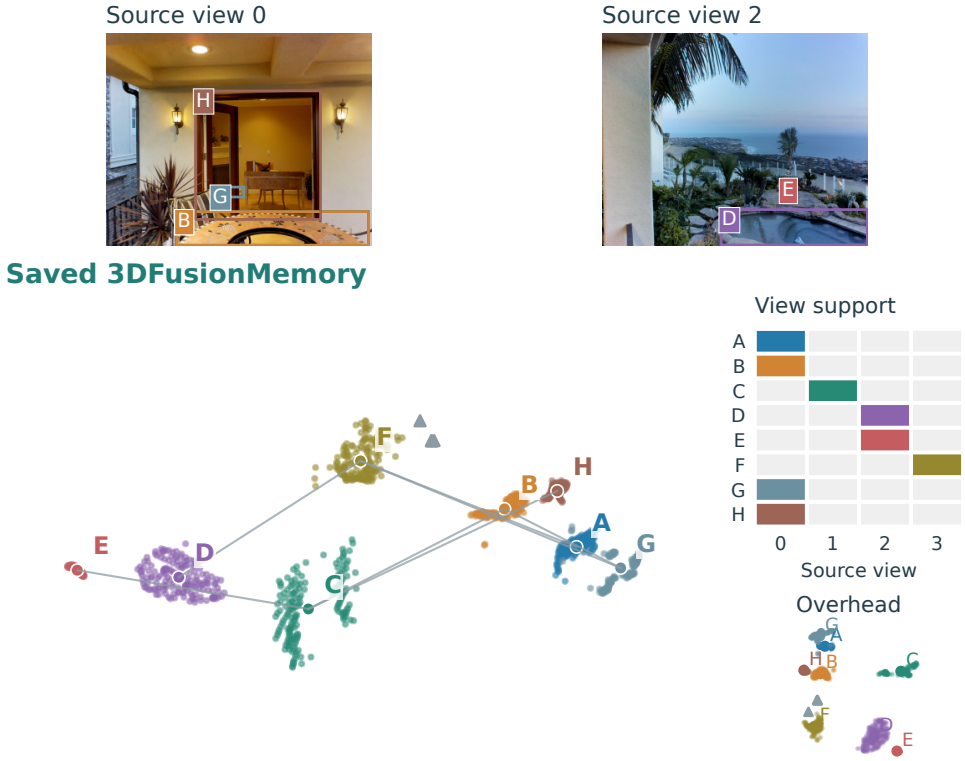
2/6 anchors have support in at least two views.

3DFusionMemory: Spatial Representation Memory 03

Representation Example · MMSI 700

A shared frame for indoor and outdoor anchors

4 camera views · 8 3D anchors · 28 stored relation pairs



Eight anchors from four views coexist in the saved coordinate frame. Each anchor has single-view support: this example shows cross-view scene organization, not repeated observation of the same object.

0/8 anchors have support in at least two views.

3DFusionMemory: Spatial Representation Memory 04

Representation Example · MMSI 711

Object and relation state from two views

2 camera views · 7 3D anchors · 21 stored relation pairs

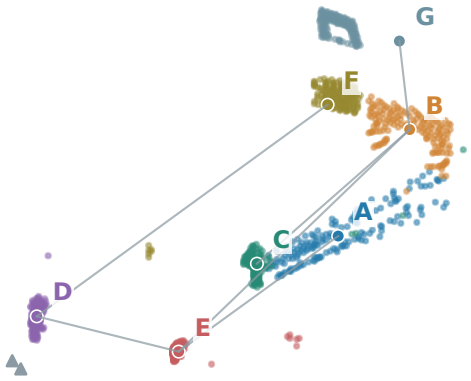
Source view 0



Source view 1



Saved 3DFusionMemory



View support



Overhead



A Sofa
B Fireplace
C Table lamp
D Dining chair
E Countertop plant
F Television
G Window

Triangles: cameras | Dashed: uncertain relation | Same geometry in both views

The two source views provide the visual context for the stored anchors and relations. Object colors are kept identical in the detections, oblique view, overhead view, and support matrix.

0/7 anchors have support in at least two views.

