MINDTRAVERSE: BENCHMARKING CROSS-SCALE SPATIAL COMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world spatial tasks often exhibit cross-scale structure: models must organize and propagate spatial context across continuous sequences of local spatial judgments and global environmental reasoning. Yet such *cross-scale spatial composition* has been largely overlooked by existing benchmarks, which either focus on single-step reasoning or fail to capture the cross-scale nature of real-world space in multi-step tasks. To investigate this underexplored problem, we construct MINDTRAVERSE. The benchmark recursively composes atomic task interfaces to create multi-step cross-scale reasoning problems validated through both geometric and visual checks. In MINDTRAVERSE, models must continuously propagate and reason over spatial context between fine-grained, object-level workspace reasoning and coarse-grained, environment-level transit reasoning. The resulting benchmark contains 6,167 multi-view questions covering 35 composition templates, six scale sequences, and 210 simulator scenes. Across 20 model configurations, even the best-performing model achieves only 50.92% overall accuracy. More importantly, endpoint accuracy does not decrease monotonically as the intended compositional dependency becomes deeper. To investigate this counterintuitive behavior, we conduct further controlled experiments and reveal three systematic deficiencies in current spatial reasoning models. First, **Structural Binding Failure**: models recover task-relevant semantic entities but struggle to organize them into coherent cross-scale structure. Second, **State–Evidence Conflict**: even when explicit spatial state is available, visual evidence is not integrated reliably and can interfere with downstream reasoning. Third, **Trajectory-Role Shortcut**: on some high-accuracy tasks, models bypass the intended cross-scale dependency and collapse onto simple trajectory roles. These results show that genuine cross-scale spatial reasoning requires not only correct final answers, but also correct spatial representation, reliable evidence integration, and faithful downstream use of that representation.

1 INTRODUCTION

Real-world spatial tasks inherently span multiple spatial scales. A robot may need to perform fine-grained operations within a local workspace, move through global connecting regions, and then enter another workspace to continue reasoning (Garrett et al., 2021; Shridhar et al., 2020; Szot et al., 2021; Li et al., 2023). These stages form an interdependent reasoning chain: a spatial result obtained at one scale may directly determine the operation required at the next. Such cross-scale state transfer connects local spatial perception to long-horizon task execution, and its failure may leave subsequent operations grounded in an incorrect spatial context. We refer to this class of problems as **cross-scale spatial composition** (Figure 1).

Although existing benchmarks have expanded spatial evaluation to multi-view perception, spatial mental modeling, and multi-hop reasoning (Yang et al., 2025; 2026c; Wang et al., 2026; 2025; Wu et al., 2026; Ke et al., 2026), they still primarily evaluate relatively independent spatial operations or multi-step reasoning within a single scale, lacking systematic evaluation of cross-scale state transfer and continuous reasoning.

To formalize this problem, we distinguish two operational scales. **Workspace-scale reasoning** (`W`) handles fine-grained, object-level spatial relations and localization within local workspaces, such as tabletops, shelves, or workstations. **Transit-scale reasoning** (`T`) handles coarser, environment-level

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.

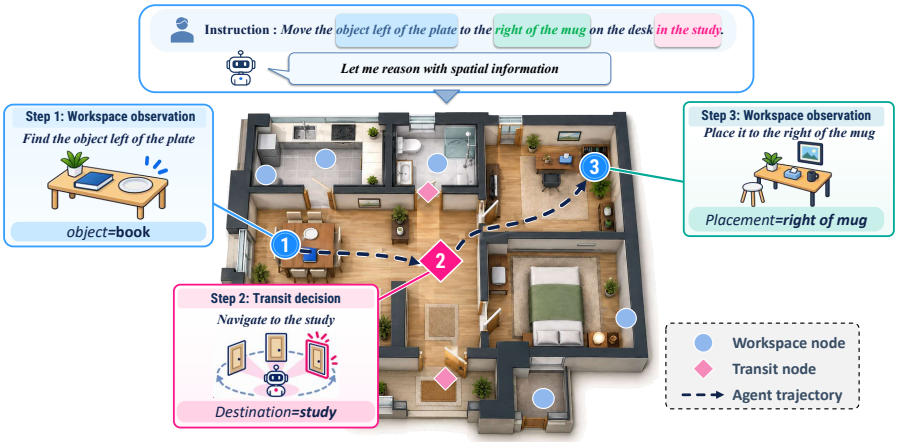


Figure 1: **Cross-scale spatial composition:** a workspace-scale result guides subsequent transit, whose outcome establishes the context for the next workspace reasoning step. The map and intermediate states are shown only for illustration.

orientation, connectivity, and movement decisions between workspaces. A cross-scale task is

$$\mathbf{s}_{t+1} = \mathcal{O}_{\sigma_{t+1}}\left(\mathbf{x}_{t+1}; \mathbf{s}_t\right), \qquad \sigma_{t+1} \in \{\mathtt{W}, \mathtt{T}\}. \tag{1}$$

Here, $\mathbf{x}_{t+1}$ denotes the observations available at the current stage, and $\mathbf{s}_t$ denotes the task-relevant spatial state formed from preceding stages.

Based on this formulation, we construct MINDTRAVERSE. We represent basic W- and T-scale spatial operations as atomic interfaces with **typed outputs**, and recursively bind each intermediate result to the subsequent operation through result-conditioned composition. The composed tasks are then grounded in simulator scenes and jointly validated geometrically and visually to ensure answer correctness and visual answerability. MINDTRAVERSE ultimately contains 6,167 questions covering 35 composition templates, six scale sequences, and 210 simulator scenes. Unlike prior multi-view or multi-step spatial benchmarks, MINDTRAVERSE explicitly couples operations across scales through result-conditioned dependencies, enabling direct diagnosis of cross-scale state transfer. Decoupling symbolic composition from grounding further enables task instantiation across compatible scenes.

Across 20 model configurations, overall performance remains limited, with even the best-performing model achieving only 50.92% accuracy. More importantly, we observe a pronounced *composition-depth-performance decoupling*: performance varies sharply across composition sequences rather than decreasing monotonically with composition depth.We analyze three perspectives: spatial representation, reasoning over spatial representations, and shortcut reliance, revealing three recurring failure modes in our diagnostics. **Structural Binding Failure.** Models recover task-relevant semantic entities but struggle to bind them into the correct cross-scale structure. Among questions requiring both transition-door and workspace grounding, only 53.2% of predicted maps correctly ground both. **State Evidence Conflict.** MLLMs struggle to stably integrate visual evidence with explicit spatial state. Even with GT Map and exact Pose provided, adding RGB reduces overall accuracy from 54.0% to 47.5%. **Trajectory Role Shortcut.** Models can collapse onto simple trajectory roles to bypass full reasoning. In the T-W-T diagnostic, multiple models almost always select specific roles such as the unvisited room or hub rather than using the complete cross-scale context. Genuine cross-scale reasoning therefore requires correct state formation, reliable evidence integration, and faithful state use beyond endpoint accuracy.

Our contributions are threefold. **1)** We introduce and formalize *cross-scale spatial composition*, a long-understudied problem that asks whether spatial state can be continuously transferred between fine-grained object-level reasoning and coarse-grained environment-level reasoning, and whether it genuinely governs downstream decisions. **2)** We construct MINDTRAVERSE, which turns cross-scale spatial dependencies into controlled, scalable, and diagnostic multi-step spatial tasks through typed atomic interfaces, result-conditioned recursive composition, and joint geometric and visual validation. **3)** Across 20 model configurations, we identify a counterintuitive cross-scale performance pattern and systematically localize current MLLM bottlenecks across spatial representation, reasoning over spatial representations, and shortcut reliance.

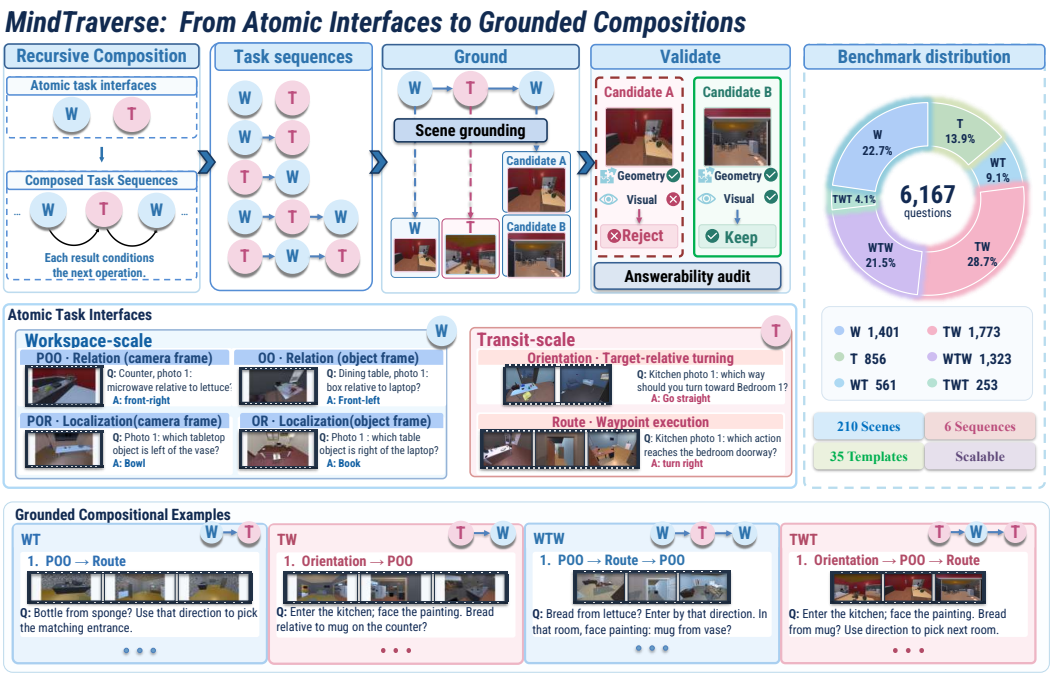


Figure 2: **MindTraverse benchmark construction:** typed workspace- and transit-scale interfaces are recursively composed into cross-scale task sequences, grounded into concrete scenes, and validated through geometric and visual checks. The pipeline yields grounded task instances across six scale sequences and 35 composition templates.

2 Related Work

Spatial intelligence benchmarks. Early spatial benchmarks focused on object relations, localization, and situated question answering from images or reconstructed 3D scenes (Yang et al., 2019; Liu et al., 2023; Azuma et al., 2022; Ma et al., 2023; Fu et al., 2024; Rajabi & Kosecka, 2024). Recent work has expanded evaluation to video and multi-view observations, spatial mental modeling, cross-view and cross-scale localization, multi-hop reasoning, and long-horizon understanding (Yang et al., 2025; Wang et al., 2026; Yang et al., 2026c; Ma et al., 2025a; Wang et al., 2025; Li et al., 2025b;c; Zhu et al., 2026; Lin et al., 2025; Yang et al., 2026b; Ke et al., 2026). SpaCE-10 (Gong et al., 2026) evaluates combinations of atomic spatial capabilities, while MindTraverse explicitly binds workspace and transit operations through intermediate results and uses matched interventions to test whether these results govern downstream decisions.

Spatial representations and reasoning. Existing methods improve spatial reasoning either by learning spatial regularities from data and training (Chen et al., 2024; Ma et al., 2025c; Cai et al., 2026; Yang et al., 2026b; Ouyang et al., 2025; Wu et al., 2025a; Yang et al., 2026a; Li et al., 2026b; Wu et al., 2025b), or by introducing explicit geometric information such as depth, scene graphs, positional encodings, reconstruction, and 3D representations (Cai et al., 2025; Cheng et al., 2024; Zhu et al., 2025; Fan et al., 2026; Ma et al., 2025b; Li et al., 2026a; Zhang et al., 2026). Rather than asking only whether these representations improve a prediction, we study whether task-relevant spatial state can be formed, carried across spatial scopes, and used to guide subsequent decisions.

3 Construction of MindTraverse

3.1 Overview

MindTraverse operationalizes cross-scale spatial composition through four stages: atomic spatial interfaces, recursive composition, scene grounding, and geometric–visual validation (Figure 2). This pipeline supports scalable benchmark construction with automated question generation over procedural scenes.

The benchmark contains 6,167 questions from 210 simulator scenes across six composition sequences. Each question provides ordered RGB observations, a natural-language instruction, and a closed set of answer options. Models receive all observations jointly, without access to maps, camera poses, task programs, or simulator state.

The key design principle is that composition is defined by *dependency*, rather than merely by task length: an intermediate spatial result must change the condition of a subsequent operation. This allows MINDTRAVERSE to control how spatial information is formed and reused as the task transitions between workspace- and transit-scale reasoning.

Although existing benchmarks cover complementary aspects of multi-view, multi-step, and multi-scale spatial reasoning, Table 1 highlights MINDTRAVERSE's explicit scale sequences and result-conditioned cross-scale handoffs.

Table 1: **MINDTRAVERSE explicitly composes dependencies across spatial scales.** Comparison of benchmark scale structure and task construction. Cross-scale handoff means that an intermediate result conditions the next-scale operation. ✓: explicitly supported; △: partial or semi-automatic support; –: not supported.

Benchmark	#Q	Multi-step	Multi-scale	Scale Seqs.	Cross-scale Handoff	Auto QA
MMSI-Bench (Yang et al., 2026c)	1,000	✓	✓	–	–	–
MindCube (Wang et al., 2026)	21,154	–	–	–	–	✓
SpaCE-10 (Gong et al., 2026)	5k+	✓	✓	–	–	△
MINDTRAVERSE	**6,167**	✓	✓	**6**	✓	✓

We next describe how these dependencies are constructed from typed atomic interfaces.

3.2 ATOMIC INTERFACES AND RECURSIVE COMPOSITION

Typed spatial outputs are recursively bound across scales. We define workspace-scale interfaces (`W`) for local object relations and localization, and transit-scale interfaces (`T`) for target-relative orientation and route selection. A `W` interface returns an entity or direction, whereas a `T` interface returns a turn or action. Executing a `T` interface also produces an arrival context, including the destination room and transit pose, which can condition a subsequent `W` operation.

These typed outputs form the variables passed between composed operations. Figure 2 illustrates these handoffs, while Algorithm 1 formalizes their recursive binding.

Algorithm 1 Schematic result-conditioned composition

Require: initial interface I_1, target sequence S
$P \leftarrow [I_1]$
for $k = 1, \ldots, |S| - 1$ **do**
  $r_k \leftarrow \textsc{SymbolicOutput}(P_k)$
  $I_{k+1} \leftarrow \textsc{SampleCompatible}(S_{k+1})$
  $b_k \leftarrow \textsc{SelectBinding}(r_k, I_{k+1})$
  $I_{k+1} \leftarrow \textsc{Bind}(I_{k+1}, r_k; b_k)$
  $P \leftarrow P \oplus I_{k+1}$
end for
return composed program P

Each intermediate result must serve as a condition for selecting the next interface, rather than merely appending another query. Repeated application yields six sequence families: `W`, `T`, `W-T`, `T-W`, `W-T-W`, and `T-W-T`. Across these sequences, we instantiate 35 operator–binding templates that vary in their atomic operations, reference frames, and intermediate results passed between interfaces.

3.3 GROUNDING COMPOSITIONS INTO SCENES

Grounding instantiates abstract compositions as concrete scenes while preserving their dependencies. An abstract composition specifies spatial roles and dependencies, but not the concrete entities that realize them. We ground each program in adapted ProcTHOR scenes and render observations with AI2-THOR (Deitke et al., 2022; Kolve et al., 2017). Abstract roles are bound to type-compatible objects, workspaces, entrances, anchors, and destinations.

A valid grounding preserves the information flow encoded by the original interface sequence: an orientation result may constrain a later route, a selected entity may determine the next target, and an arrival state may establish the workspace context for a subsequent operation.

