SpatialAct: Probing Spatial Reasoning-to-Action Capabilities of VLM Agents in 3D Scenes

Tianhui Liu[1] **Jie Feng**[2†] **Zhiheng Zheng**[3] **Shengyuan Wang**[3]
Yiming Guo[2] **Yanxin Xi**[4] **Hangyu Fan**[3] **Yong Li**[3†] **Pan Hui**[1,4†]

[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]Zhongguancun Academy
[3]Tsinghua University
[4]Helsinki University

Abstract

Humans can effortlessly perceive spatial layouts, form cognitive representations, reason about spatial relations, and translate such reasoning into actions in everyday 3D environments. Although recent vision-language models (VLMs) have shown promising performance on observation-conditioned spatial perception and reasoning tasks, it remains unclear whether they can build coherent spatial understanding, act upon it, and refine their actions through multi-turn feedback. To study this problem, we introduce **SpatialAct**, a simulator-grounded benchmark for probing *action-conditioned spatial reasoning* in 3D scenes. Starting from the most challenging setting, Multi-turn Interactive Refinement, we further design its decomposed counterpart, Single-step Error Detection and Fix, together with five fundamental spatial ability tasks to diagnose the underlying causes of model failures. Experiments reveal a clear reasoning-to-action gap: current VLMs can perform well on isolated spatial reasoning tasks, but struggle to maintain coherent spatial beliefs and produce reliable actions during multi-turn feedback, substantially underperforming humans. These results suggest that current VLM agents still lack robust spatial state tracking under action-induced environment changes, even when low-level control is abstracted away.

1 Introduction

Vision-language models are increasingly moving from passive perception toward agentic interaction in 3D environments and the real world [14, 25, 39, 36]. This shift is driven by emerging applications such as indoor scene understanding and editing, embodied manipulation, 3D world generation, and interactive simulation [1, 3, 5, 20, 28, 33, 36, 40]. In these settings, spatial intelligence is no longer limited to recognizing objects or answering questions about spatial relations. A spatially capable agent must build and maintain a coherent understanding of the environment, decide how to intervene, observe how the environment changes after its own action, and continue reasoning over the updated state. This ability is fundamental for bridging visual-spatial understanding with real-world and simulator-grounded agency.

Recent spatial reasoning benchmarks have made substantial progress in evaluating the visual-spatial capabilities of VLMs. Existing benchmarks examine spatial understanding from different perspectives, including 2D spatial relations, classical spatial cognition, 3D spatial reasoning, and psychometric-inspired basic spatial abilities [8, 12, 23, 26, 27, 31, 38]. Other efforts further extend spatial evaluation

[†]Corresponding author, email: fengj12ee@hotmail.com, liyong07@tsinghua.edu.cn, panhui@ust.hk.

Preprint.

Table 1: Comparison of spatial reasoning and embodied benchmarks.

Benchmark	Dimension	Modality	Scenario	QA Size	Eval Type	Multi-view	Multi-turn	Env. Feed.	Objective
SpatialEval [23]	2D	Image	Mix	4.6k	MCQ	✗	✗	✗	Understanding
CoreCognition [12]	2D	Mix	Mix	1.5k	MCQ	✗	✗	✗	Understanding
BSA [31]	Mix	Image	Classic	312	MCQ	✗	✗	✗	Understanding
Spatial457 [27]	Mix	Image	Outdoor	23.7k	Mix	✗	✗	✗	Understanding
VIS-Bench [32]	3D	Video	Indoor	5k	Mix	✓	✗	✗	Understanding
MINDCUBE [34]	3D	Image	Mix	21.1k	MCQ	✓	✗	✗	Understanding
CityCube [30]	3D	Image	Outdoor	5.0k	MCQ	✓	✗	✗	Understanding
OpenEQA [14]	3D	Video	Indoor	1.6k	Open-ended	✓	✗	✗	Exploration
THEORY OF SPACE [37]	3D	Image	Indoor	2.7k	Open-ended	✓	✓	✗	Exploration
MetaVQA [25]	3D	Image	Outdoor	9.7k	MCQ	✗	✓	✓	Action
SpatialAct	3D	Image	Mix	4.3k	Mix	✓	✓	✓	Action

to richer and more realistic settings, such as all-scale spatial reasoning from millimeters to kilometers [21], 3D scenes and videos [32], limited-view spatial mental modeling [34], cross-view urban reasoning [30], embodied question answering [14, 39], and active spatial belief construction [37]. These benchmarks have revealed important limitations of current VLMs in perceiving, remembering, and reasoning about space.