3DFusionMemory: Spatial Representation Memory 05

Representation Example · MindCube around / 033d379d

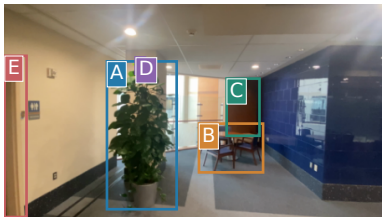
Scene context retained around an anchor

3 camera views · 5 3D anchors · 10 stored relation pairs

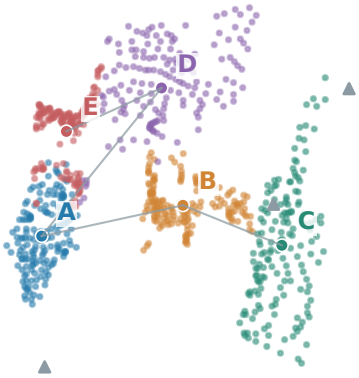
Source view 0



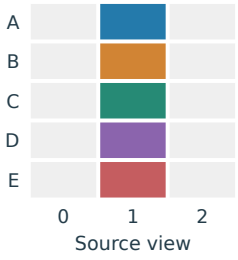
Source view 1



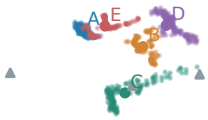
Saved 3DFusionMemory



View support



Overhead



A Potted plant
B Table and chairs
C Wall
D Window
E Structural pillar

Triangles: cameras | Dashed: uncertain relation | Same geometry in both views

All five 3D anchors are supported by view 1, while three camera poses are retained. This distinguishes a multi-camera scene state from multi-view support for an individual object.

0/5 anchors have support in at least two views.

3DFusionMemory: Spatial Representation Memory 06

Representation Example · MindCube around / 0d8a8816

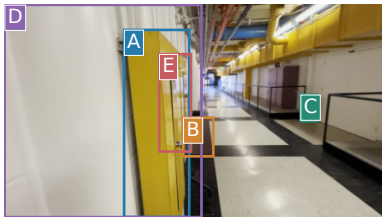
A second saved indoor layout

3 camera views · 5 3D anchors · 10 stored relation pairs

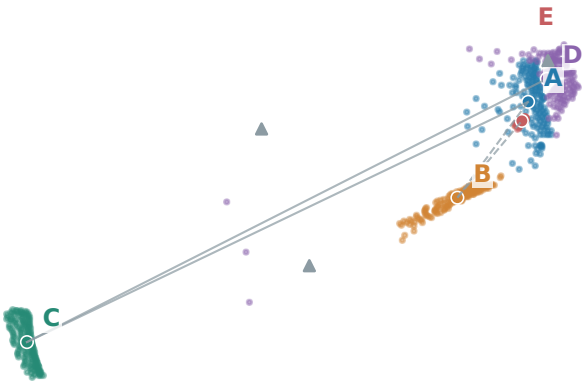
Source view 0



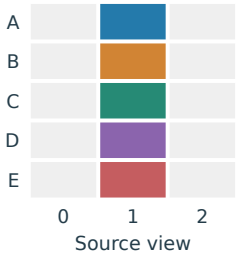
Source view 1



Saved 3DFusionMemory



View support



Overhead



A Yellow cabinets
B Bench
C Trash bin
D Wall
E Cabinet handles

Triangles: cameras | Dashed: uncertain relation | Same geometry in both views

Object identities and their supporting views use the same colors throughout. Blank cells indicate views with no recorded support; the overhead inset shows the same geometry from above.

0/5 anchors have support in at least two views.

C.2 Spatial Skill Library

Skill Library: Reusable Rotation Estimation MMSI 32

Skill Library · Answer Correction

Question

The images are taken continuously from a first-person perspective. In which direction are you rotating?

A. Left B. Right C. Up D. Down



Image 1



Image 2

Direct A. Left × Incorrect

SpatialBrain

Skill
camera-camera-rotation-template
Returned direction
down
Total rotation
1.095 degrees
Primary axis
x

Skill output

1. Invoke camera-camera-rotation-template for the two ordered views.
2. The successful skill returns down, with primary axis x and total rotation about 1.095 degrees.
3. Pass the skill result to final-answer generation. Return D: down, instead of the Direct answer A: left.

Answer D. Down ✓ Correct

The rotation is small. This is reuse of an existing executable skill at state v0, not evidence that a new skill was learned on this item.