However, structural and syntactic validity alone does not guarantee that these dependencies can be grounded in visual evidence, nor that they can be recovered from that evidence. We therefore add *bridge observations* at each `W-T` or `T-W` boundary to connect the upstream exit state with the downstream entry context. When room identity is required, we additionally include *room-semantic observations* that expose diagnostic objects associated with the room. These views ensure that cross-scale handoffs are supported by RGB evidence rather than existing only in symbolic state.

The grounded program records both the abstract dependency and the evidence required to execute it visually. Different compositions, additional views, or language variants constructed from the same ProcTHOR scene remain realizations of the same underlying scene. The same abstract program can be grounded in compatible scenes, yielding diverse tasks with a shared composition structure.

3.4 Geometric Validity and Visual Answerability

Geometry determines the answer; visual evidence makes it recoverable. Geometric validation ensures that each grounded program induces a unique and consistent answer. We verify object and target uniqueness, stability of the required spatial relations, and consistency of route states, then replay the composition to derive the answer. Ambiguous or inconsistent instances are rejected.

Geometric correctness alone does not ensure visual answerability. Every dependency used by the program must also be **supported by RGB observations**: query entities must be recognizable, entrances and destinations visually associable, and cross-scale transitions connected by sufficient shared or overlapping evidence. When the observations are insufficient, we add connecting views or reject the candidate. Simulator reachability or connectivity is never treated as visual evidence. As an additional audit, we **manually review every question** from a fixed 60-scene subset and verify that the public observations support the complete reasoning sequence required by the grounded program. Finally, visual evidence must also be interpretable in the **reference frame specified by each operation**. Visibility alone is insufficient: an orientation anchor may be visible without defining the query frame, and movement headings follow the grounded program rather than incidental camera rotations. We therefore validate both the visibility and recoverability of the required spatial relations under the task-defined reference frame.

4 Dissecting Cross-Scale Spatial Composition

The six native sequences in MindTraverse provide controlled variations in how spatial state is formed, transferred, and reused across workspace and transit scales. We first characterize the overall behavior of models across these composition structures, and then dissect the cross-scale reasoning chain along three key stages: *spatial representation formation*, *reasoning over spatial representations*, and *dependency faithfulness*. Specifically, we ask whether models can form task-relevant spatial representations that connect local workspaces with the surrounding transition structure, whether explicit spatial state can be reliably integrated with visual evidence, and whether final decisions are genuinely governed by the intended cross-scale dependencies. Matched controls and counterfactual interventions are used to isolate these stages of failure.

4.1 Experimental Setup

We evaluate 20 direct-answer model configurations on all 6,167 questions from 210 scenes (Table 2), covering frontier models, general open-weight models, and models with additional spatial training.

Evaluation protocol. In the standard evaluation, each question provides the instruction, answer options, and ordered RGB observations. Sequence accuracy is computed over all questions in each sequence, while Overall accuracy is computed over all 6,167 questions. Model configurations use their own prompt wrappers, decoding settings, and inference budgets and are therefore not strictly compute-matched. Further diagnostic analyses primarily use Qwen3.7-Flash on Tiny600, which contains 100 questions per sequence.

Table 2: **Composition-Depth-Performance Decoupling.** Accuracy across 20 model configurations and six scale sequences. Bold/underline: best/second-best model scores. Peach: best Overall; blue: diagnostic backbone. Some models do not support disabling thinking (†), resulting in different reasoning budgets across models (Appendix B). Citations refer to models or model families.

Method	Accuracy (%) ↑						
	Overall	Single scale		One transition		Two transitions	
		W	T	W-T	T-W	W-T-W	T-W-T
Chance level (random)	27.71	28.77	23.33	25.00	28.29	29.16	31.13
Human baseline	89.49	93.15	82.81	84.62	91.67	94.20	74.29
Frontier models							
Gemini 3.7 Flash†	**50.92**	**45.68**	**56.89**	**93.94**	**46.87**	**39.68**	**51.38**
Qwen3.7-Flash (Qwen Team, 2026b)	33.24	36.33	33.29	81.82	23.07	20.71	45.06
GPT-5.6 Luna	29.12	30.41	26.52	55.61	23.63	22.30	<u>46.25</u>
Seed 2.0 Mini (Bytedance Seed, 2026)	25.52	31.41	23.13	47.59	16.47	20.56	41.50
General open-weight models							
GLM 5.3 Flash† (GLM-5-Team et al., 2026)	<u>37.73</u>	34.76	<u>36.92</u>	<u>88.95</u>	30.68	27.66	45.45
Qwen3.5-27B (Qwen Team, 2026a)	31.46	39.04	27.10	77.18	21.04	18.75	42.29
Qwen3.5-4B (Qwen Team, 2026a)	27.71	32.12	25.00	25.13	24.42	27.51	42.29
Kimi K2.5 (Kimi Team et al., 2026)	28.70	36.62	26.87	59.00	17.99	20.03	44.27
NVILA-8B (Liu et al., 2025)	30.58	26.48	36.33	37.61	<u>31.92</u>	26.76	28.85
LLaVA-OneVision-Qwen2-7B (Li et al., 2025a)	28.49	24.20	26.87	47.77	28.14	25.70	32.02
Qwen3-VL-2B-Instruct (Bai et al., 2025)	27.03	29.19	30.14	33.87	26.28	23.28	14.23
Qwen3-VL-4B-Instruct (Bai et al., 2025)	23.64	25.05	22.43	52.23	15.91	17.16	44.66
Qwen3-VL-8B-Instruct (Bai et al., 2025)	26.14	28.05	28.62	48.13	18.10	20.56	43.87
Spatially trained open-weight methods							
Molmo2-ER (Fang et al., 2026)	34.30	<u>40.90</u>	33.53	74.15	24.37	21.92	<u>46.25</u>
SenseNova-SI (Cai et al., 2026)	33.10	40.83	32.01	57.58	24.87	25.77	35.57
SpaceR (Ouyang et al., 2025)	29.16	31.48	21.38	29.77	29.05	<u>30.99</u>	32.41
VLM-3R (Fan et al., 2026)	25.51	25.27	35.28	42.42	19.97	21.09	18.18
Spatial-MLLM (Wu et al., 2025a)	25.17	21.63	30.37	40.64	19.80	23.89	37.15
VST (Yang et al., 2026a)	23.63	23.55	25.00	55.61	15.74	16.10	43.08
GeoThinker-7B (Li et al., 2026a)	23.09	19.91	26.99	28.16	22.05	21.62	31.23

4.2 Composition-Depth-Performance Decoupling

Each recursive extension adds one result-conditioned operation to the intended dependency chain. Across 20 model configurations, we observe a clear *Composition-Depth-performance decoupling*: even the best-performing model achieves only 50.92% overall accuracy, while performance does not decrease monotonically with composition depth and instead varies sharply across composition sequences (Table 2). For example, Gemini scores 93.9%, 39.7%, and 51.4% on `W-T`, `W-T-W`, and `T-W-T`, respectively, while GLM shows the same pattern at 88.9%, 27.7%, and 45.5%. Notably, `W-T-W` and `T-W-T` contain the same number of stages but still differ substantially in accuracy.

This decoupling shows that the actual difficulty of cross-scale tasks cannot be characterized by composition depth alone, and model performance does not vary monotonically with composition depth. We therefore dissect the cross-scale reasoning chain along three stages: *spatial representation formation*, *reasoning over spatial representations*, and *dependency faithfulness*, asking whether spatial state is formed correctly, reliably integrated with visual evidence, and genuinely used to control downstream decisions.

4.3 Finding I: Structural Binding Failure

We use predicted maps as an explicit diagnostic representation to examine whether models can recover task-relevant entities and organize them into cross-scale spatial structure.

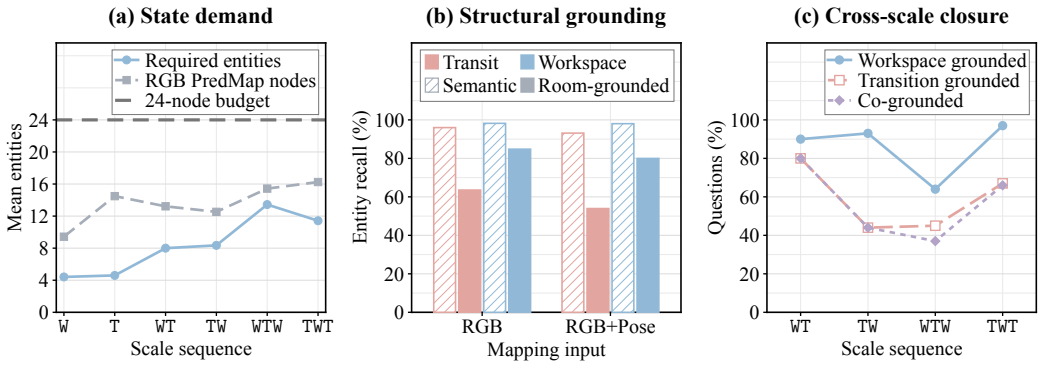


Figure 3: Spatial representation diagnostics on Tiny600. We compare (a) required entities and predicted nodes, (b) semantic and room-grounded recall of transition doors and workspaces, and (c) joint door–workspace grounding, to localize where cross-scale spatial structure breaks down.

Structural Binding Failure. *Models recover many task-relevant entities, yet often fail to bind them into a coherent room–door–workspace structure.*

Map capacity is not the bottleneck. We recover a *choice-complete* entity set $E_{\text{req}}(q)$ from each private task program, containing the executed chain and the physical alternatives required for the decision, including rooms, doors, workspaces, orientation anchors, and queried or candidate objects. As composition extends across scales, both the size and diversity of the required spatial state increase (Figure 3(a)). However, predicted maps generate at least as many nodes as the task requires on average, and the largest extracted requirement remains below the 24-node map budget. Limited node capacity therefore does not explain the missing task-relevant content.

Entities are recovered, but structural bindings are not. Cross-scale representation requires not only recovering relevant entities, but also binding them into the correct spatial topology. Semantic matching checks whether an entity is recovered by type and category, whereas room-grounded matching further requires that it be associated with the correct semantic room. Figure 3(b) reveals a substantial gap between the two: transition doors and workspaces are often recovered semantically, while their room associations are markedly less reliable.

To test whether this gap primarily arises from multi-view localization errors, we repeat map construction with exact camera position and orientation. Yet the semantic-to-grounded gap persists even with privileged pose, especially for transition doors. Exact pose is therefore insufficient to recover the structural bindings required for cross-scale reasoning.

Binding failures prevent cross-scale closure. Cross-scale composition ultimately requires the local workspace and its corresponding transition structure to be grounded jointly. We analyze the 455 Tiny600 questions that require both a transition door and a workspace. Figure 3(c) shows that workspace grounding remains relatively strong, whereas transition grounding is substantially weaker, leaving only 53.2% of applicable maps with both sides jointly grounded. In `W-T-W`, joint co-grounding falls further to 37.0%.

Predicted maps therefore often recover the local workspace without reliably embedding it into the correct room–door transition structure. This reveals a more specific bottleneck in cross-scale representation: models do not simply miss task-relevant entities; they struggle to preserve the structural bindings that connect spatial information across scales. We next ask whether this bottleneck limits downstream reasoning when the correct spatial state is explicitly provided.

4.4 Finding II: State Evidence Conflict

State–Evidence Conflict. *Visual evidence can undermine reasoning even when correct spatial state is explicitly available.*

Table 3: **Explicit state helps, but RGB can induce State–Evidence Conflict.** Tiny600 interventions on spatial state and visual evidence. Blue: representation controls; peach: state–evidence integration; gray: privileged GT Map + pose.

Map and pose			RGB input		Accuracy (%) ↑						
PredMap	GT Map	Pose	Mapping	Reasoning	W	T	W-T	T-W	W-T-W	T-W-T	Overall
Representation controls											
✓	—	—	✓	—	42.0	25.0	86.0	37.0	27.0	38.0	42.5
✓	—	✓	✓	—	31.0	42.0	53.0	34.0	28.0	32.0	36.7
—	✓	✓	—	—	52.0	39.0	60.0	59.0	58.0	56.0	**54.0**
State–evidence integration											
—	✓	✓	—	—	52.0	39.0	**60.0**	59.0	**58.0**	56.0	**54.0**
—	✓	✓	—	✓	40.0	27.0	**83.0**	52.0	**30.0**	53.0	**47.5**

State quality affects downstream reasoning. Under shared reasoning rules, replacing the pose-conditioned PredMap with a complete GT Map improves Tiny600 accuracy by 17.3 percentage points (from 36.7% to 54.0%; Table 3, representation controls), while PredMap without pose reaches 42.5%. Both PredMap variants use RGB for mapping; the RGB columns distinguish mapping inputs from images supplied to downstream reasoning.

Supplied state recovers downstream reasoning. On 37 matched W-T-W episodes, providing the source relation and each room's target relation with transit RGB raises joint room-and-relation accuracy from 4/37 with Full RGB to 33/37 (Table 4). Once the local relations are supplied, the model can usually select the destination and retrieve its target relation. We next examine what RGB contributes when explicit state is available.

Table 4: **Supplied state recovers downstream reasoning.** W-T-W state controls on 37 matched episodes. Correct requires both room and relation.

Condition	Correct
Full RGB	4/37
Supplied state	33/37

Visual evidence can conflict with explicit state. Holding the simulator-derived GT Map and exact Pose fixed, adding RGB reduces Tiny600 accuracy from 54.0% to 47.5% (Table 3, state–evidence integration). The effect also reverses across compositions: workspace–transit (W-T) improves from 60% to 83%, whereas workspace–transit–workspace (W-T-W) drops from 58% to 30%. Explicit state therefore helps downstream decisions, but its integration with visual evidence remains composition-dependent. We next ask what information actually controls the final prediction.

4.5 Finding III: Trajectory Role Shortcut

Trajectory Role Shortcut. *Cross-scale decisions can collapse onto model-specific trajectory roles. In the tested Qwen3.7-Flash episodes, trajectory role becomes a behavioral control variable that can bypass the intended spatial dependency.*

We study 77 fixed, evidence-complete T-W-T episodes to test whether predictions follow the intended spatial result or simpler trajectory role cues. In these tasks, the initial-room identity changes the agent's trajectory role but not the correct downstream answer, enabling a controlled intervention on trajectory role dependence.

Trajectory roles emerge as alternative decision rules. Original Direct predictions concentrate on simple trajectory roles, although the dominant role differs across models. Qwen3.5-27B, Qwen3.7-Flash, and VST select the *unvisited* room on 96.1%, 97.4%, and 98.7% of episodes, respectively, whereas VLM-3R selects the current *hub* on 77.9% of episodes (Figure 4). Notably, the hub cannot be the correct target under the T-W-T task construction. These structured, model-specific preferences suggest that trajectory roles can provide alternative rules to the intended spatial dependency.

Evidence removal preserves Qwen predictions. Under the same intervention setting, removing all RGB or masking all bridge views preserves Qwen's semantic prediction on 100% of episodes (Table 5). Thus, the tested decisions remain unchanged even after the visual evidence is removed.

Counterfactual role swaps reveal behavioral control. We modify only the initial-room identity in the instruction while keeping RGB, spatial facts, direction-to-room mapping, answer options, and the correct answer fixed. Qwen follows the newly unvisited room on 98.7% of paired episodes (Table 5). This counterfactual intervention identifies trajectory role as an alternative behavioral control variable that can bypass the intended workspace-to-transit dependency.

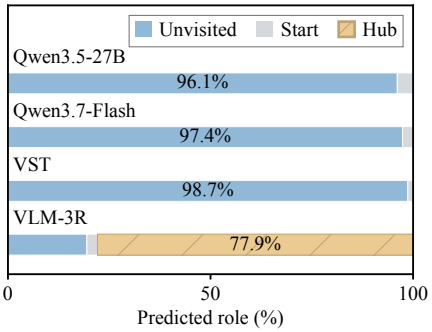


Figure 4: **Trajectory–Role Shortcut signals emerge across models.** Original Direct role distributions on 77 `T-W-T` episodes. Qwen/VST favor the unvisited role; VLM-3R favors the hub.