Nevertheless, existing evaluations still leave an important gap. As summarized in Table 1, most spatial reasoning benchmarks evaluate models as observers. The model is given an image, a video, or a set of views, and is asked to answer questions about the observed scene. Even when the input becomes multi-view, egocentric, or temporally extended, the model output usually does not directly alter the environment that the model must subsequently reason about. In contrast, embodied benchmarks involve actions and feedback, but their evaluation often entangles high-level spatial reasoning with low-level control, navigation, or manipulation [25, 28, 36]. Therefore, a missing middle ground remains between passive spatial question answering and full embodied control: evaluating whether VLM agents can perform high-level semantic spatial actions, observe the resulting state transitions, and maintain consistent reasoning across multiple rounds of interaction.

We refer to this capability as **action-conditioned spatial reasoning**. Unlike conventional observation-conditioned reasoning, action-conditioned spatial reasoning requires a model to reason not only about the current spatial state, but also about how its own action changes that state and how future decisions should adapt to the updated environment. To make this evaluation concrete and controllable, we instantiate action-conditioned spatial reasoning as interactive 3D layout refinement. This formulation is motivated by the fact that spatial validity, including collision avoidance, boundary consistency, and orientation plausibility, is a fundamental requirement for usable 3D scenes [1, 3, 4, 5, 7, 20, 33, 40]. At the same time, layout refinement provides executable high-level actions and objectively verifiable state changes, allowing us to isolate spatial reasoning from low-level motor control.

To this end, we introduce **SpatialAct**, a simulator-grounded benchmark for probing reasoning-to-action capabilities of VLM agents in 3D scenes. **SpatialAct** abstracts away low-level control and focuses on high-level semantic spatial actions, where models issue executable commands such as moving, rotating, or scaling objects with specific parameters. These commands are parsed and executed in a 3D simulator, and the updated multi-view renderings are then returned to the model for continued reasoning and action. **SpatialAct** covers three types of 3D scenarios, including *Abstract Geometric*, *Urban Architectural*, and *Indoor Scene*, with 333 scenes and 4,355 QA pairs across three question formats. Each scene is rendered from both top-view and isometric-view perspectives, while procedural generation, simulator execution, and dynamically injected spatial errors enable controllable evaluation and reduce the risk of static benchmark contamination.

SpatialAct follows a hierarchical diagnostic design that connects basic spatial understanding to multi-turn action-conditioned refinement. It first evaluates five Basic Spatial Abilities, including object meaning, spatial relations, spatial orientation, mental rotation, and spatial visualization. It then tests whether models can identify and repair spatial errors in one step through Single-step Error Detection and Fix. Finally, it challenges models to iteratively repair abnormal spatial configurations through closed-loop simulator feedback in Multi-turn Interactive Refinement. Empirically, our experiments reveal a substantial reasoning-to-action gap. Although current strong VLMs perform well on isolated spatial tasks and reach around 80% accuracy in several basic categories, the strongest VLM only achieves 0.411 Repair Rate and 0.206 Scene Success Rate in multi-turn simulator-grounded refinement. In contrast, human participants achieve 0.911 Repair Rate and 0.763 Scene Success Rate. These results suggest that current VLMs may recognize local spatial relations, but still struggle to