Skill Library: Camera-frame Reasoning MMSI 42

Skill Library · Answer Correction

Question

When you took the photo in Figure 1, where was the iron refrigerator located relative to you?

A. Front left B. Rear left C. Rear right D. Front right



Image 1



Image 2

Direct B. Rear left × Incorrect

SpatialBrain

Skill
camera-object-direction-template

Returned direction
left

Camera frame
x = -1.377; z = +0.566

Mapped option
A: front left

Skill output

1. Invoke camera-object-direction-template for the iron refrigerator in the two views, using camera 0 as the reference.
2. The skill returns a camera-frame displacement with x = -1.377 and z = +0.566: left and in front.
3. The skill label is left; the final answer uses the returned coordinates to select the diagonal option A: front left.

Answer A. Front left ✓ Correct

The skill returns a coarse direction; its coordinates resolve the diagonal answer option.

C.3 SPATIALCOGNITION

SpatialCognition: Experience-Guided Routing MMSI 748

SpatialCognition · Efficiency Improvement

Question

Which is longer: the long side of the statue's base in Figure 1 or the width of the bench in Figure 2?

A. They are the same length
B. The long side of the statue's base in Figure 1 is longer
C. Sometimes the long side of the statue's base in Figure 1 is longer, sometimes the width of the bench in Figure 2 is longer
D. The width of the bench in Figure 2 is longer



Image 1



Image 2

Direct D. The width of the bench in Figure 2 is longer ✓ Correct

SpatialBrain

Shared meta-skill guard
Prefer 2D size evidence

Earlier same-family events
27; current target excluded

Policy change
v10: agent / v11: direct

Executed route
Direct; memory off; no skill

Versioned policy and recorded routing decision

1. The visual-measurement meta-skill guard prefers 2D evidence over scale-ambiguous 3D extents. This guard is shared by both arms.
2. After 27 earlier same-family events, the v11 meta-memory recommendation changes from agent to direct. Apply it without exploration.
3. Answer D from the input views, without memory or skill execution. The fixed-policy run instead makes two detector calls.

Answer D. The width of the bench in Figure 2 is longer ✓ Correct

What evolves: the decision to call tools
Accumulated experience changes the route from agent execution to direct reasoning and turns off skill hints. The shared 2D-measurement guard remains unchanged, and the resulting path uses no tools.

Fixed-policy control → SpatialBrain (both correct): 37,884 → 3,632 tokens (−90.4%); 2 → 0 non-final top-level calls; 38.0 → 9.4 s observed.

The standalone Direct answer is also correct: 3,697 versus 3,632 tokens, only a 1.8% difference; both use zero tools.

SpatialCognition: Adaptive Skill Selection MMSI 348

SpatialCognition · Policy Evolution

Question

What kind of angle is approximately formed by the projections on the ground of the forward directions of the camera when taking the first and second pictures?

A. Obtuse angle
B. 0 degrees
C. Right angle
D. Acute angle

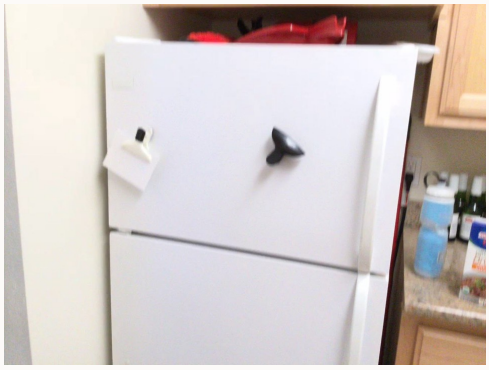


Image 1

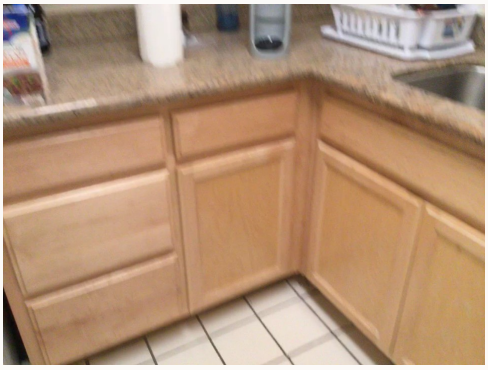


Image 2

Direct D. Acute angle ✓ Correct

SpatialBrain

Skill
camera-camera-rotation-template

Policy
v5 / executable

Returned angle
47.6 degrees

Skill output

1. Earlier same-family experience changes the existing rotation skill from hint-only (v1) to executable (v4).
2. With 18 earlier same-family events, the v5 policy selects and executes this skill.
3. The skill returns 47.6 degrees, mapped to an acute angle.