Shortcut reliance is model-dependent. The strength of this effect varies across models. For Gemini 3.7 Flash, agreement with the Full RGB baseline is 66.2% after removing RGB and 68.8% after masking all bridge views, while role following under the identity-swap intervention reaches 42.9%, substantially below Qwen's 98.7%. This suggests that both the shortcut surrogate and the strength of its behavioral control can vary across models.

A complementary `W-T` binding intervention further indicates fragile dependency use. Reversing the relation-to-action rule reduces Qwen accuracy from 70.0% to 26.3%, while 76.3% of actions remain unchanged. As inverse-rule execution is also weak in text-only controls, we treat this as evidence of fragile dependency use rather than a second clean shortcut.

Table 5: **Qwen exhibits a Trajectory–Role Shortcut.** `T-W-T` evidence-removal and identity-swap controls on 77 episodes. Gray: *Full-RGB agreement*; peach: *unvisited-role following*.

Model	Intervention	Rate (%)
Evidence removal		
Qwen3.7-Flash	No RGB	100.0
Qwen3.7-Flash	All bridges masked	100.0
Gemini 3.7 Flash	No RGB	66.2
Gemini 3.7 Flash	All bridges masked	68.8
Identity swap		
Qwen3.7-Flash	Start-room identity	98.7
Gemini 3.7 Flash	Start-room identity	42.9

Other task structures expose weaker shortcut signals through restricted answer spaces. Across all 6,167 questions and 35 templates, Qwen and Gemini achieve 23.38% and 27.13% text-only accuracy, respectively. `W-T` reduces the terminal decision to left/right, yielding 46.35% and 52.94% text-only accuracy. These structural constraints can narrow the answer space without recovering the intended spatial dependency, but they do not isolate a causal behavioral surrogate as cleanly as the `T-W-T` trajectory role intervention.

5 Discussion and Conclusion

Why cross-scale spatial composition matters. Real-world spatial tasks do not unfold at a single scale. Genuine spatial intelligence requires models to continuously organize local perception into cross-scale spatial states that can support subsequent decisions.

Implications for spatial reasoning systems. On MindTraverse, current models remain limited overall; even on higher-accuracy tasks, we observe clear shortcut reliance, suggesting that existing models do not yet solve cross-scale spatial composition reliably. Our diagnostics further show that explicit spatial representations can improve downstream reasoning, partial structural connections can facilitate spatial state construction, yet visual evidence can still conflict with an existing spatial state. These results point to two directions for future spatial reasoning systems: stronger 3D priors for structured map construction, and training strategies that explicitly suppress potential shortcuts so that downstream decisions depend on the spatial relations required by the task.

Scope and limitations. MindTraverse is currently evaluated in household simulator scenes with fixed joint observations, and our conclusions therefore do not establish direct transfer to real-world images, active embodied exploration, or manipulation. Spatial representation and matched-intervention analyses are primarily conducted on different subsets with Qwen3.7-Flash. Future work should extend these diagnostics to more models and real-world interactive settings.

Ethics Statement

Human evaluation provides a reference baseline for spatial reasoning in simulated household scenes. Multiple participants answered 495 questions from 15 scenes using the same public RGB views, questions, and answer options as the models. Participants subsequently revisited incorrectly answered questions and provided feedback on the causes of their errors. Appendix E describes the human evaluation and the visual and logical answerability review.

Reproducibility Statement

Appendices A–C document benchmark construction and validation, experimental protocols and model settings, scoring and statistical procedures, and aggregate results with chance adjustment. Appendices D–F describe matched controls, dependency interventions, human evaluation, and representation metrics; Appendices H and I provide prompts, intervention specifications, and representative examples. Appendix G documents the experimental records, reconciliation of template-, sequence-, and benchmark-level counts, and offline rescoring of three external runs. We will release the benchmark data, construction, evaluation, and diagnostic code, prompt templates, model settings, per-template results, and reproduction commands, as outlined in Appendix G.

AI Use Statement

We used large language models (LLMs), including OpenAI's ChatGPT, as auxiliary tools to assist with writing, editing, translation, literature retrieval, code development, and paper presentation, as well as to provide feedback on experimental design and result interpretation. All AI-assisted content and suggestions were critically reviewed and verified by the authors, and AI-assisted code was tested for correctness. The authors take full responsibility for all implementations, statements, analyses, and conclusions presented in this paper.

References

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 19107–19117. IEEE, 2022.

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

Bytedance Seed. Seed2.0 model card: Towards intelligence frontier for real-world complexity. *arXiv preprint arXiv:2607.00248*, 2026.

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. IEEE, 2025.

Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Tongxi Zhou, Qingping Sun, Hui En Pang, Jiaqi Li, Oscar Qian, 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7879–7890, June 2026.

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, June 2024.

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 *NeurIPS*, 2024.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. ProcTHOR: Large-scale embodied AI using procedural generation. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

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, June 2026.

Haoquan Fang, Jiafei Duan, Donovan Clay, Sam Wang, Shuo Liu, Weikai Huang, Xiang Fan, Wei-Chuan Tsai, Shirui Chen, Yi Ru Wang, Shanli Xing, Jaemin Cho, Jae Sung Park, Ainaz Eftekhar, Peter Sushko, Karen Farley, Angad Wadhwa, Cole Harrison, Winson Han, Ying-Chun Lee, Eli VanderBilt, Rose Hendrix, Suveen Ellawela, Lucas Ngoo, Joyce Chai, Zhongzheng Ren, Ali Farhadi, Dieter Fox, and Ranjay Krishna. MolmoAct2: Action reasoning models for real-world deployment, 2026. URL https://arxiv.org/abs/2605.02881.

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 Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXIII*, volume 15081 of *Lecture Notes in Computer Science*, pp. 148–166. Springer, 2024.

Caelan Reed Garrett, Rohan Chitnis, Rachel Holladay, Beomjoon Kim, Tom Silver, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. Integrated task and motion planning. *Annual review of control, robotics, and autonomous systems*, 4(1):265–293, 2021.

GLM-5-Team, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, Huilong Chen, Jiajie Zhang, Jian Jiao, Jiaqi Guo, Jingsen Wang, Jingzhao Du, Jinzhu Wu, Kedong Wang, Lei Li, Lin Fan, Lucen Zhong, Mingdao Liu, Mingming Zhao, Pengfan Du, Qian Dong, Rui Lu, Shuang-Li, Shulin Cao, Song Liu, Ting Jiang, Xiaodong Chen, Xiaohan Zhang, Xuancheng Huang, Xuezhen Dong, Yabo Xu, Yao Wei, Yifan An, Yilin Niu, Yitong Zhu, Yuanhao Wen, Yukuo Cen, Yushi Bai, Zhongpei Qiao, Zihan Wang, Zikang Wang, Zilin Zhu, Ziqiang Liu, Zixuan Li, Bojie Wang, Bosi Wen, Can Huang, Changpeng Cai, Chao Yu, Chen Li, Chengwei Hu, Chenhui Zhang, Dan Zhang, Daoyan Lin, Dayong Yang, Di Wang, Ding Ai, Erle Zhu, Fangzhou Yi, Feiyu Chen, Guohong Wen, Hailong Sun, Haisha Zhao, Haiyi Hu, Hanchen Zhang, Hanrui Liu, Hanyu Zhang, Hao Peng, Hao Tai, Haobo Zhang, He Liu, Hongwei Wang, Hongxi Yan, Hongyu Ge, Huan Liu, Huanpeng Chu, Jia'ni Zhao, Jiachen Wang, Jiajing Zhao, Jiamin Ren, Jiapeng Wang, Jiaxin Zhang, Jiayi Gui, Jiayue Zhao, Jijie Li, Jing An, Jing Li, Jingwei Yuan, Jinhua Du, Jinxin Liu, Junkai Zhi, Junwen Duan, Kaiyue Zhou, Kangjian Wei, Ke Wang, Keyun Luo, Laiqiang Zhang, Leigang Sha, Liang Xu, Lindong Wu, Lintao Ding, Lu Chen, Minghao Li, Nianyi Lin, Pan Ta, Qiang Zou, Rongjun Song, Ruiqi Yang, Shangqing Tu, Shangtong Yang, Shaoxiang Wu, Shengyan Zhang, Shijie Li, Shuang Li, Shuyi Fan, Wei Qin, Wei Tian, Weining Zhang, Wenbo Yu, Wenjie Liang, Xiang Kuang, Xiangmeng Cheng, Xiangyang Li, Xiaoquan Yan, Xiaowei Hu, Xiaoying Ling, Xing Fan, Xingye Xia, Xinyuan Zhang, Xinze Zhang, Xirui Pan, Xu Zou, Xunkai Zhang, Yadi Liu, Yandong Wu, Yanfu Li, Yidong Wang, Yifan Zhu, Yijun Tan, Yilin Zhou, Yiming Pan, Ying Zhang, Yinpei Su, Yipeng Geng, Yong Yan, Yonglin Tan, Yuean Bi, Yuhan Shen, Yuhao Yang, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yurong Wu, Yutao Zhang, Yuxi Duan, Yuxuan Zhang, Zezhen Liu, Zhengtao Jiang, Zhenhe Yan, Zheyu Zhang, Zhixiang Wei, Zhuo Chen, Zhuoer Feng, Zijun Yao, Ziwei Chai, Ziyuan Wang, Zuzhou Zhang, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yuxiao Dong, and Jie Tang. Glm-5: from vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763.

Ziyang Gong, Wenhao Li, Xianzheng Ma, Songyuan Li, Zhaokai Wang, Songze Li, Jiayi Ji, Xue Yang, Gen Luo, Junchi Yan, and Rongrong Ji. Space-10: A compre-

hensive benchmark for multimodal large language models in compositional spatial intelligence. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 55900–55932, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/5b84864ff8474fd742c66f219b2eaac1-Paper-Conference.pdf.

Fucai Ke, Zhixi Cai, Boying Li, Long Chen, Beibei Lin, Weiqing Wang, Pari Delir Haghighi, Gholamreza Haffari, and Hamid Rezatofighi. View2space: Studying multi-view visual reasoning from sparse observations. *arXiv preprint arXiv:2603.16506*, 2026.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Ziwei Chai, Y Charles, HS Che, Cheng Chen, et al. Kimi k2.5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An Interactive 3D Environment for Visual AI. *arXiv*, 2017.

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*, 2025a. ISSN 2835-8856. URL https://openreview.net/forum?id=zKv8qULV6n.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *Conference on Robot Learning*, pp. 80–93. PMLR, 2023.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, Weiming Lu, and Yueting Zhuang. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models, 2025b. URL https://arxiv.org/abs/2505.21500.

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*, 2026a.

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 *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=KtrFXlvgrK.

Yun Li, Yiming Zhang, Tao Lin, XiangRui Liu, Wenxiao Cai, Zheng Liu, and Bo Zhao. Sti-bench: Are mllms ready for precise spatial-temporal world understanding? In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5622–5632. IEEE, 2025c.

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, 2025. URL https://arxiv.org/abs/2512.10863.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvila: Efficient frontier visual language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4122–4134. IEEE, 2025.

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, October 2025a.

Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso de Melo, Jianwen Xie, and Alan L. Yuille. Spatialreasoner: Towards explicit and generalizable 3d spatial reasoning. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025b.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17249–17260. IEEE, 2025c.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=IDJx97BC38.

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.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.7: The agent frontier, May 2026b. URL https://qwen.ai/blog?id=qwen3.7.

Navid Rajabi and Jana Kosecka. Gsr-bench: A benchmark for grounded spatial reasoning evaluation via multimodal llms. *arXiv preprint arXiv:2406.13246*, 2024.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10737–10746. IEEE, 2020.

Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. Habitat 2.0: Training home assistants to rearrange their habitat. *Advances in neural information processing systems*, 34:251–266, 2021.

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. Spatial mental modeling from limited views. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=0FhrtdKLtD.

Wenqi Wang, Reuben Tan, Pengyue Zhu, Jianwei Yang, Zhengyuan Yang, Lijuan Wang, Andrey Kolobov, Jianfeng Gao, and Boqing Gong. SITE: towards spatial intelligence thorough evaluation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 9058–9069. IEEE, 2025.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 13569–13597. Curran Associates, Inc., 2025a. doi: 10.52202/085713-0457. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/142b166c8b18b43774c4f0e0c1e9948e-Paper-Conference.pdf.

Haoning Wu, Xiao Huang, Yaohui Chen, Ya Zhang, Yanfeng Wang, and Weidi Xie. Spatialscore: Towards comprehensive evaluation for spatial intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31029–31041, 2026.

Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan

Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025b.

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

Kaiyu Yang, Olga Russakovsky, and Jia Deng. Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2051–2060. IEEE, 2019.

Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, et al. Visual spatial tuning. In *European Conference on Computer Vision*, pp. 192–211. Springer, 2026a.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis L Brown II, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. Cambrian-s: Towards spatial supersensing in video. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=rBFDvZu6pb.

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 C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 157051–157088, 2026c. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/fe5d4bd3e2af823701b5d8f70a3c7602-Paper-Conference.pdf.

Zaibin Zhang, Yuhan Wu, Lianjie Jia, Yifan Wang, Zhongbo Zhang, Yijiang Li, Binghao Ran, Fuxi Zhang, Zhuohan Sun, Zhenfei Yin, et al. Think3d: Thinking with space for spatial reasoning. *arXiv preprint arXiv:2601.13029*, 2026.

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, October 2025.

Rui Zhu, Xin Shen, Shuchen Wu, Chenxi Miao, Xin Yu, Yang Li, Weikang Li, Deguo Xia, and Jizhou Huang. Video-msr: Benchmarking multi-hop spatial reasoning capabilities of mllms. *arXiv preprint arXiv:2601.09430*, 2026.

MindTraverse: Appendix

Appendix guide. Construction and validation: Section A; protocols: Section B; aggregate results: Section C; dependency experiments: Section D; human evaluation: Section E; representation metrics: Section F; reproducibility: Section G; prompts: Section H; examples: Section I.

A Benchmark Specification and Validation

A.1 Spatial Scopes and Public Inputs

MindTraverse tests whether the result of one spatial operation can condition an operation in another context. Workspace (`W`) operations concern local objects; transit (`T`) operations concern orientation and movement between contexts. `W`, `T`, `W-T`, `T-W`, `W-T-W`, and `T-W-T` specify their order. The benchmark contains 6,167 questions in 210 scene instances from 185 source houses. Questions, scene instances, and underlying houses are distinct sampling units.

Standard model inputs are the question, options, and ordered RGB views. Geometry, instance identities, camera poses, and task programs support construction and scoring. Named diagnostic conditions separately supply specified maps, poses, or intermediate states.

A.2 Typed Atomic Interfaces

A relation query returns a direction, localization returns an entity, and transit returns a turn, movement direction, or next-room choice. Executing transit also establishes the arrival room and pose for a later `W` operation.

Table A1: **Six atomic interfaces supply reusable outputs.** Frames are specified by the question.

Interface	Scope	Operation and frame	Output
POO	`W`	Relate objects in the observer frame.	Direction
OO	`W`	Relate a target to an oriented object.	Direction
POR	`W`	Select an object in an observer-frame direction.	Entity
OR	`W`	Select an object in an oriented object's frame.	Entity
Orientation	`T`	Orient toward a goal from the current state.	Turn direction
Route	`T`	Resolve the queried route action or destination.	Action or room

For object centers $C(A), C(B)$ in the ground plane, set $\Delta = C(B) - C(A)$. With unit forward direction $f = (f_x, f_z)$, the relation policy uses

$$s_{\text{left}} = \frac{f_x \Delta_z - f_z \Delta_x}{\|\Delta\|}, \qquad s_{\text{front}} = \frac{f_x \Delta_x + f_z \Delta_z}{\|\Delta\|}. \tag{A1}$$

The signed bearing selects a direction sector. Localization requires a unique eligible object in that sector. An anchor-relative frame uses the anchor's wall normal, which can differ from the camera direction in the image showing it. Object-frame queries require a supported semantic front.