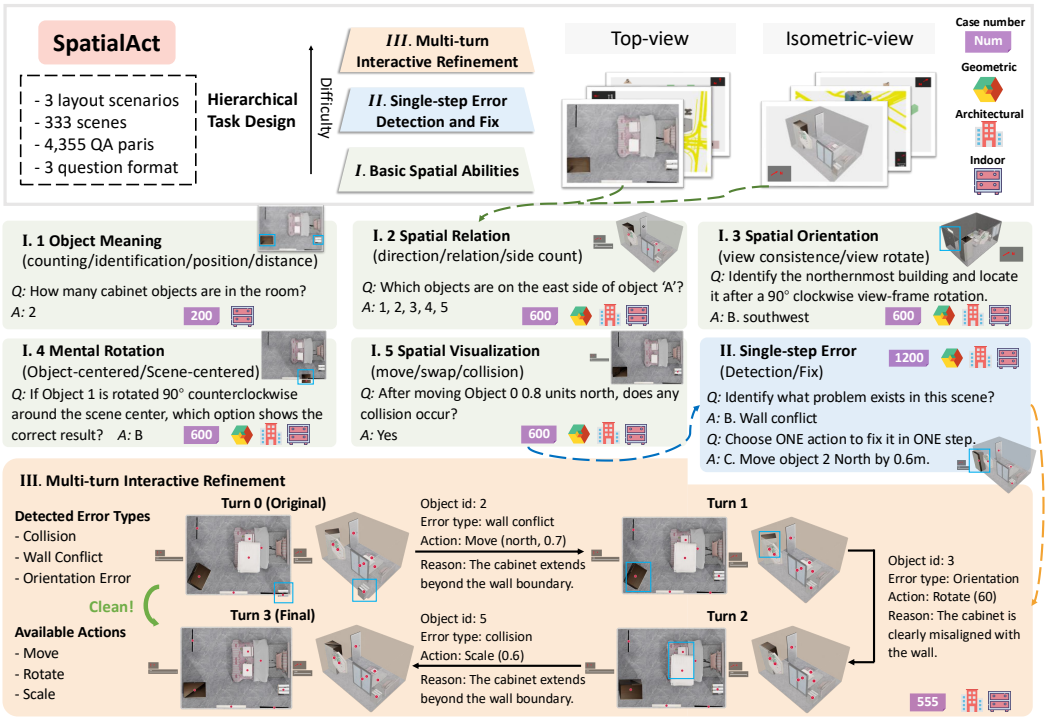


Figure 1: Framework of SpatialAct, which introduces a three-level hierarchical task design to systematically evaluate and diagnose agentic VLMs' 3D spatial reasoning and action capabilities.

maintain coherent spatial beliefs and produce reliable actions across long-horizon state transitions. Our contributions are threefold.

- We formalize **action-conditioned spatial reasoning** as a missing evaluation axis for VLM agents, where high-level simulator-executable actions change the 3D environment state that future reasoning depends on.
- We introduce **SpatialAct**, a simulator-grounded benchmark with 333 scenes and 4,355 QA pairs across *Abstract Geometric*, *Urban Architectural*, and *Indoor Scene* scenarios, organized into a hierarchical design covering Basic Spatial Abilities, Single-step Error Detection and Fix, and Multi-turn Interactive Refinement.
- We conduct a systematic evaluation of leading proprietary and open-source VLMs, revealing a clear reasoning-to-action gap between basic spatial understanding and stable multi-turn spatial refinement. We will release the benchmark, simulator workflow, evaluation platform, and analysis toolkit to support future research on spatially grounded VLM agents. Our codes and datasets are open-sourced via `https://github.com/tsinghua-fib-lab/SpatialAct`.

2 Methods

We introduce **SpatialAct** to evaluate agentic VLMs in 3D environments from reasoning to action. As shown in Figure 1, SpatialAct covers three types of layout scenarios: Abstract Geometric, Urban Architectural, and Indoor Scene, where the latter two are constructed from 333 Scenes. In total, SpatialAct contains 4,355 QA pairs spanning 3 question formats, providing a unified testbed for evaluating spatial understanding across controlled geometric layouts and daily-life 3D scenes.

2.1 Benchmark Construction

To evaluate the 3D spatial reasoning and action capabilities of VLMs across diverse scenarios, we construct a scene-level dataset as illustrated in Figure 2. The pipeline consists of data collection, quality control, and QA pair generation. Each scene is represented by both a top-view and an isometric-view, enabling models to access complementary spatial information from different perspectives.