Answer D. Acute angle ✓ Correct

What evolves: skill selection

Earlier events change the recommendation from hint-only to executable. The target uses the v5 policy to execute the recommended rotation skill, without exploratory routing.

The 18 earlier events exclude this target. The skill already exists, and the fixed-policy control also calls it. The change is in the selection policy, not the skill code or this item's correctness.

Additional execution traces.

Tool Trajectory: Viewpoint Transformation MMSI 3

Tool-grounded reasoning · Answer Correction

Question

When you took the second photo, where was the toilet in relation to you?

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



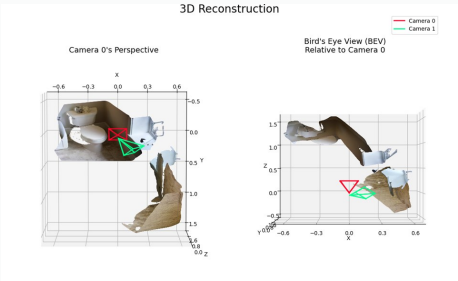
Image 1



Image 2

Direct A. back right × Incorrect

SpatialBrain



Reconstruction and estimated camera poses

1. Detect the toilet in image 1.
2. Reconstruct both views and project the toilet into 3D.
3. Transform it into the second camera frame: x = -0.630, z = -0.038. The toilet is behind and to the left.

Answer D. back left ✓ Correct

The back/front component is small.

Tool Trajectory: Coordinate-convention Alignment MMSI 18

Tool-grounded reasoning · Answer Correction

Question

The camera coordinate system is defined as +Y up, -Z forward, and is a right-handed system. How is the viewpoint in the second image obtained from the viewpoint in the first image by moving in its corresponding camera coordinate system?

A. Move in the positive Z direction and positive Y direction
B. Move in the negative Z direction and positive Y direction
C. Move in the negative X direction and negative Y direction
D. Move in the positive X direction and negative Y direction

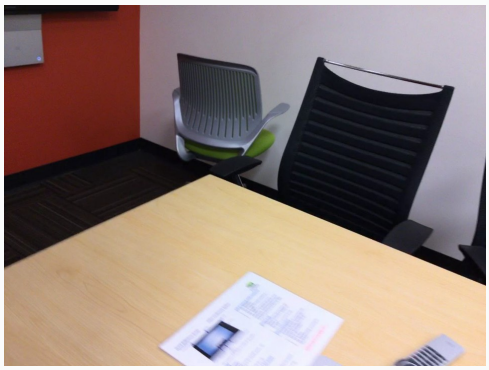


Image 1

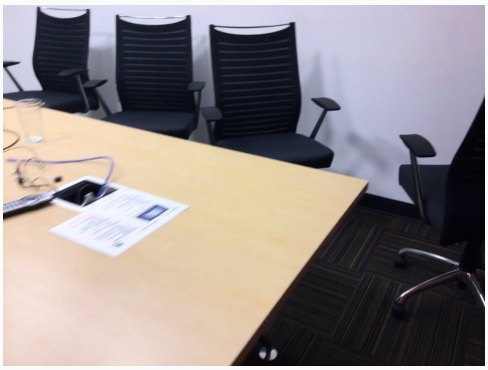
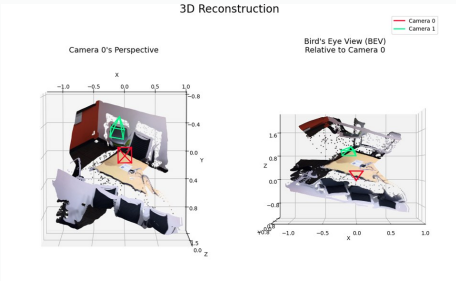


Image 2

Direct D. Move in the positive X direction and negative Y direction × Incorrect

SpatialBrain



Reconstruction and estimated camera poses

1. Reconstruct both views and recover their camera centers.
2. Express the second center in the first camera frame, using the specified +Y-up, -Z-forward convention.
3. The displacement is (-0.083, +0.509, -1.060): negative Z and positive Y. Return B.