A.3 Result-Conditioned Recursive Composition

Let P_i be the program prefix through stage i, with result z_i and active context c_i. Binding b_i constructs the next context and operands:

$$(c_{i+1}, u_{i+1}) = b_i(c_i, z_i; \eta_i), \qquad z_{i+1} = o_{i+1}(\mathcal{S}, c_{i+1}, u_{i+1}). \tag{A2}$$

Here η_i is the binding rule and frame information, and $\mathcal{S}$ is scene state available to the construction executor. These variables describe the dependency; they are excluded from standard model inputs.

B01 applies a workspace direction to a route in the declared transit frame. B02 maps a selected entity to its semantic goal using the supported object–room association or an explicit public destination rule. B03 executes transit and binds the arrival context to a workspace query. Compatibility requires matching output types, reference frames, candidate sets, and scene roles. Reapplying these bindings yields `W-T-W` and `T-W-T`; the final output can vary within a sequence. `W-T-W` can return a direction or entity, whereas `T-W-T` can return a room or action.

Each recursive extension adds a result-conditioned consumer to the intended dependency chain. Depth counts these successive operations, rather than words in the question. However, deeper recursive

chains retain all dependencies of their prefixes and add further reasoning steps, so task difficulty should increase with depth in principle.

The final `W-T-W` query refers to the room reached through the selected entrance; it does not repeat the resolved room name. Initial locations, candidate-room inventories, and reference frames remain explicit. The implementation instantiates a finite template inventory.

A.4 Grounding Programs to Scenes

Algorithm A1 expands the main-text schematic. In its composition loop, r_k is the typed symbolic output of the current prefix P; grounding and geometric execution determine its value.

Algorithm A1 Complete construction of a grounded spatial task

Require: target sequence S, operator–binding template τ, scene inventory G
Ensure: public RGB views V, question q, options O, answer label y, or rejection

$o_1 \leftarrow \textsc{InstantiateAtomic}(S_1, \tau)$; $P \leftarrow [o_1]$
for $k = 1, \ldots, |S| - 1$ **do**
  $r_k \leftarrow \textsc{TypedSymbolicOutput}(P)$
  $b_k \leftarrow \textsc{Binding}(\tau, k)$; $o_{k+1} \leftarrow \textsc{InstantiateAtomic}(S_{k+1}, \tau)$
  if $b_k =$ B01: direction $\rightarrow$ route **then**
    Resolve the transit reference pose
    Map r_k to a compatible entrance or route state in that pose
  else if $b_k =$ B02: entity $\rightarrow$ semantic goal **then**
    Bind the entity role to r_k
    Bind the semantic goal using the entity's destination rule
  else if $b_k =$ B03: transit $\rightarrow$ workspace **then**
    Bind the arrival-room role to the transit result r_k
    Select a compatible workspace, anchor, and reference frame
  end if
  Check type compatibility and candidate-role availability in G
  if either check fails **then**
    return rejection
  end if
  Append o_{k+1} to P with the resolved binding
end for
$P_G \leftarrow \textsc{GroundRoles}(P, G)$
$(z_1, \ldots, z_{|S|}, y) \leftarrow \textsc{ExecutePrivateGeometry}(P_G)$
Assemble the candidate RGB views V, question q, and options O
Check uniqueness, geometric margins, and reference-frame consistency
Check required bridge and room-marker visual evidence in V
Check question–option alignment, image order, and absence of duplicate public tasks
if any check fails **then**
  return rejection
else
  return export of (V, q, O, y)
end if

Scene preparation supplies rooms, entrances, workspaces, objects, orientation anchors, and rendered RGB. The compiler consumes their geometry and projection records, assigns entities to template roles, computes intermediate states and answers, and verbalizes the resulting task. Missing evidence rejects a candidate. C033/C035 additionally specify an imagined transport and placement; the executor evaluates that hypothetical operation without providing a photograph of the imagined final arrangement.

A.5 Geometric and Referential Checks

Geometric checks prevent a well-formed sentence from being assigned an unstable or ambiguous label. Relation queries reject coincident objects and bearings too near a sector boundary. Four-action route turns use the nearest canonical action with a margin from a competing action. A route turn is computed from the incoming and outgoing headings; it is not the horizontal pixel position of an entrance. Route variants can use a different terminal movement vocabulary, so the four-action rule is not a universal option-count assumption.

Referential checks operate on public distinctions. Two private instance IDs do not make two visually indistinguishable objects unambiguous. The compiler checks same-scope competing instances and

registered confusable categories; explicit, distinct appearance descriptors are required when such objects are both named. Object-frame queries require a supported orientation cue, and anchor-relative queries require a usable anchor in the correct room. These conditions concern whether a reference can be grounded, separately from whether a model will ground it correctly.

Table A2: **Geometry and RGB checks constrain admissible questions.** Marker and overlap thresholds are stricter than object thresholds; they do not measure human recognizability.

Applicable check	Quantity	Threshold
Workspace relation	Minimum distance to sector boundary	7°
Four-action route turn	Minimum distance to midpoint boundary	15°
Object / anchor	Minimum bounding-box side	12 px
Object / anchor	Minimum visible mask area	256 pixels
Marker / overlap witness	Minimum bounding-box side	24 px
Marker / overlap witness	Minimum visible mask area	512 pixels
Marker / overlap witness	Minimum box area / image area	0.25%
RGB checks	Normalized center in each image axis	[0.05, 0.95]

A.6 Visual Evidence Checks

Visual evidence is assessed at the object, association, and task levels. At the object level, the renderer records projected bounding boxes, visible instance-mask area, image position, and a framing flag. The public-object check requires a legible, fully framed object with a safe image-center location and sufficient projected size. Table A2 lists the audited contract parameters. These are operational visibility checks, not calibrated probabilities of human recognition.

At the association level, seeing the two endpoints separately is insufficient. For an anchor-relative query, an image containing the anchor must connect to the workspace evidence through a direct co-view or an ordered overlap chain with shared recognizable non-structural objects. For applicable destination queries, the validator similarly requires a room marker and workspace in one view, or a marker–landmark view linked to a landmark–workspace view. It checks the actual projections in the question's public image subset and the relevant room context. Private room membership alone cannot supply this visual link. The `T-W-T` hub task has a dedicated evidence contract rather than reusing every destination-workspace check unchanged.

Named room markers use a stricter size test than ordinary query objects and must be fully framed. A static overlap witness uses the same stricter size and center thresholds but may be partially framed, since its role is to connect two views rather than identify the named room marker. Additional context or connecting photographs can establish a missing association during scene preparation. The question compiler only accepts evidence present in its input observations; a possible line of sight in simulator geometry does not count as an observed image.

At the task level, each question from the first 60 scenes is manually reviewed using the public images, question, and options (Section E). Reviewers check both the sufficiency of the visual evidence and the logical steps needed to derive the answer, including the reference frames and associations required to carry intermediate results between stages.

A.7 Realized Templates and Public Export

The 35 templates in Table A3 vary the atomic operation, reference frame, and intermediate result passed between stages.

Export checks require distinct nonempty options, a valid answer index, unique images, and consistent image order. Deduplication uses finalized question text, options and answer, ordered images, and public context within the task-type/sequence group. Labels support scoring; private scene programs, poses, branch IDs, and evidence-role annotations are excluded from standard inputs. Selected illustrated examples appear in Section I.

A.8 Scene Coverage

The 210 scene instances come from 185 source houses. Table A5 summarizes their layout, observations, and reuse. Room-graph edges are eligible doors connecting two recorded rooms; for a disconnected graph, maximum distance means the largest finite shortest-path distance within a component. These graph distances describe simulator topology, while visual connectivity is checked

Table A3: **Operator–binding patterns define 35 realized templates.** Stable template IDs; N: question count; chance is option-count adjusted.

ID	Sequence	Operators	Bindings	N	Chance (%)
C001	W	POO	—	976	25.0
C002	W	OO	—	103	25.0
C003	W	POR	—	312	41.6
C004	W	OR	—	10	36.7
C005	T	Orientation	—	393	25.0
C006	T	Route	—	463	21.9
C007	W-T	POO→Route	B01	527	25.0
C008	W-T	OO→Route	B01	34	25.0
C013	T-W	Orientation→POO	B03	1188	25.0
C014	T-W	Orientation→OO	B03	155	25.0
C015	T-W	Orientation→POR	B03	322	41.3
C016	T-W	Orientation→OR	B03	23	40.6
C017	T-W	Route→POO	B03	58	25.0
C018	T-W	Route→OO	B03	12	25.0
C019	T-W	Route→POR	B03	15	40.0
C021	W-T-W	POO→Route→POO	B01, B03	778	25.0
C022	W-T-W	POO→Route→OO	B01, B03	102	25.0
C023	W-T-W	POO→Route→POR	B01, B03	268	42.4
C024	W-T-W	POO→Route→OR	B01, B03	6	33.3
C025	W-T-W	OO→Route→POO	B01, B03	58	25.0
C026	W-T-W	OO→Route→OO	B01, B03	13	25.0
C027	W-T-W	OO→Route→POR	B01, B03	21	42.9
C028	W-T-W	OO→Route→OR	B01, B03	1	33.3
C033	W-T-W	POR→Route→POO	B02, B03	42	25.0
C035	W-T-W	POR→Route→POR	B02, B03	34	36.8
C045	T-W-T	Orientation→POO→Route	B03, B01	159	33.3
C046	T-W-T	Orientation→OO→Route	B03, B01	17	33.3
C047	T-W-T	Route→POO→Route	B03, B01	7	33.3
C048	T-W-T	Route→OO→Route	B03, B01	3	33.3
C049	T-W-T	Orientation→POR→Orientation	B03, B02	37	25.0
C050	T-W-T	Orientation→POR→Route	B03, B02	22	25.0
C051	T-W-T	Orientation→OR→Orientation	B03, B02	2	25.0
C052	T-W-T	Orientation→OR→Route	B03, B02	2	25.0
C053	T-W-T	Route→POR→Orientation	B03, B02	2	25.0
C054	T-W-T	Route→POR→Route	B03, B02	2	25.0

Table A4: **The benchmark covers six sequences and 35 templates.** Full-6167; chance averages uniform choice over each question's options.

Sequence	Questions	Scenes	Templates	Chance (%)
W	1,401	189	4	28.77
T	856	206	2	23.33
W-T	561	183	2	25.00
T-W	1,773	173	7	28.29
W-T-W	1,323	113	10	29.16
T-W-T	253	84	10	31.13

on each question's images. Scene-ledger object counts include entities outside the task-required set, and recorded scene images need not all be supplied on every question.

Table A5: **Scene structure varies across the benchmark.** 210 scene instances from 185 houses; graph distances use eligible room-connecting doors and do not measure visual connectivity.

Quantity	N	Min	Median	Mean	Max
Rooms	210	3	4.0	5.13	10
Eligible room-connecting doors	210	2	3.0	4.01	9
Objects in the scene ledger	210	98	156.5	179.09	344
Rendered image records	210	14	27.0	26.83	31
Room-graph components	210	1	1.0	1.12	3
Maximum finite room-graph distance	210	2	3.0	3.46	7
Questions per scene	210	4	25.0	29.37	88
Scene variants per source house	185	1	1.0	1.14	3

B Experimental Protocol and Statistical Setup

B.1 Scoring and Model Settings

Accuracy is the number of correct final options divided by the complete question count. Each run uses its fixed answer normalizer. Uniform chance averages $1/K_q$ over questions; `T-W-T` role-balanced accuracy instead averages recall over observed target roles. The main table contains 20 configurations, none trained on MindTraverse in this evaluation. CoT and cognitive mapping (CM) are additional Qwen3.7-Flash prompts (Table A8).

Hosted models use temperature 0 and a 512-token primary output limit. Seed, GPT, and Kimi use no reasoning mode. Gemini and GLM use their provider's lowest reasoning setting; Qwen3.5-27B and Qwen3.7-Flash disable thinking. The archived Qwen3.7-Flash Direct run uses the spatial prompt-control instructions and a 1,024-token limit. Other public Direct runs use the answer-only instructions in Section H.1. Hosted runs retain the original view order and resolution, using PNG or quality-95 JPEG.

Qwen3-VL-2B/4B/8B, NVILA-8B, and LLaVA-OneVision-Qwen2-7B use BF16, deterministic decoding, thinking disabled, and a 64-token limit. Qwen3-VL uses 12,544–100,352 pixels per view; NVILA and LLaVA preserve aspect ratio with maximum edges of 448 and 384 pixels. VST and SpaceR use NF4 with BF16 compute and 12,544–100,352 pixels per view. Spatial-MLLM uses BF16. SenseNova-SI uses BF16 with up to four patches per image and sixteen in total; VLM-3R uses bilinear spatial pooling at stride two and grid newline placement. These settings are not matched for compute or precision.

Externally evaluated models. Each of the following runs uses one H100 PCIe GPU, one question per call, all public frames in order, greedy decoding, and a 64-token output limit.

Table A6: **Inference settings for three external runs.** TF: Transformers; PT: PyTorch. All use greedy decoding with a 64-token output limit.

Model	Checkpoint	Precision / libraries
Molmo2-ER	allenai/Molmo2-ER	BF16 TF 4.56.2 PT 2.7.1+cu126
Qwen3.5-4B	Qwen/Qwen3.5-4B	BF16 TF 5.16.1 PT 2.7.1+cu126
GeoThinker-7B	lihy285/GeoThinker	NF4; BF16 compute TF 4.57.0 PT 2.5.1+cu124

Molmo2-ER uses its 378×378 processor setting and 14-pixel patches. Qwen3.5-4B uses 16-pixel patches, merge size two, and disabled thinking. GeoThinker uses NF4 double quantization with BF16 compute; its Qwen-VL branch uses 12,544–451,584 pixels, while its geometry branch uses the 518-pixel path, an eight-view setting, and spatial downsampling by two. Repetition penalties are 1.0, 1.0, and 1.05, respectively. All three use the Primary Direct user instruction without an explicit system message; GeoThinker's native chat template supplies its default system instruction.

B.2 TINY600 MAP AND CAMERA-STATE PROTOCOL

Tiny600 contains 100 questions per sequence across 210 scenes. It retains 60 initial diagnostic items and adds 540 selected with a fixed seed using composition, scene coverage, and answer-position balance, without model outputs. Its balanced averages differ from Full-6167's natural-frequency averages.

PredMap is a task-conditioned JSON map generated from the ordered RGB and question. The two conditions share a flat schema and 24-node budget; +Pose also supplies exact photo position, yaw, facing, and horizon. Map generation uses 8,192 output tokens, and downstream reasoning uses 4,096. GT Map replaces predicted state with simulator entities and derived anchor frames, without using the answer or private task program to construct the input. Adding RGB to the GT Map+Pose condition changes the input-modality sentence and appends the original views; the map, pose, question, options, and reasoning rules stay fixed.

Table A7: **GT maps improve overall accuracy; RGB effects vary.** Tiny600 accuracy (%) under shared rules. GT Map and pose are privileged; the no-Pose GT-map condition is post hoc.

Condition	W	T	W-T	T-W	W-T-W	T-W-T	Overall
RGB → PredMap	42.0	25.0	86.0	37.0	27.0	38.0	42.5
RGB + Pose → PredMap	31.0	42.0	53.0	34.0	28.0	32.0	36.7
GT Map + Pose, no RGB	52.0	39.0	60.0	59.0	58.0	56.0	54.0
RGB + GT Map + Pose	40.0	27.0	83.0	52.0	30.0	53.0	47.5
RGB + GT Map, no Pose	44.0	24.0	81.0	39.0	27.0	49.0	44.0

GT Map+Pose improves over pose-conditioned PredMap by 17.3 pp (scene-clustered 95% CI $[12.3, 22.4]$). Adding RGB changes Overall by -6.5 pp ($[-11.2, -1.7]$); over the 400 cross-scale items, the change is -3.8 pp ($[-9.7, 2.3]$). The fifth, post-hoc condition removes supplied camera state while retaining GT Map and RGB; its 44.0% Overall score is separate from the primary map controls. Adding Pose to that condition changes Overall by $+3.5$ pp (95% CI $[-1.0, 8.0]$); the cross-scale effect is $+5.5$ pp ($[0.2, 10.8]$, exact McNemar $p = 0.0527$). The RGB decrease is supported overall; its cross-scale-only interval includes zero. These interventions characterize state–evidence integration under the shared protocol, without isolating coordinate conversion as its cause.