Answer B. Move in the negative Z direction and positive Y direction ✓ Correct

Tool Trajectory: Landmark-based Verification MMSI 61

Tool-grounded reasoning · Answer Correction

Question

Assuming the large art display area is on the south wall, where is the ingredient storage area located in the room?

A. Northwest corner
B. Northeast corner
C. Southeast corner
D. Southwest corner



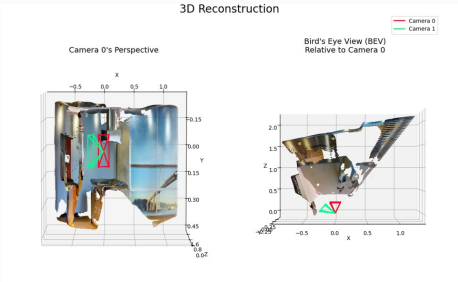
Image 1



Image 2

Direct A. Northwest corner × Incorrect

SpatialBrain



Reconstruction of the art wall and kitchen

1. Detect the wall art and refrigerator; select the art box from three candidates.
2. Estimate the art orientation and project the refrigerator into the reconstructed room.
3. Use the south-wall art as the reference. North = +1.784 and east = +1.119 place the refrigerator in the northeast corner.

Answer B. Northeast corner ✓ Correct

Tool Trajectory: Temporal Correspondence MMSI 82

Tool-grounded reasoning · Answer Correction

Question

The pictures were taken in sequence. In which direction are the two grippers moving?

A. The left gripper did not move, the right gripper moved to the left
B. The left gripper moved to the right, the right gripper moved to the right
C. The left gripper moved to the right, the right gripper did not move
D. The left gripper moved to the left, the right gripper did not move.



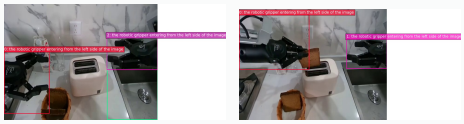
Image 1



Image 2

Direct C. The left gripper moved to the right, the right gripper did not move × Incorrect

SpatialBrain



Detections in the two frames

1. Detect each gripper in both frames and select the corresponding boxes.
2. The selected horizontal centers shift by +57.3 pixels for the left gripper and +17.7 pixels for the right gripper.
3. Both shifts exceed the respective stationary thresholds (9.8 and 7.7 pixels). Return B: both move right.

Answer B. The left gripper moved to the right, the right gripper moved to the right ✓ Correct

Detections contain multiple candidates; the calculation uses the selected box correspondences.

Tool Trajectory: Reduced Tool Use MMSI 289

Tool-grounded reasoning · Correction and Compute Comparison

Question

The images are taken continuously from a first-person perspective. In which direction are you moving?

A. Front left
B. Front right
C. Right
D. Forward



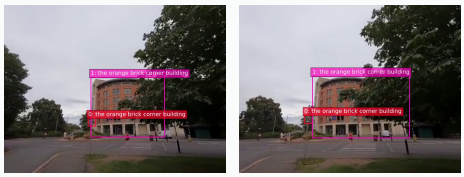
Image 1



Image 2

Direct C. Right × Incorrect

SpatialBrain



Landmark detections in the two input frames

1. Request landmark detections for the brick building and roadside sign in both frames.
2. Proceed from the two detection calls without a reconstruction or Python call.
3. Return B: front right. The fixed-policy run reaches the same answer after an additional reconstruction and two Python calls.

Answer B. Front right ✓ Correct

Fixed-policy control → SpatialBrain (both correct): 88,760 → 29,415 tokens (−66.9%); 5 → 2 non-final top-level calls; 318.6 → 29.7 s observed.

Tool Trajectory: Skill Fallback and Recovery MMSI 880

Tool-grounded reasoning · Compute Comparison

Question

The picture is a continuous first-person perspective shot. In which direction is the spoon moving?