B.3 STATISTICAL UNITS AND UNCERTAINTY

Diagnostic binomial intervals use Wilson's method. Tiny600 paired effects use 10,000 scene-clustered bootstrap draws; the correlations in Section F.3 use 2,000. Every question from a sampled scene is retained. The 24-item interface suite has one item per scene and uses 20,000 paired draws. The full T-W-T category audit uses 5,000 scene-clustered draws; one-scene strata remain descriptive without cluster intervals. Paired binary outcomes additionally use exact two-sided McNemar tests where reported. Intervention pairs stay together during resampling. Repeated conditions on the same episodes are not independent replications, and each interval refers to its own diagnostic set.

B.4 FIXED-BUDGET PROMPT CONTROLS

These Qwen3.7-Flash controls retain the original prompts and output budgets and precede the Tiny600 Map→Reason protocol.

Table A8: **Prompt elaboration alone does not improve RGB accuracy.** Qwen3.7-Flash, Full-6167, fixed output budget; GT-map conditions use privileged input.

Condition	Correct	N	Accuracy (%)
RGB Direct	2050	6167	33.24
RGB CoT	2004	6167	32.50
RGB CM	1961	6167	31.80
GT Map	1897	6167	30.76
GT Map + Pose	1828	6167	29.64
GT Map + Pose + CoT	2093	6167	33.94

Relative to Direct, CoT changes accuracy by -0.75 pp (95% CI $[-1.90, 0.47]$), and cognitive mapping by -1.44 pp ($[-2.55, -0.30]$). GT Map+Pose+CoT improves over GT Map+Pose by 4.30 pp ($[2.18, 6.48]$). The comparisons characterize these specific prompts.

They do not isolate the effect of a uniform increase in reasoning budget or of adding reference-frame rules alone.

C Aggregate Results and Chance Adjustment

The main table reports all 20 configurations on all six sequences. Every run covers 6,167 questions, and correct counts reconcile across the 35 templates. Offline rescoring reproduces the three externally evaluated runs. Overall accuracy weights questions equally; an unweighted mean across sequences answers a different question because their sample counts differ. For example, Gemini's question-weighted and sequence-macro accuracies are 50.92% and 55.74%; Qwen's are 33.24% and 40.05%. The complete per-template counts and scores for all configurations will accompany the benchmark as machine-readable tables. Template sample sizes remain visible in Table A3.

C.1 Chance-Adjusted Accuracy

For sequence s, K_q options on question q, and accuracy A_s, define

$$b_s = \frac{1}{N_s} \sum_{q \in s} \frac{1}{K_q}, \qquad \widetilde{A}_s = \frac{A_s - b_s}{1 - b_s}. \tag{A3}$$

Zero denotes uniform random choice; negative values denote lower accuracy. The normalization accounts for option counts, while wording and answer-category priors are examined separately in Section D.6.

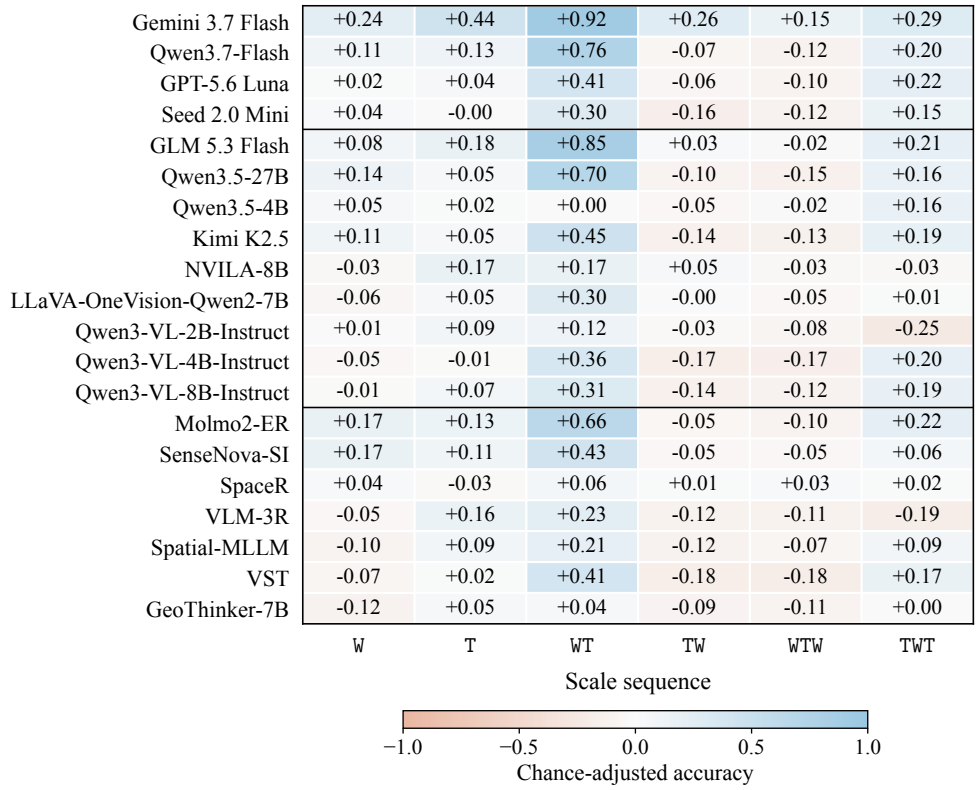


Figure A1: **Sequence differences persist after chance adjustment.** All 20 configurations; blue and peach indicate values above and below uniform chance. Values use Equation A3.

D Does Performance Reflect Cross-Scale Dependency?

Matched controls distinguish state formation, use of supplied state, and sensitivity to variables that should control the answer. Each suite retains its own question set and denominator.

D.1 MATCHED W-T-W STATE FORMATION AND ORACLE CLOSURE

The suite contains 37 distinct houses selected without model predictions. Source and candidate-target relations are at least 12° from a sector boundary; the threshold was set before model calls after an initial 15° threshold yielded fewer than the required 30 houses. Target rooms have distinct semantic labels and relations. Correct-room enumeration is balanced 19/18. Source, transit, and correct-target probes use subsets of Full W-T-W observations; the other-target probe uses additional views of the alternative branch from the same scene.

Table A9: **Supplied-state W-T-W controls.** 37 matched episodes; descriptive Wilson 95% CIs.

Condition	Correct	%	95% CI
Source workspace	7/37	18.92	[9.5, 34.2]
Transit given source direction	35/37	94.59	[82.3, 98.5]
Correct target workspace	3/37	8.11	[2.8, 21.3]
Other target workspace	2/37	5.41	[1.5, 17.7]
Full W-T-W joint	4/37	10.81	[4.3, 24.7]
All required atoms	1/37	2.70	[0.5, 13.8]
Supplied-state joint	33/37	89.19	[75.3, 95.7]
Room-oracle lookup	37/37	100.00	[90.6, 100.0]

The conditional composition test required at least 15 episodes with all required atoms correct. Only 1/37 qualified, triggering the supplied-state control. It supplies the source relation and the target relation for each candidate room alongside transit RGB. The task is to select the destination and retrieve its supplied relation; joint accuracy reaches 33/37. A text-only control supplies the correct room and tests lookup, reaching 37/37. These results show that supplied workspace relations support downstream room selection and relation retrieval. Joint success requires both the destination room and final relation to be correct. The other-branch atomic query scores 2/37, and no episode has all four atomic queries correct.

D.2 W-T RULE AND INTERFACE PROTOCOLS

The 80-scene suite includes 56 observer-frame and 24 object-frame relation-to-route items. Original, paraphrased, and opposite-side rules share the RGB, source question, options, and option order. Only the binding instruction changes.

Table A10: **Reversing the W-T rule reduces accuracy.** Qwen3.7-Flash; 80 matched scenes.

Rule	Correct	Accuracy (%)
Same-side	56/80	70.00
Paraphrase	56/80	70.00
Opposite-side	21/80	26.25

Original and paraphrased answers agree on 63/80 items, with 48/80 jointly correct; original and opposite-side rules have 7/80 jointly correct pairs. Separate explicit-mapping controls reach 58/80 and 39/80 with RGB, and 78/80 and 34/80 without RGB, under same/opposite rules. Thus ordinary inverse-rule execution also contributes to the difference. With the correct source relation supplied alongside RGB, the explicit same/opposite rules score 73/80 and 32/80. These are separate prompts from the original-rule comparison.

A separate 24-scene binary interface suite uses transparent names or neutral Gate A/B labels. Accuracy is 22/24, 19/24, and 24/24 for transparent, neutral, and neutral with an oracle source relation, respectively (Wilson 95% CIs $[74.2, 97.7]$, $[59.5, 90.8]$, and $[86.2, 100.0]\%$). The oracle supplies the required relation while the public rule determines which gate to choose. Selection uses a fixed seed without model answers; LEFT/RIGHT each occurs twelve times, with six identity and six swapped mappings within each relation. All conditions use two options and a fresh transparent baseline. Transparent-minus-neutral accuracy is $+12.5$ pp (paired 95% CI $[-4.2, 29.2]$, McNemar $p = 0.375$); oracle-minus-neutral is $+20.8$ pp ($[4.2, 37.5]$, $p = 0.0625$). The transparent-interface effect remains uncertain; the oracle verifies lookup on these 24 items.

D.3 T-W-T State Correctness and Reuse

The state suite retains 77 of 80 C045 scenes based on evidence completeness. State contains a workspace relation and direction-to-room mapping; Same/Other rules are applied afterward. Targeted views use private evidence roles for oracle routing; random subsets match their image count.

Table A11: **Joint T-W-T state recovery remains rare.** 77 episodes; correct workspace relations, transit mappings at the GT workspace direction, and both. Targeted views use privileged selection.

State condition	N	Workspace	Transit	Both
Full, rule-isolated	77	7	11	1
Targeted, rule-isolated	77	10	15	1
Random, rule-isolated	77	6	9	1
Full, Same-entangled	77	7	8	0
Full, Other-entangled	77	6	11	1

Table A12: **Paired T-W-T accuracy remains low after state isolation.** 77 pairs; PairAcc requires both rules correct. Program execution uses predicted state only.

State / executor	N	Same	Other	Both	PairAcc (%)
Entangled / Qwen	77	44	28	13	16.88
Isolated / Qwen	77	40	34	16	20.78
Entangled / program	77	6	7	4	5.19
Isolated / program	77	8	8	8	10.39

Full and targeted routing each produce only 1/77 jointly correct states. State isolation changes Qwen PairAcc from 13/77 to 16/77 (paired 95% CI $[-6.49, 14.29]$ pp), and program PairAcc from 4/77 to 8/77 ($[-6.49, 12.99]$ pp). These intervals include zero. The deterministic executor uses predicted state and cannot restore missing geometry.

Of 17 executable isolated states, Qwen follows supplied state on 17/17 Same and 7/17 Other queries. Correct GT state yields 154/154 correct program answers, validating the executor interface. Reusing a frozen state guarantees invariance by construction; state correctness and paired answer correctness remain separate measurements. Changing only the final rule preserves the entangled workspace relation on 69/77 pairs and the full direction-to-room mapping on 44/77. Targeted-image marginal-effect intervals include zero, so this suite does not establish reliable gains from oracle view routing.

D.4 T-W-T Evidence and Role Interventions

The evidence suite freezes 77 episodes, one per scene, selected for public evidence completeness without model answers. Bridge masking preserves image slots, order, and resolution, replacing selected views with neutral images. No RGB supplies only the original instruction and options. Qwen's Full RGB and No RGB answers both select the unvisited room on every item, scoring 43/77 with 50% balanced accuracy over observed target roles. Rotating text-only options preserves room identity on 76/77 and option index on 0/77 items.

The identity intervention changes only the stated initial-room identity. Hub, workspace geometry, exit mapping, images, options, and correct answer stay fixed. This initial transit is virtual and does not define the workspace query pose. Paired role following requires the old unvisited choice before the swap and the new unvisited choice afterward, using all 77 pairs.

Gemini uses temperature 0, low reasoning, and a 512-token output limit. Its original/swapped identity conditions choose the unvisited room on 53/77 and 54/77 items. In a separate evidence-removal batch, Full/No RGB/masked accuracies are 51/77, 43/77, and 39/77; No RGB chooses unvisited on 77/77. The two batches' Full RGB answers agree on 58/77 (75.3%, Wilson CI $[64.6, 83.6]\%$). This stability reference is across batches, rather than a same-batch repeated-input control. This variability prevents attributing all prediction changes solely to evidence removal.

A geometry-by-rule control moves the workspace object and reverses the rule in a 2×2 design, with original cameras and 12 houses passing geometric and visual checks before model calls. Qwen chooses the same unvisited room in all four cells: each cell scores 6/12, and geometry-flip, rule-flip, and four-cell joint success are all 0/12 (Wilson 95% upper bound 24.2%). In the 77-episode workspace negative control, accuracy changes from 10/77 to 12/77 under bridge masking. Together

Table A13: **Evidence and narrative interventions differ across models.** Evidence rows measure agreement with each model's Full RGB; identity rows measure paired unvisited-role following. Wilson 95% CIs. Gray: evidence removal; peach: identity intervention.

Model	Intervention	Count	Rate (%)	95% CI (%)
Evidence removal: agreement with Full RGB				
Qwen3.7-Flash	No RGB	77/77	100.0	[95.2,100.0]
Qwen3.7-Flash	All bridges masked	77/77	100.0	[95.2,100.0]
Gemini 3.7 Flash	No RGB	51/77	66.2	[55.1,75.8]
Gemini 3.7 Flash	All bridges masked	53/77	68.8	[57.8,78.1]
Identity swap: paired unvisited-role following				
Qwen3.7-Flash	Initial-room identity	76/77	98.7	[93.0,99.8]
Gemini 3.7 Flash	Initial-room identity	33/77	42.9	[32.4,54.0]

these controls identify narrative-role dependence in the tested Qwen episodes; they do not specify an internal switch to a fallback policy. The composite-minus-workspace drop is $+2.6$ pp (paired 95% CI $[-5.2, 10.4]$), failing the preregistered bridge-dependence criterion. Exact sample agreement alone does not establish population equivalence.

Additional evidence conditions. Removing bridge views, or masking source bridges, destination bridges, or room signatures separately, preserves 76/77 Full RGB answers in each condition. Masking every image gives 39/77 correct and 67/77 agreement with Full RGB. Keeping only workspace views gives 42/77 correct and 76/77 agreement; keeping only destination evidence gives 39/77 and 73/77. Post-hoc anti-novelty reminders yield 35/77 text-only and 31/77 full-RGB accuracy. These conditions measure sensitivity to evidence and instructions; they are not improved benchmark methods. Their exact input changes are specified in Section H.4.

D.5 Full-Family T-W-T Answer-Category Audit

The audit covers all 253 T-W-T questions, ten templates, and 84 scenes using original Direct predictions from all 20 configurations. The 186 B01 items return rooms; 67 B02 items return actions. Room roles are derived from the program and shuffle-aligned option identities: start, hub, or unvisited. Action labels retain left, straight, right, and U-turn. These annotations are used for scoring, not supplied to models.

Table A14: **T-W-T uses room and action answer interfaces.** All 253 questions; target counts and retrospective majority-category accuracy (%) from observed labels.