A. Out of the bowl
B. Toward the person
C. Into the bowl
D. Toward the pot



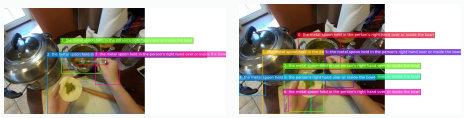
Image 1



Image 2

Direct C. Into the bowl ✓ Correct

SpatialBrain



Spoon detections in the two ordered frames

1. Try the object-motion skill. Insufficient slot information returns control to the planner; no nested tool is executed.
2. Run one spoon-detection call on each frame, then answer C: into the bowl.
3. The fixed-policy run also answers C, but makes eight detector calls after the same skill fallback.

Answer C. Into the bowl ✓ Correct

Fixed-policy control → SpatialBrain (both correct): 90,898 → 52,417 tokens (−42.3%); 9 → 3 non-final top-level calls; 110.5 → 69.4 s observed.

Both runs encounter ambiguous detections. The fallback remains included in the call and token accounting.

Tool Trajectory: Compact Visual Measurement — MMSI 715

Tool-grounded reasoning · Compute Comparison

Question

Which is higher, the fabric on the left or the sink on the right?

A. The same height
B. The sink
C. The fabric
D. Sometimes the fabric is higher, sometimes the sink is higher



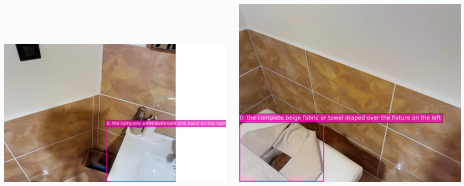
Image 1



Image 2

Direct B. The sink ✓ Correct

SpatialBrain



Detected sink (left) and fabric (right)

1. Detect the sink in image 1 and the fabric in image 2.
2. Compare normalized box-center heights: sink 0.792, fabric 0.825.
3. Return the sink. This trajectory uses 2D measurements without 3D reconstruction.

Answer B. The sink ✓ Correct

Fixed-policy control → SpatialBrain (both correct): 64,712 → 58,586 tokens (−9.5%); 3 → 3 non-final top-level calls; 196.1 → 102.6 s observed.

Both runs use the same memory-off/hint route. Direct is also correct and uses only 3,534 tokens.

Tool Trajectory: Compute Trade-offs — MMSI 416

Tool-grounded reasoning · Compute Comparison

Question

This is a set of photos I took at home. In the last photo, in which direction is the exit area relative to me?

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

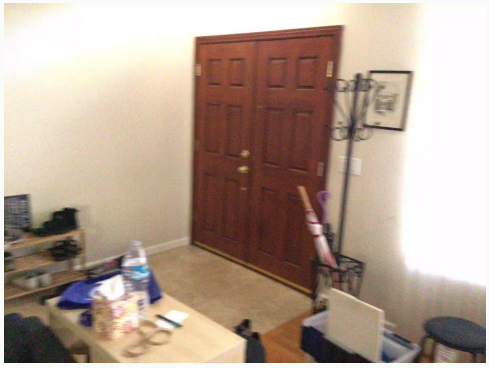


Image 1

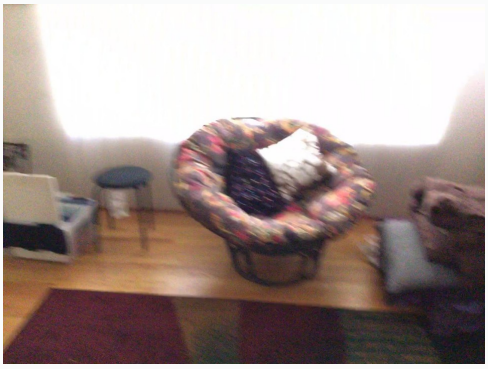
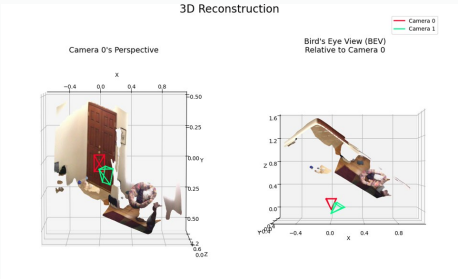


Image 2

Direct A. front left ✓ Correct

SpatialBrain



Reconstruction from the two input views

1. Detect the exit in both images.
2. Reconstruct both views and project the exit into 3D.
3. Transform it into the last camera frame: x = -0.852, z = +0.987. The exit is ahead and to the left.