Template	N	Scenes	Target counts	Majority (%)
Room targets: start / hub / unvisited				
C045	159	80	79 / 0 / 80	50.3
C046	17	13	9 / 0 / 8	52.9
C047	7	5	5 / 0 / 2	71.4
C048	3	2	1 / 0 / 2	66.7
Action targets: left / straight / right / U-turn				
C049	37	24	11 / 9 / 13 / 4	35.1
C050	22	13	8 / 4 / 6 / 4	36.4
C051	2	2	0 / 0 / 2 / 0	100.0
C052	2	2	0 / 0 / 2 / 0	100.0
C053	2	1	1 / 0 / 0 / 1	50.0
C054	2	1	1 / 0 / 0 / 1	50.0

Hub is never a correct room target. Qwen chooses unvisited on 184/186 items (98.9%; scene-clustered 95% CI $[97.2, 100.0]\%$), answering 1/94 start-target and 91/92 unvisited-target questions correctly. Its accuracy is 49.5% and balanced accuracy is 50.0% (95% CI $[48.5, 51.6]\%$). VLM-3R and Qwen3-VL-2B-Instruct instead select hub on 150/186 and 142/186 items. On the 67 action questions, Qwen scores 22/67 (32.8%) and never selects straight, although it is the target on 13/67. Its accuracy difference from the majority-action reference is -1.5 pp (95% CI $[-23.7, 19.0]$). These distributions describe broader behavior; they do not establish the same causal mechanism across all templates or

Table A15: **Room choices often favor the unvisited role.** All 186 B01 questions; accuracy and choice shares (%). B.Acc.: mean recall over observed targets; H: choice entropy (bits); majority: 50.5%. Blue: Qwen3.7-Flash; gray: target distribution.

Model	Acc. ↑	B.Acc. ↑	Start	Hub	Unvisited	H
Target distribution	—	—	50.5	0.0	49.5	—
Gemini 3.7 Flash	55.4	55.7	23.7	0.0	74.7	0.80
Qwen3.7-Flash	49.5	50.0	1.1	0.0	98.9	0.09
GPT-5.6 Luna	51.6	52.1	2.2	0.0	97.8	0.15
Seed 2.0 Mini	48.9	49.5	0.5	0.0	99.5	0.05
GLM 5.3 Flash	48.9	49.2	18.8	0.5	79.0	0.75
Qwen3.5-27B	47.8	48.3	3.8	0.0	96.2	0.23
Qwen3.5-4B	48.4	48.8	10.2	4.3	85.5	0.72
Kimi K2.5	47.8	48.3	7.0	0.0	93.0	0.37
NVILA-8B	30.6	31.0	1.1	37.1	61.8	1.03
LLaVA-OneVision-Qwen2-7B	33.3	33.4	23.1	31.7	45.2	1.53
Qwen3-VL-2B-Instruct	9.7	9.7	8.1	76.3	15.6	1.01
Qwen3-VL-4B-Instruct	49.5	50.0	0.0	0.0	100.0	0.00
Qwen3-VL-8B-Instruct	49.5	49.9	7.5	0.0	92.5	0.39
Molmo2-ER	50.5	51.0	4.8	2.2	93.0	0.43
SenseNova-SI	38.7	38.8	25.8	23.7	50.5	1.49
SpaceR	33.3	33.5	19.4	28.0	52.2	1.46
VLM-3R	12.9	13.0	2.2	80.6	17.2	0.81
Spatial-MLLM	38.7	39.0	5.9	28.5	65.6	1.16
VST	50.5	51.1	1.1	0.0	98.9	0.09
GeoThinker-7B	30.6	30.9	2.2	36.0	61.8	1.08

Table A16: **Action preferences vary across models.** All 67 B02 questions; accuracy and choice shares (%). B.Acc.: mean recall over observed targets; majority: 34.3%. Blue: Qwen3.7-Flash; gray: target distribution.

Model	Acc. ↑	B.Acc. ↑	Left	Straight	Right	U-turn
Target distribution	—	—	31.3	19.4	34.3	14.9
Gemini 3.7 Flash	40.3	37.9	34.3	6.0	26.9	28.4
Qwen3.7-Flash	32.8	31.8	43.3	0.0	41.8	14.9
GPT-5.6 Luna	31.3	26.4	40.3	1.5	52.2	6.0
Seed 2.0 Mini	20.9	18.1	31.3	14.9	29.9	23.9
GLM 5.3 Flash	35.8	37.4	37.3	16.4	20.9	22.4
Qwen3.5-27B	26.9	20.9	59.7	6.0	28.4	6.0
Qwen3.5-4B	25.4	21.5	25.4	17.9	32.8	23.9
Kimi K2.5	34.3	27.6	26.9	4.5	52.2	16.4
NVILA-8B	23.9	19.0	25.4	17.9	35.8	20.9
LLaVA-OneVision-Qwen2-7B	28.4	23.0	38.8	0.0	43.3	17.9
Qwen3-VL-2B-Instruct	26.9	23.6	31.3	22.4	22.4	23.9
Qwen3-VL-4B-Instruct	31.3	23.3	19.4	1.5	64.2	14.9
Qwen3-VL-8B-Instruct	28.4	21.7	41.8	7.5	43.3	7.5
Molmo2-ER	34.3	28.4	31.3	20.9	47.8	0.0
SenseNova-SI	26.9	20.8	50.7	3.0	40.3	6.0
SpaceR	29.9	27.6	50.7	10.4	32.8	6.0
VLM-3R	32.8	27.9	28.4	9.0	62.7	0.0
Spatial-MLLM	32.8	25.6	35.8	7.5	55.2	1.5
VST	22.4	17.1	43.3	1.5	44.8	10.4
GeoThinker-7B	32.8	32.8	14.9	50.7	32.8	1.5

models. Choice shares use the full subset denominator. Balanced accuracy averages exact-answer recall over observed target categories; the majority-category reference uses observed labels. Room and action interfaces are analyzed separately. The 77-episode causal suite is a subset of C045, whereas the Text-only audit below covers all templates.

D.6 TEXT-ONLY AUDIT ACROSS ALL TEMPLATES

Qwen3.7-Flash and Gemini 3.7 Flash receive only question text and options for all 6,167 questions, 210 scenes, and 35 templates. Request records confirm zero image parts. Both use identical instructions (Section H.5) and temperature 0; Qwen disables thinking with a 1,024-token limit, while Gemini uses low reasoning and 512 tokens. Every question remains in the accuracy denominator.

Table A17: **Text-only accuracy remains high on `W-T` and `T-W-T`.** All 6,167 questions; accuracy (%). Gray: answer-prior references.

Model / reference	`W`	`T`	`W-T`	`T-W`	`W-T-W`	`T-W-T`	Overall
Qwen3.7-Flash	19.77	20.91	46.35	19.51	20.94	40.71	23.38
Gemini 3.7 Flash	25.41	28.04	52.94	21.21	21.69	46.25	27.13
Uniform choice	28.77	23.33	25.00	28.29	29.16	31.13	27.71
Majority index	27.27	24.53	27.09	27.86	29.78	32.81	27.16

Uniform choice averages $1/K_q$ over questions. Majority index uses the most frequent correct position within each column and is a retrospective reference.

Qwen scores 1,442/6,167 and Gemini 1,673/6,167 (Wilson 95% CIs $[22.34, 24.46]\%$ and $[26.03, 28.25]\%$). They agree on 2,617/6,167 answers, including all 159 C045 and 17 C046 items. The sequence scores are reported independently against the uniform-choice and majority-position references. Their agreement detects shared text-driven choices; the identity intervention separately tests causal role dependence.

Table A18: **Text-only behavior varies across all 35 templates.** Model columns: correct counts (accuracy, %); agreement: identical Qwen/Gemini choices.

Template	Seq.	N	Uniform (%)	Index (%)	Qwen	Gemini	Agreement
C001	`W`	976	25.00	27.97	130 (13.32)	214 (21.93)	394/976
C002	`W`	103	25.00	29.13	9 (8.74)	15 (14.56)	30/103
C003	`W`	312	41.56	44.55	137 (43.91)	125 (40.06)	180/312
C004	`W`	10	36.67	50.00	1 (10.00)	2 (20.00)	5/10
C005	`T`	393	25.00	26.21	71 (18.07)	120 (30.53)	105/393
C006	`T`	463	21.90	24.62	108 (23.33)	120 (25.92)	170/463
C007	`W-T`	527	25.00	26.76	247 (46.87)	283 (53.70)	177/527
C008	`W-T`	34	25.00	35.29	13 (38.24)	14 (41.18)	10/34
C013	`T-W`	1188	25.00	26.18	154 (12.96)	195 (16.41)	535/1188
C014	`T-W`	155	25.00	26.45	30 (19.35)	27 (17.42)	58/155
C015	`T-W`	322	41.30	38.51	132 (40.99)	123 (38.20)	149/322
C016	`T-W`	23	40.58	47.83	10 (43.48)	11 (47.83)	10/23
C017	`T-W`	58	25.00	31.03	9 (15.52)	10 (17.24)	31/58
C018	`T-W`	12	25.00	50.00	3 (25.00)	3 (25.00)	5/12
C019	`T-W`	15	40.00	40.00	8 (53.33)	7 (46.67)	12/15
C021	`W-T-W`	778	25.00	29.69	96 (12.34)	106 (13.62)	315/778
C022	`W-T-W`	102	25.00	34.31	15 (14.71)	17 (16.67)	53/102
C023	`W-T-W`	268	42.41	42.91	113 (42.16)	110 (41.04)	105/268
C024	`W-T-W`	6	33.33	50.00	4 (66.67)	1 (16.67)	1/6
C025	`W-T-W`	58	25.00	34.48	5 (8.62)	6 (10.34)	18/58
C026	`W-T-W`	13	25.00	46.15	0 (0.00)	3 (23.08)	5/13
C027	`W-T-W`	21	42.86	52.38	6 (28.57)	12 (57.14)	4/21
C028	`W-T-W`	1	33.33	100.00	1 (100.00)	0 (0.00)	0/1
C033	`W-T-W`	42	25.00	35.71	10 (23.81)	5 (11.90)	11/42
C035	`W-T-W`	34	36.76	41.18	27 (79.41)	27 (79.41)	22/34
C045	`T-W-T`	159	33.33	35.85	80 (50.31)	80 (50.31)	159/159
C046	`T-W-T`	17	33.33	47.06	8 (47.06)	8 (47.06)	17/17
C047	`T-W-T`	7	33.33	57.14	2 (28.57)	2 (28.57)	7/7
C048	`T-W-T`	3	33.33	33.33	2 (66.67)	2 (66.67)	3/3
C049	`T-W-T`	37	25.00	32.43	5 (13.51)	14 (37.84)	16/37
C050	`T-W-T`	22	25.00	40.91	4 (18.18)	8 (36.36)	7/22
C051	`T-W-T`	2	25.00	50.00	0 (0.00)	1 (50.00)	0/2
C052	`T-W-T`	2	25.00	50.00	0 (0.00)	0 (0.00)	2/2
C053	`T-W-T`	2	25.00	50.00	1 (50.00)	0 (0.00)	0/2
C054	`T-W-T`	2	25.00	100.00	1 (50.00)	2 (100.00)	1/2

Uniform: per-question random-choice expectation. Index: retrospective majority position. Small-N templates are descriptive.

The full template table retains sparse strata. Uniform chance averages $1/K_q$; majority index uses observed labels. Agreement compares choices independently of correctness. Text-only scores alone do not show that the other templates are free of shortcuts.

Restricted action targets in `W-T`. All 561 `W-T` items offer four actions, but their targets are 326 left and 235 right turns, with no straight or U-turn targets. Uniform choice between left/right gives 50%; always choosing left gives 58.11%. These are retrospective target-distribution references. Qwen and Gemini agree on only 187/561 choices, unlike C045/C046. The restricted answer space permits shortcuts without establishing the same unvisited-room mechanism as `T-W-T`.

E Human Evaluation and Answerability Review

Each question from the first 60 scenes underwent a manual answerability review. Reviewers used the same public inputs available to the evaluated models: the question, answer options, and ordered RGB views. They inspected each question from visual and logical perspectives to verify that the available evidence supports the reasoning needed to derive an answer.

E.1 Visual Evidence Review

For each question, reviewers checked whether the referenced entities and relevant spatial cues could be identified in the public images. The review also considered whether the views supported the required reference frames and cross-view associations. These checks follow the visual-evidence criteria described in Section A.6.

E.2 Logical Solvability Review

Reviewers followed the operations and dependencies stated in each question to check that its answer could be derived from the available evidence. This included checking that intermediate results could be carried into the subsequent spatial context under the stated task rules. Geometric validity and reference-frame conventions provide the construction basis for this review (Section A.5).

The questions in this 60-scene review passed the visual and logical answerability checks. This question-quality review is distinct from testing whether a model actually uses the intended cross-scale dependency, which is assessed through the matched controls in Section D.

E.3 Human Baseline

The human baseline in Table 2 aggregates responses from multiple participants on a subset of 495 questions from 15 scenes. Participants received the same public inputs as the models: ordered RGB views, the question, and answer options. Human accuracy is 89.49% (443/495 correct).

The subset covers all six scale sequences and 31 templates. Table A19 reports accuracy and question counts by sequence.

Table A19: **Humans achieve 89.49% overall accuracy.** Pooled responses on a 495-question subset from 15 scenes.

	`W`	`T`	`W-T`	`T-W`	`W-T-W`	`T-W-T`	**Overall**
Questions	73	64	26	228	69	35	495
Correct	68	53	22	209	65	26	443
Accuracy (%) ↑	93.15	82.81	84.62	91.67	94.20	74.29	89.49

E.4 Participant Review of Errors

Participants revisited incorrectly answered questions and reported that the errors arose from their own reasoning. Their feedback did not identify incorrect ground-truth answers or insufficient visual evidence in the questions they revisited. This post-hoc participant feedback complements the answerability review; it is not a separate independent audit.

F Representation Metrics and Error Analysis

We analyze errors at three observable interfaces: retaining the required entities, associating them with the right room context, and computing a relation in the requested frame. The metrics below

inspect representations or controlled answers; they do not expose a model's internal computation. The task-required inventory is preferable to comparing predicted map size with every simulator object, many of which are irrelevant to a question.

F.1 Entity Inventory and Matching

The choice-complete set includes entities on the executed program and the physical candidates needed to resolve its decisions. The smaller core set contains only the correct chain and its operands. Entities are deduplicated by physical identity. Rooms, doors, workspaces, orientation anchors, query objects, physical alternatives, room-signature objects, named route waypoints, and transported objects contribute to the count. Camera pose, coordinates, directions, relations, actions, and images do not. These attributes matter to spatial state but are not additional physical entities.

The extraction audit covers 6,167 unique questions, with no missing referenced GT-map entities or structural-rule violations. Table A20 reports the Full-6167 inventory. The largest choice-complete set contains 17 entities, below the tested 24-node limit. This does not imply that every question receives enough correctly selected or geometrically accurate nodes.

Table A20: **Cross-scale tasks require heterogeneous entity sets.** Full-6167; core counts follow the correct chain. Entity-type counts are means.

(a) Choice-complete and core inventories

Seq.	Questions	Mean	Median	Range	Mean core
W	1,401	4.35	4	4–6	4.00
T	856	4.60	5	4–7	4.60
W-T	561	8.00	8	8–8	6.00
T-W	1,773	8.31	8	8–10	8.00
W-T-W	1,323	13.43	13	13–17	11.00
T-W-T	253	11.40	11	10–13	9.09

(b) Physical entity types

Seq.	Rooms	Doors	Workspaces	Anchors	Objects
W	1.00	0.00	1.00	0.00	2.35
T	2.00	1.00	0.59	0.00	1.01
W-T	3.00	2.00	1.00	0.00	2.00
T-W	2.00	1.00	1.00	1.00	3.31
W-T-W	3.00	2.00	2.00	1.00	5.43
T-W-T	3.00	2.00	1.00	1.00	4.40

Predicted IDs are freely named, so matching does not equate them with canonical IDs. Maximum one-to-one bipartite matching uses role-aware type and category compatibility. Rooms require a compatible room category; doors require a door type; workspaces admit compatible support-surface types; anchors and objects require compatible normalized categories. Room-grounded matching additionally requires association with the correct semantic room category; a door needs at least one compatible endpoint. Neither metric verifies unique instance identity, both door endpoints, coordinates, or orientation. A category visible in the question can also yield a semantic match without visual recognition.

Let $M_k(q)$ be the match count under criterion k. We use

$$N_{\text{req}}(q) = |E_{\text{req}}(q)|, \quad R_k(q) = \frac{M_k(q)}{N_{\text{req}}(q)}, \quad P_k(q) = \frac{M_k(q)}{\max(N_{\text{pred}}(q), 1)}, \quad k \in \{\text{sem}, \text{room}\}. \tag{A4}$$

The precision and recall summaries in Tables A21 and A22 use per-question ratios. An empty map has zero matches and zero precision. Table A21 also reports the mean per-question node-to-requirement ratio; it is not the ratio of the two displayed means.

F.2 Structural Grounding and Co-grounding

Figure 3(b) pools entity occurrences across the 500 T-bearing questions, giving 801 required doors and 559 workspaces per condition. Each recall is the total matched count divided by the corresponding required count. A deterministic matching first maximizes semantic cardinality and then grounded

Table A21: **More nodes do not ensure higher task-relevant precision.** Tiny600 means; P_{sem} and P_{room}: semantic and room-grounded precision (%, ↑).

PredMap	N_{req}	Nodes	Nodes/Req.	P_{sem}	P_{room}
NoPose	8.37	13.56	1.85	55.4	42.5
+Pose	8.37	11.72	1.60	62.4	45.7

pairs; both totals agree with the original independent matchings. Pure `W` is excluded because its maps often omit room nodes and parent links.

Question-level co-grounding requires all needed doors and all needed workspaces to satisfy room-grounded matching in the same map. It tests semantic room associations, rather than unique room instances, both door endpoints, or a continuous path. There are 455 applicable questions: 100 in each cross-scale sequence and 55 in `T`. Figure 3(c) displays only the four cross-scale sequences. In the `T` subset, RGB PredMaps ground all doors on 27.3% of questions, all workspaces on 83.6%, and both on 27.3%. Across all 455, these rates are 55.2%, 85.7%, and 53.2%, respectively. The corresponding joint rate for `W-T-W` is 37.0%.

F.3 Coverage Associations and Their Scope

Table A22 retains all analyzed slices, including the pooled result. Pure `W` maps often omit room nodes or parent links; their low room-grounded recall changes the pooled association's sign. The negative size–recall relationship is observed after excluding `W` and within native cross-scale tasks. It remains an association across heterogeneous families, not an intervention on context size or a causal scaling law.

Recall–accuracy correlations are weak in the cross-scale and `W-T-W/T-W-T` subsets, with 95% intervals that include zero. In the non-`W` NoPose subset, semantic recall has a small negative correlation with accuracy ($r = -0.088$, 95% CI $[-0.177, -0.003]$).

Table A22: **Coverage associations depend on the evaluation slice.** Tiny600; ρ: Spearman correlation; r: point-biserial correlation. Scene-clustered 95% CIs.

(a) Required-state size versus recall

Map	Slice	n	Scenes	ρ_{sem}	95% CI	ρ_{room}	95% CI
NoPose	All	600	210	$+0.079$	$[-0.002, +0.167]$	$+0.212$	$[+0.131, +0.294]$
NoPose	Excluding `W`	500	210	-0.296	$[-0.393, -0.197]$	-0.179	$[-0.275, -0.091]$
NoPose	Cross-scale	400	209	-0.146	$[-0.235, -0.055]$	-0.182	$[-0.270, -0.089]$
NoPose	`W-T-W/T-W-T`	200	142	-0.194	$[-0.318, -0.068]$	-0.240	$[-0.364, -0.117]$
+Pose	All	600	210	-0.023	$[-0.101, +0.052]$	$+0.133$	$[+0.061, +0.205]$
+Pose	Excluding `W`	500	210	-0.384	$[-0.460, -0.309]$	-0.311	$[-0.383, -0.239]$
+Pose	Cross-scale	400	209	-0.162	$[-0.260, -0.064]$	-0.212	$[-0.300, -0.115]$
+Pose	`W-T-W/T-W-T`	200	142	-0.347	$[-0.469, -0.222]$	-0.393	$[-0.512, -0.259]$

(b) Recall versus answer correctness

Map	Slice	n	Scenes	r_{sem}	95% CI	r_{room}	95% CI
NoPose	All	600	210	-0.063	$[-0.133, +0.017]$	$+0.025$	$[-0.050, +0.100]$
NoPose	Excluding `W`	500	210	-0.088	$[-0.177, -0.003]$	$+0.034$	$[-0.047, +0.115]$
NoPose	Cross-scale	400	209	-0.030	$[-0.137, +0.068]$	$+0.060$	$[-0.037, +0.156]$
NoPose	`W-T-W/T-W-T`	200	142	$+0.100$	$[-0.035, +0.206]$	$+0.099$	$[-0.034, +0.219]$
+Pose	All	600	210	$+0.062$	$[-0.021, +0.142]$	$+0.064$	$[-0.017, +0.143]$
+Pose	Excluding `W`	500	210	$+0.052$	$[-0.035, +0.142]$	$+0.042$	$[-0.043, +0.133]$
+Pose	Cross-scale	400	209	$+0.025$	$[-0.064, +0.115]$	$+0.031$	$[-0.067, +0.124]$
+Pose	`W-T-W/T-W-T`	200	142	-0.015	$[-0.144, +0.111]$	-0.075	$[-0.213, +0.060]$

F.4 Axis-Level Localization

The fixed 55-item workspace subset separates lateral and depth judgments from a joint direction answer (Table A23). Independent lateral accuracy is 44/55, whereas independent depth accuracy is 15/55; deriving axes from the joint output gives 43/55 and 15/55. The empirical majority baseline for depth is 26/55, so the depth result should not be presented as strong evidence of successful geometric inference.

Table A23: **Depth judgments trail lateral judgments.** 55 questions; independent versus joint-derived axes. Macro recall averages target categories.

Axis condition	Correct	Accuracy (%)	Macro recall (%)
Lateral, independent	44/55	80.0	63.8
Lateral, joint-derived	43/55	78.2	59.7
Depth, independent	15/55	27.3	29.3
Depth, joint-derived	15/55	27.3	34.9

This diagnostic localizes an atomic workspace difficulty under the recorded view and frame conditions. It does not establish that depth errors arise only during a scale transition, nor that depth alone explains `W-T-W` failures. The cross-scale argument instead relies on testing whether an upstream state is transferred into the correct downstream context, as in Section D. A distinct matched reference-frame-prompt intervention would be needed to attribute an improvement to explicit coordinate-conversion instructions.

G Reproducibility

Each experimental record identifies the model configuration, prompt, input condition, question set, output budget, prediction, and scoring rule. Template-level correct and evaluated counts reconcile with sequence and Overall totals for all 20 configurations. Offline rescoring of the retained Molmo2-ER, Qwen3.5-4B, and GeoThinker-7B outputs reproduces all seven reported scores per model, covering 6,167 unique questions with no missing, extra, or duplicate IDs. This verification uses the original outputs without new inference. Their model and processor settings are listed in Section B.1.

We will release the benchmark data and construction, evaluation, and diagnostic code, together with complete prompt templates, model settings, per-template results, and reproduction commands. Run records preserve the association between each result and its actual evaluated inputs. Simulator geometry, poses, and task programs support construction, scoring, and diagnostic-input generation. They are excluded from standard model inputs; named privileged conditions supply only the specified maps, poses, or states. This input distinction does not restrict access to the research code. Statistical units and uncertainty calculations appear in Section B.

H Prompts and Intervention Specifications

Instruction boxes reproduce the archived prompt text. Bracketed fields denote per-question inputs; shared blocks are printed once. Model-specific message wrappers and complete auxiliary templates accompany the code.

H.1 Primary Evaluation and Prompt Controls

The primary evaluation uses the following answer-only instructions.

Primary Direct / system

You are an answer-only benchmark model. Output only the final selected option JSON.

Primary Direct / user

You are answering a multi-view spatial intelligence benchmark. Act as the answer-only model and use only the question, options, and public RGB images.
Do not access, infer, or request any BEV map, coordinates, private GT, question-generation program, or hidden metadata.
Choose the option that best matches the question. Output exactly one JSON object and no explanation, reasoning, Markdown, or other text:
{"answer_index": 0, "answer_text": "exact option text"}
answer_index is a zero-based option index, and answer_text must copy the selected option verbatim.

Question: [PUBLIC QUESTION]
Options:
[LETTERED PUBLIC OPTIONS]

Qwen Direct controls. The Qwen Direct run and its matched interventions use the spatial-control system instruction: only the ordered RGB, question, and options are available, and references to the first photo are literal. Its user protocol is:

Diagnostic Direct / user

There are [N] public RGB images in the user message, in the supplied order.

Reasoning protocol:
1. Read the question and options carefully.
2. Inspect the relevant image(s), preserving the benchmark's stated viewpoint/anchor and frame order.
3. Decide internally. Do not show a rationale or a map.
4. [FINAL OUTPUT CONTRACT]

Question:
[PUBLIC QUESTION]

Options:
[LETTERED PUBLIC OPTIONS]

Here [FINAL OUTPUT CONTRACT] requires a final JSON object with a zero-based `answer_index` and verbatim `answer_text`, as above. CoT adds short spatial checks before that final object; CM first requests a 10-by-10 inferred cognitive map. These prompts precede Map→Reason.

H.2 TINY600 MAP CONSTRUCTION AND SHARED REASONING

The Map→Reason conditions use shared reasoning instructions. The two PredMap conditions share RGB, the flat schema, and a 24-node construction budget; only supplied camera state differs. The map-stage limit is 8,192 tokens and the reasoning limit is 4,096. GT-map conditions skip map generation and pass the supplied map to the same reasoner (Section B.2). Each attached RGB view is preceded by "[IMAGE_ID] [public image ID]"; the user text follows the images.

Map construction / system

You are the MAP CONSTRUCTION stage of a controlled multi-view spatial-reasoning experiment. You receive ordered public RGB images from one static indoor scene and a question with options. Depending on the declared condition, exact camera poses are either supplied or deliberately withheld.

Build a compact, parseable, task-conditioned spatial map. Camera poses are authoritative and must be copied exactly. Object positions, room membership, topology, cross-view identity, workspace membership, and anchor geometry must be inferred from visible RGB evidence. Never answer the multiple-choice question. Never output an option letter/index or a final directional relation copied from an option. Unknown facts must be null or listed in unresolved; do not hallucinate them.

Only an intrinsically directional object with clear visual/semantic evidence may have intrinsic_facing_xz. An ordinary table, desk, counter, or workspace does not acquire an intrinsic facing merely because it supports objects. For a wall anchor, anchor_observer_facing_xz is the direction in which a virtual observer standing in front of the anchor faces the wall. It is distinct from both the anchor asset's front and the oblique camera direction of an RGB image that happens to show the anchor.

Use compact/minified JSON. Keep at most 24 task-relevant entity nodes. For each entity, cite only its single best evidence image unless a second view is needed for identity. Keep every unresolved record under 16 words.

Output exactly one JSON object and no surrounding prose or markdown.

Map construction / flat representation and schema

Use one shared X/Z coordinate frame for every room, door, workspace, anchor, object, and camera. In the authoritative-pose condition it is the AI2-THOR world metric frame. In the no-pose condition it is the single inferred relative gauge declared below. Do not create a workspace-local coordinate frame. Preserve continuous camera facing; do not quantize it to four cardinal directions.

Required schema:
{
"schema_version": "cross_scale_flat_metric_map.1",
"coordinate_frame": {
"id": "ai2thor_world_xz_v1|inferred_relative_xz_v1",

Map construction / flat representation and schema (continued)

"unit": "meter|relative_unit",
"yaw_zero": "+Z",
"facing_from_yaw": "[sin(yaw), cos(yaw)]"
},
"views": [
{"image_id": "image_01", "position_xz_m": [x,z],
"yaw_deg": number, "facing_xz": [fx,fz], "horizon_deg": number}
],
"nodes": [
{"id": "unique_id", "kind": "room|door|workspace|anchor|object",
"category": "name", "position_xz_m": [x,z] or null,
"parent_id": "id" or null, "intrinsic_facing_xz": [fx,fz] or null,
"anchor_observer_facing_xz": [fx,fz] or null,
"footprint_half_extents_xz_m": [hx,hz] or null,
"evidence": "image_id" or null}
],
"edges": [
{"kind": "contains|connects|same_instance", "source": "id",
"target": "id", "evidence": "image_id" or null}
],
"unresolved": ["short uncertainty record"]
}

Map only task-relevant entities and the evidence required to connect them. Merge the same physical entity across views, while keeping distinct instances separate. Do not add left_of/right_of/front_of/behind or a reasoning chain.

Map construction / user with supplied pose

Build the spatial map for this Cross-scale benchmark sample.

[MAP CONSTRUCTION RULE]
[FLAT REPRESENTATION AND SCHEMA]

Map only the entities, viewpoints, containment, connectivity, identity, and geometry needed to solve the supplied question. Preserve every spatial scale that the question actually connects, but do not solve any relation or action during map construction.

[CAMERA CONDITION: AUTHORITATIVE POSE]
Copy every supplied camera pose exactly; never re-estimate it.
[AUTHORITATIVE PHOTO POSES]
[PHOTO POSES JSON]
[/AUTHORITATIVE PHOTO POSES]

Question:
[PUBLIC QUESTION]

Options:
[LETTERED PUBLIC OPTIONS]

Remember: output the map JSON only. Do not answer the question.

The no-pose condition replaces the complete authoritative-pose block above with the following. It supplies no camera coordinates, room IDs, or private image roles.

Map construction / no-pose replacement

[CAMERA CONDITION: NO POSE]
No camera position, yaw, facing, room_id, or private image role is supplied.
Infer a mutually consistent relative camera layout only from the ordered RGB.
Fix image_01 as the arbitrary gauge origin [0,0] facing [0,1]; this gauge
choice is not evidence about the physical scene. Infer all other view poses
relative to it. Use coordinate_frame.id=inferred_relative_xz_v1 and record
uncertain geometry in unresolved rather than pretending it is metric GT.
[/CAMERA CONDITION]

Frozen-map reasoning / shared system

You are the REASONING stage of a controlled cross-scale spatial experiment. You receive one frozen spatial map, a public question, and its options, but no RGB. Use only the frozen map and question. Never repair the map, invent missing geometry, or infer an answer from option order.

Follow [REASONING PROTOCOL]. Keep the procedure general:

1. For each stage of a multi-stage question, first identify the active spatial scale, pose/reference frame, reference entity, subject, and requested output. Preserve the stated A/B order and preserve each intermediate result until the next stage consumes it.
2. A photo-defined Person frame uses that image's stored pose. A request to face a wall anchor uses the virtual pose facing straight toward that anchor, perpendicular to its wall; an oblique evidence-photo yaw does not define this frame. An Object frame is used only when the question explicitly asks from that object's own perspective and the map supplies a reliable intrinsic facing. Ordinary tables/workspaces do not define a facing.
3. For B relative to A, use delta=position(B)-position(A). With the active unit facing, left=cross2d(facing,delta) and front=dot(facing,delta), where cross2d([fx,fz],[dx,dz])=fx*dz-fz*dx. Positive left means left, negative means right; positive front means front, negative means behind. Combine both signs for diagonal directions and select the closest direction represented by the options. For localization, evaluate all candidate objects and select the unique one satisfying the requested relation.
4. For movement, start from the explicitly specified pose, process targets in textual order, and update position and facing after each movement before the next decision. A room target is approached through its real connected door, not its geometric center. Keep literal straight segments as given. Distinguish a turn action from a final endpoint direction by the wording and options.
5. Across scales, do not skip the binding step. A Local direction used as an entrance instruction keeps its left/right side; a localized object used as a semantic goal must first be identified and then matched to its ordinary room use; a Global destination selects the concrete room, workspace, and anchor frame for the following Local query. When a direction selects a room, match it to the physically aligned outgoing door in the active pose rather than using a fixed direction-to-room prior.
6. A requested object placement is a mental state update, not a new image. Place the identified object at the stated relation to the reference object, then solve the final relation/localization from the updated coordinates.