Answer A. front left ✓ Correct

Fixed-policy control → SpatialBrain (both correct): 123,214 → 111,609 tokens (−9.4%); 4 → 5 non-final top-level calls; 258.9 → 262.3 s observed.

This trajectory tests an exploratory route with memory disabled; fewer tokens do not imply fewer calls or shorter time.

D Detailed Benchmark Results

We report all task categories for the Direct/SpatialBrain comparison in Table 1. Both methods use GPT-5.6-sol. Missing or unparsed answers score zero, and overall scores weight each item equally. Differences (Δ, percentage points) are computed before rounding. These tables compare the complete framework with direct answers; component conditions are reported separately where available. Tables A5, A6, A7, and A8 give the corresponding category breakdowns for MMSI, OmniSpatial, SAT, and SpaceDG.

D.1 MindCube

Table A3 uses the same 920/1,050 SpatialBrain run (87.6%) as Table 1 and the Full result in Table A4. Component settings are given in Appendix A.

Table A3: **MindCube results by task category** (accuracy %).

Task category	N	GPT-5.6-sol	SpatialBrain	Δ
Among	600	55.8	84.7	+28.8
Around	250	78.8	86.4	+7.6
Rotation	200	87.0	98.0	+11.0
Overall	1,050	67.2	87.6	+20.4

Table A4: **MindCube component results by task category** (accuracy %). Values shown in Figure 6; all conditions cover 1,050 items. Settings are given in Appendix A.

Condition	MindCube Overall	Among	Around	Rotation
Local target N	1,050	600	250	200
GPT-5.6-sol	67.2	55.8	78.8	87.0
w/o Skill Library	76.7	72.0	73.2	95.0
w/o 3DFusionMemory	75.2	71.0	70.8	93.5
w/o SpatialCognition	78.8	74.0	75.2	97.5
SpatialBrain	**87.6**	**84.7**	**86.4**	**98.0**

D.2 MMSI

Table A5: **MMSI results by task category** (accuracy %).

Task category	N	GPT-5.6-sol	SpatialBrain	Δ
Position (Cam.–Cam.)	93	65.6	69.9	$+4.3$
Position (Cam.–Obj.)	86	38.4	72.1	$+33.7$
Position (Cam.–Reg.)	83	60.2	63.9	$+3.6$
Position (Obj.–Obj.)	94	45.7	48.9	$+3.2$
Position (Obj.–Reg.)	85	44.7	36.5	-8.2
Position (Reg.–Reg.)	81	54.3	54.3	$+0.0$
Motion (Cam.)	74	66.2	73.0	$+6.8$
Motion (Obj.)	76	51.3	46.1	-5.3
Attribute (Appr.)	66	40.9	62.1	$+21.2$
Attribute (Meas.)	64	60.9	56.2	-4.7
MSR	198	58.1	52.5	-5.6
Overall	1,000	53.8	57.1	$+3.3$

D.3 OMNISPATIAL

Table A6: **OmniSpatial results by task category** (accuracy %).

Task category	N	GPT-5.6-sol	SpatialBrain	Δ
Allocentric	376	59.6	63.8	$+4.3$
Egocentric	102	86.3	80.4	-5.9
Hypothetical	83	67.5	68.7	$+1.2$
Motion analysis	346	72.0	76.0	$+4.0$
Manipulation	74	77.0	71.6	-5.4
Traffic analysis	85	77.6	69.4	-8.2
Localization	105	79.0	80.0	$+1.0$
Geospatial strategy	110	86.4	82.7	-3.6
Geometric reasoning	155	48.4	45.8	-2.6
Pattern recognition	97	55.7	60.8	$+5.2$
Overall	1,533	68.3	69.1	$+0.8$

D.4 SAT

All 150 items use fixed option order; this is not circular evaluation.

Table A7: **SAT results by task category** (accuracy %).

Task category	N	GPT-5.6-sol	SpatialBrain	Δ
Object movement	23	100.0	100.0	$+0.0$
Ego movement	23	100.0	100.0	$+0.0$
Action consequence	37	89.2	89.2	$+0.0$
Perspective	33	84.8	84.8	$+0.0$
Goal / aim	34	88.2	88.2	$+0.0$
Overall	150	91.3	91.3	$+0.0$

The equal category totals do not imply identical predictions: six items are corrected and six regress.