Use one concise reasoning_step for each genuine reasoning stage in the question. Even if the map is insufficient or contradictory, make the best forced choice: answer_index must be zero-based and answer_text must copy one supplied option exactly.

Output exactly one compact JSON object and no surrounding prose or markdown.

Frozen-map reasoning / user and output schema

Reason over this frozen map and answer the question.

[FROZEN SPATIAL MAP]
[FROZEN MAP JSON]
[/FROZEN SPATIAL MAP]

Question:
[PUBLIC QUESTION]

Options:
[LETTERED PUBLIC OPTIONS]

[REQUIRED OUTPUT SCHEMA]

```
{
"map_sufficiency": "sufficient|insufficient|contradictory",
"reasoning_steps": [
{"stage": 1, "scale": "local|global", "map_id": "map id",
"reference_frame": "camera:image_01|anchor:id|object:id|world",
"operation": "Relation|Localization|Orientation|Route",
"derived_state": "state passed to the next stage or final readout"}
],
"answer_index": 0,
"answer_text": "exact option text"
}
```

answer_index is zero-based and answer_text must exactly copy the selected option.

Both PredMap reasoners receive their fixed predicted map and no RGB; inferred camera state remains in the no-pose PredMap. GT Map+Pose without RGB uses the same system text. Adding RGB replaces only the following input sentence fragment. The post-hoc GT Map+RGB condition without supplied pose also replaces the opening of rule 2. These controlled variants do not constitute a matched test of explicit frame rules against an ordinary reasoning prompt.

RGB reasoning / input sentence replacement

and its options, together with ordered public RGB images. Use the frozen map, question, and supplied RGB evidence.

No supplied pose / rule 2 opening replacement

2. No camera position, yaw, facing, or horizon is supplied. For a photo-defined Person frame, infer that camera's position and facing in the map frame from the supplied RGB and visible map entities. A request to

H.3 Supplied-State Reasoning

The `W-T-W` control supplies source and room-keyed target relations alongside the original transit images. Images precede the user message and retain their order. Both output fields must be correct for joint success.

Supplied-state closure / system

Execute an oracle-state Local-to-Global-to-Local closure. Use the supplied source Local state to ground a destination room in the ordered public RGB evidence, then retrieve the target Local state associated with that room. Return exactly one JSON object: {"selected_room": "ROOM", "final_relation": "DIRECTION"}. Do not explain.

Supplied-state closure / user

The Local states below are verified oracle facts; do not infer them again from the RGB.
Source Local direction instruction: [VERIFIED SOURCE RELATION].
Target-state table:
- If the selected room is [ROOM A], its target Local relation is [RELATION A].
- If the selected room is [ROOM B], its target Local relation is [RELATION B].

The first supplied RGB image shows the two possible entrances. The remaining ordered RGB images are the public path and room evidence from the original episode. Apply the source direction instruction to the entrance, determine which concrete destination room is reached, then use that room as the key into the target-state table. The room answer must be exactly one of: [ROOM A], [ROOM B].

The room-oracle control replaces transit grounding with the verified room and requests only the relation stored under that key, without RGB. Atomic workspace probes request one direction in the stated frame; the transit probe supplies the correct source direction and requests the reached room. These conditions and their image subsets are defined in Section D.1.

H.4 Evidence and Trajectory-Role Interventions

Qwen's evidence and role interventions use Diagnostic Direct. Gemini's identity and evidence controls use Primary Direct. Full RGB attaches all public views. Masking replaces the selected views with neutral images of the same resolution, preserving slots and order; removal deletes those views. No RGB attaches no images and supplies no map or intermediate state. Workspace-only and destination-only conditions retain their corresponding views in the original order. Image counts describe the actual input slots.

Evidence masking / image caption

Public image slot [i]/[N] is intentionally masked and contains no scene evidence.

The identity intervention applies the following clause replacement.

Initial-room swap / replaced clause

Before: you are currently in the [ORIGINAL INITIAL ROOM]. After: you are currently in the [NEW INITIAL ROOM].

The new initial identity is the originally unvisited room. The hub, workspace geometry, exit mapping, RGB, options, and correct answer stay fixed; punctuation and case are preserved. The initial transit is virtual and does not define the workspace query pose. The changed identity is thus irrelevant to the retained spatial solution. In the option-rotation control, the first option moves to the end and option labels are assigned anew.

The workspace negative control removes the final room query and asks for the recovered direction. The separate post-hoc reminder appends:

Post-hoc anti-novelty / question suffix

Important: The correct destination may be either the room where you started or the third room not yet visited. Returning to the starting room is allowed. Do not prefer an unvisited room merely because it is new; use the spatial relation and visible entrance mapping.

The geometry-by-rule control independently changes object geometry and the Same/Other rule, preserving cameras and option order. Geometry changes are rerendered without supplying a description of the displacement to the model. The two rule templates are:

Geometry-by-rule / same question

[PUBLIC STEM WITHOUT FINAL ROOM QUERY] After completing the first move, treat the room you have just entered as the hub. Binding rule: among the two destination rooms other than this hub room, choose the concrete room reached through the entrance whose direction best matches the recovered local relation. Which room should you enter next?

Geometry-by-rule / other question

[PUBLIC STEM WITHOUT FINAL ROOM QUERY] After completing the first move, treat the room you have just entered as the hub. Binding rule: among the two destination rooms other than this hub room, first identify the room reached through the entrance whose direction best matches the recovered local relation, then choose the other destination room instead. Which room should you enter next?

`W-T` rule controls. The original and inverse rules preserve images, source question, and options. The inverse condition uses:

`W-T` rule / opposite-side question

[SOURCE WORKSPACE QUERY]. Use the recovered direction only as an intermediate result. For this question, choose the entrance on the opposite lateral side: a left, left-front, or left-back relation selects the right entrance; a right, right-front, or right-back relation selects the left entrance.

The explicit-mapping follow-up prints the corresponding same/opposite mapping table. The binary interface suite replaces direction-revealing entrance names with Gate A/B and supplies their mapping; its oracle condition also supplies the source relation without RGB.

H.5 Full-Benchmark Text-only Input

Both models receive the same text-only instructions and zero image parts. The system instruction specifies:

Text-only / input contract (verbatim excerpt)

You are evaluated on the text-only control of a multi-view spatial-intelligence benchmark. You receive only one multiple-choice question and its options. The public RGB images referenced by the question are intentionally unavailable in this condition.

Text-only / input contract (verbatim excerpt) (continued)

Use only the supplied question text and options. Do not claim to inspect an image, and do not access, request, or invent a BEV map, metric coordinates, private generation program, answer key, or hidden metadata. Select the best option that can be inferred or guessed from the textual input alone.

The user instruction requests an internal decision using text alone and a final option JSON with a zero-based index and verbatim option text. Output budgets and reasoning settings are specified in Section D.6.

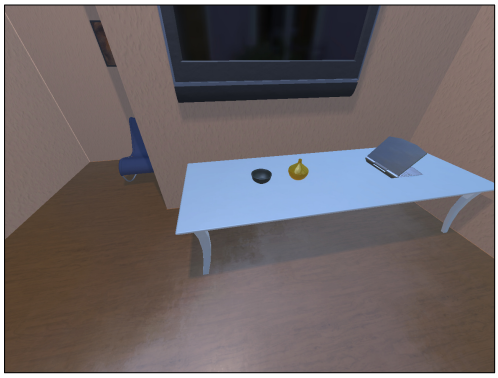
I Representative Illustrated Tasks

Five examples cover `W-T`, `T-W`, `W-T-W`, and the B01/B02 variants of `T-W-T`. They are drawn from the first 20 scenes. Each example includes all public RGB views in their original input order. Questions and options are reproduced from the example records. The full illustrated template catalog will accompany the benchmark.

C007: POO → Route `W-T`

B01; 4 public RGB views (complete input).

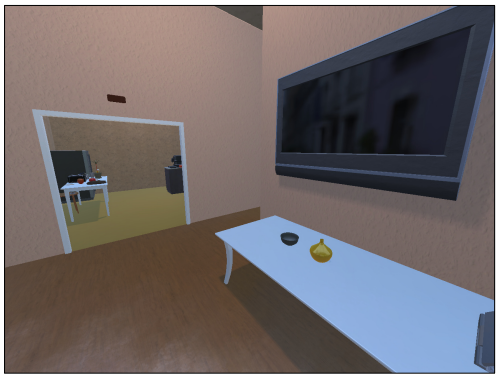
View 1



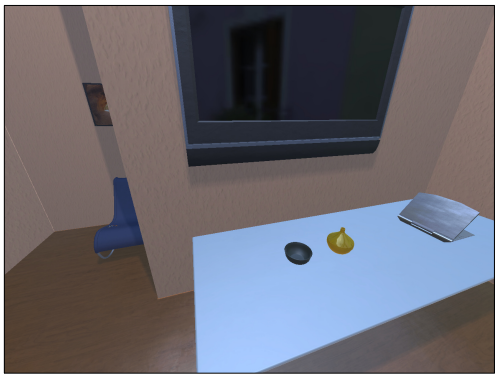
View 2



View 3



View 4



Question. The scene contains a living room, a bedroom, and a kitchen; the current location is the living room. In the first photo, following the viewing direction of that photo, let A be the laptop and B be the bowl; determine the position of B relative to A; use the resulting direction as both the directional instruction and selection criterion to choose the entry direction consistent with this instruction from the two entrances shown in the next photo.

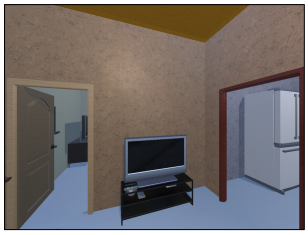
Options: **A.** Turn around **B.** Turn left **C.** Go straight **D.** Turn right

Transit-to-workspace and workspace–transit–workspace. The arrival context sets the frame and operands for the final local query.

C013: Orientation → POO `T-W`

B03; 14 public RGB views (complete input).

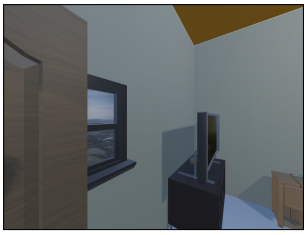
View 1



View 2



View 3



View 4



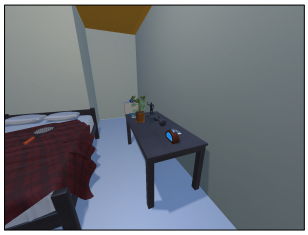
View 5



View 6



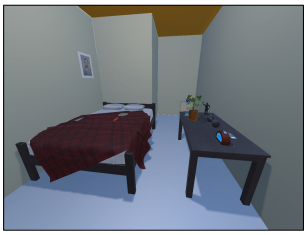
View 7



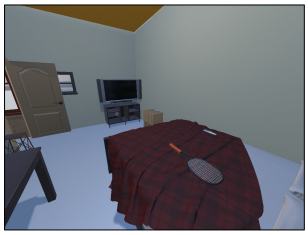
View 8



View 9



View 10



View 11



View 12



View 13



View 14



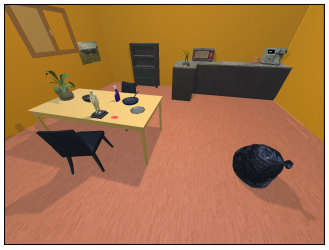
Question. The scene contains a living room, a bedroom, and a kitchen; you are currently in the living room. Go to the bedroom; upon arrival, face the painting on the wall. On the workbench in the bedroom, with this orientation, let A be the dumbbell and B be the statue; determine which direction B is relative to A?

Options: **A.** front-right **B.** right **C.** left **D.** front-left

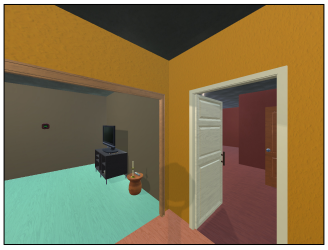
C021: POO → Route → POO `W-T-W`

B01, B03; 8 public RGB views (complete input).

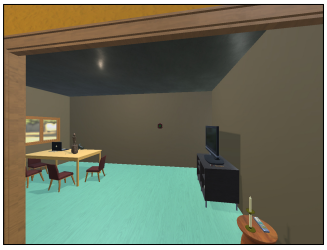
View 1



View 2



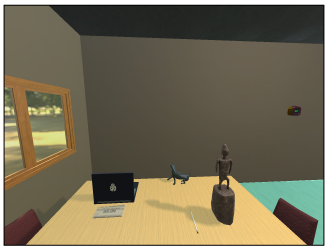
View 3



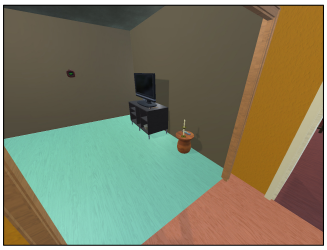
View 4



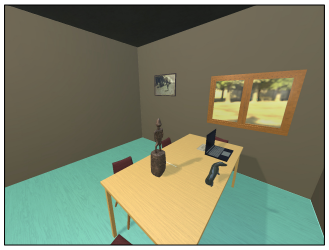
View 5



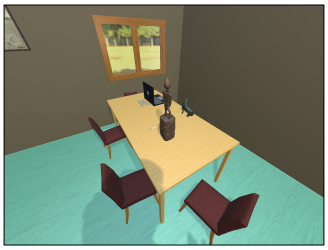
View 6



View 7



View 8



Question. The scene contains a kitchen, a bedroom, and a living room; you are currently in the kitchen. In the first photo, facing the viewing direction, let A be the plate and B be the spray bottle; determine the position of B relative to A. Use this resulting direction as both the directional instruction and the basis for selecting the target area. From the two entrances shown in the next photo, choose the entry direction consistent with this instruction. Upon arrival, face the painting on the wall. In the target workspace, maintaining this orientation, let A be the newspaper and B be the laptop; in which direction is B located relative to A?

Options: **A.** front-left **B.** front-right **C.** right **D.** left

Transit–workspace–transit. The intermediate workspace output conditions either a room choice (B01) or an entity's semantic destination (B02).

C045: Orientation → POO → Route `T-W-T`

B03, B01; 2 public RGB views (complete input).

View 1



View 2



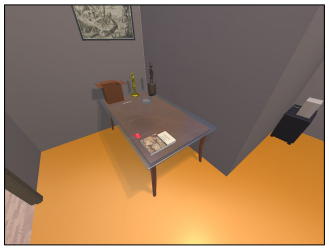
Question. The scene contains a bedroom, a living room, and a kitchen; you are currently in the bedroom. First, enter the kitchen. Upon arrival, face the painting on the wall. On the countertop, with this orientation, let A be the potato and B be the bowl; determine the position of B relative to A. Based on the direction you obtained in the local scene, which room should you enter next?

Options: **A.** Living room **B.** Bedroom **C.** Kitchen

C049: Orientation → POR → Orientation `T-W-T`

B03, B02; 3 public RGB views (complete input).

View 1



View 2



View 3



Question. The scene contains a bedroom, kitchen, and living room; you are currently in the bedroom. First, enter the living room. Upon arrival, face the painting on the wall. On the workbench, in this direction, what desktop object is located to the right-rear of the box? After identifying this object, which room should it belong to in daily life? Then, from your current position, go to that room; which direction should you turn?

Options: **A.** Turn around **B.** Turn left **C.** Go straight **D.** Turn right