D.5 SPACEDG

The fixed seed-0 1,000-item subset uses accuracy for categorical items and mean relative accuracy (MRA) for numeric items. Numeric questions receive fractional credit, so these scores are not integer correct-answer counts.

Table A8: **SpaceDG results by task category** (accuracy / MRA %).

Task category	N	GPT-5.6-sol	SpatialBrain	Δ
Camera translation	81	56.5	64.0	$+7.4$
Camera rotation	74	82.4	91.9	$+9.5$
Camera object relative direction	174	60.3	64.9	$+4.6$
Cross view cardinal direction	79	36.7	50.6	$+13.9$
Object proxy cardinal direction	86	32.6	40.7	$+8.1$
Inter object distance	27	58.5	24.8	-33.7
Camera object distance estimation	142	47.3	43.6	-3.7
Object size comparison	74	85.1	79.7	-5.4
Object bounding size estimation	96	50.2	43.7	-6.5
Object counting	85	54.1	41.2	-12.9
Object existence estimation	82	73.2	81.7	$+8.5$
Overall	1,000	56.9	57.9	$+1.0$

E EXPERIENCE UPDATE DETAILS

This section separates observation fusion within a scene from the admission of resource-use strategies across tasks. Neither operation updates model weights.

Confidence-weighted scene fusion. Reconstruction provides a common scene frame. For a matched object o, $\boldsymbol{\mu}_o \in \mathbb{R}^3$ is its current center and W_o its accumulated confidence weight. A projection from view v supplies center $\widetilde{\boldsymbol{\mu}}_{o,v}$ and weight $w_{o,v}$:

$$\boldsymbol{\mu}_o^+ = \frac{W_o \boldsymbol{\mu}_o + w_{o,v}\widetilde{\boldsymbol{\mu}}_{o,v}}{W_o + w_{o,v}}, \qquad W_o^+ = W_o + w_{o,v}. \tag{A1}$$

The superscript $+$ denotes the updated state. The view-specific center is the confidence-weighted centroid of projected points, or their arithmetic mean if total confidence is zero; $w_{o,v}$ is their mean confidence. The first valid projection initializes the estimate. Equation A1 requires $W_o + w_{o,v} > 0$. These weights express relative support, not calibrated probabilities. This update combines observations of the same scene object, not coordinates from different tasks.

Conservative strategy admission. In the executable configuration, consider one route or memory factor with default action a_0 and alternative a. Statistics are conditioned on the task signature and use only earlier blocks. Eligibility requires

$$\begin{gathered} \min\{n_a, n_{a_0}\} \geq n_{\min}, \qquad n_u^{\text{cost}} = n_u \;\; (u \in \{a, a_0\}), \\ R_a^{\text{upper}} \leq \epsilon, \qquad C_a^{\text{upper}} \leq (1-\delta) C_{a_0}^{\text{lower}}. \end{gathered} \tag{A2}$$

For action u, n_u counts observations and n_u^{cost} complete cost records; $n_{\min}$ is the minimum support. Incomplete costs cannot establish a saving. R_a^{upper} is a Wilson upper estimate of the process-violation rate, including unverified actions; ϵ is its tolerance. The cost bounds use the sample mean plus or minus 1.96 standard errors, with the lower bound floored at zero. $\delta \in [0, 1)$ is the required fractional saving. Thus equation A2 compares a candidate's upper cost with the default's lower cost, not just two point estimates.

The controller checks the specific context before a broader task family; supported context-specific risk can veto a broader recommendation. At each level it chooses the eligible alternative with the lowest upper cost, retaining the default if none qualifies. These are process-feedback gates, not correctness rewards or guarantees of answer accuracy. They describe the executable configuration, not every historical run in the paper.

What is learned and what is fixed. The initial library, task-signature schema, tool interfaces, validation checks, and update rules are specified before evaluation. Experience changes support statistics, associations between task conditions and strategies, and procedural availability. The executable configuration can also compile and validate new procedures. Some failure-check text comes from predefined rules: observing a failure adds its applicability and support, rather than inventing an unrestricted instruction. A selected skill, an exposed hint, an executed procedure, a verified return, and a correct answer are distinct events. Process observations cannot establish answer correctness without the separate evaluation labels, which are not used by the update loop.