Data Collection SpatialAct consists of three types of scenarios: Abstract Geometric, Urban Architectural, and Indoor Scene. The Abstract Geometric layouts are procedurally generated in

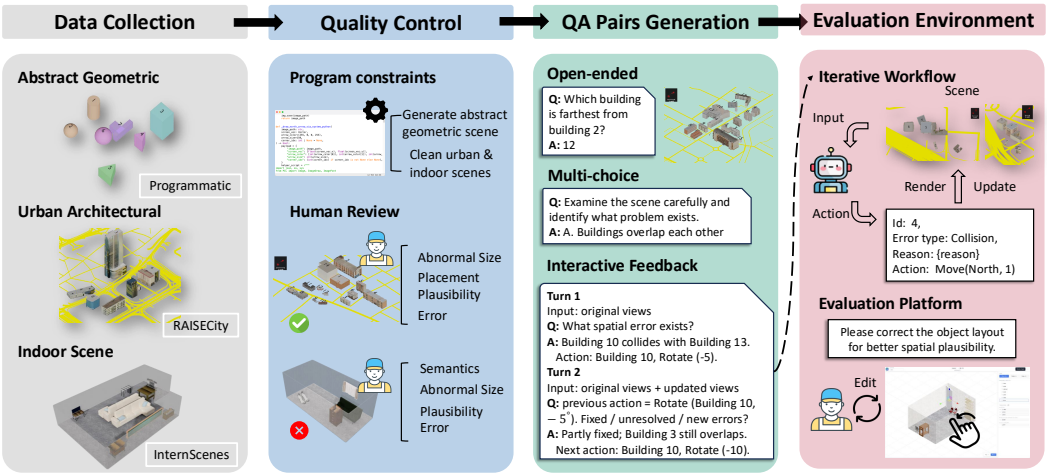


Figure 2: Benchmark construction pipeline, including data collection, quality control, QA pair generation, and evaluation environment setup.

a controllable manner. Each scene typically contains 5-7 geometric objects, which are randomly sampled from a predefined set of candidates, including cube, cuboid, cylinder, prism, sphere, L-shape, and U-shape, with varying sizes. For more complex tasks such as Mental Rotation, we further constrain each scene to contain at least one complex L-shaped or U-shaped object, ensuring sufficient structural complexity for spatial reasoning. The Urban Architectural layouts are derived from RAISECity [24], a multimodal agent framework for reality-aligned 3D world generation. We use both its white-box building models and textured building models to construct mixed urban scenes, while restricting each scene to contain no more than 20 buildings. The Indoor Scene layouts are built from InternScenes [41], which provides diverse indoor environments with movable furniture meshes. Considering the higher visual and structural complexity of indoor environments, we select scenes containing 5–15 objects for benchmark construction.

Quality Control For Abstract Geometric scenes, the layouts are generated procedurally under predefined constraints, and therefore no additional filtering is required. For Urban Architectural and Indoor Scene layouts, although the original scene-generation pipelines already provide relatively well-structured scenes, our benchmark focuses on error patterns in spatial configurations. We therefore first apply programmatic cleaning to remove existing abnormal cases from the original scenes. We then conduct manual filtering to examine abnormal element sizes, the plausibility of spatial arrangements, and potential error patterns. For Indoor Scene layouts, we additionally check semantic consistency, as indoor objects carry rich semantic categories. These inspections ensure the quality and reliability of the final benchmark scenes.

Task Design and QA Pairs Generation SpatialAct follows a hierarchical task design centered on Multi-turn Interactive Refinement, the most challenging setting in our benchmark. In this task, we first inject layout errors of varying difficulty and correction steps into clean, error-free scenes obtained from the previous filtering stage. Models are required to iteratively repair abnormal errors in a scene through multi-turn action-based interaction. To facilitate analysis of the underlying causes of model performance, we further design two levels of diagnostic tasks. First, Single-step Error Detection and Fix decomposes the interactive refinement process into a one-step setting, focusing on whether models can detect a spatial error and choose the corresponding correction. Second, we introduce five basic spatial ability tasks: Object Meaning, Spatial Relation, Spatial Orientation, Mental Rotation, and Spatial Visualization. These dimensions are adapted from prior work on basic spatial abilities [31], which identifies five key dimensions based on psychometric theories. While the original tasks are mainly designed around classical geometry and cube-like scenarios, we reinterpret their theoretical meanings and redesign the questions for Abstract Geometric scenes, further extending them to more common daily-life scenarios, including Urban Architectural and Indoor Scene environments. Each task category contains multiple QA subtypes, each emphasizing different aspects of spatial cognition. Detailed QA subtype definitions are provided in the Appendix A.2. Depending on the characteristics of each task, QA pairs are formulated in three formats: open-ended questions, multiple-choice questions (MCQs), and multi-turn feedback-based interaction.

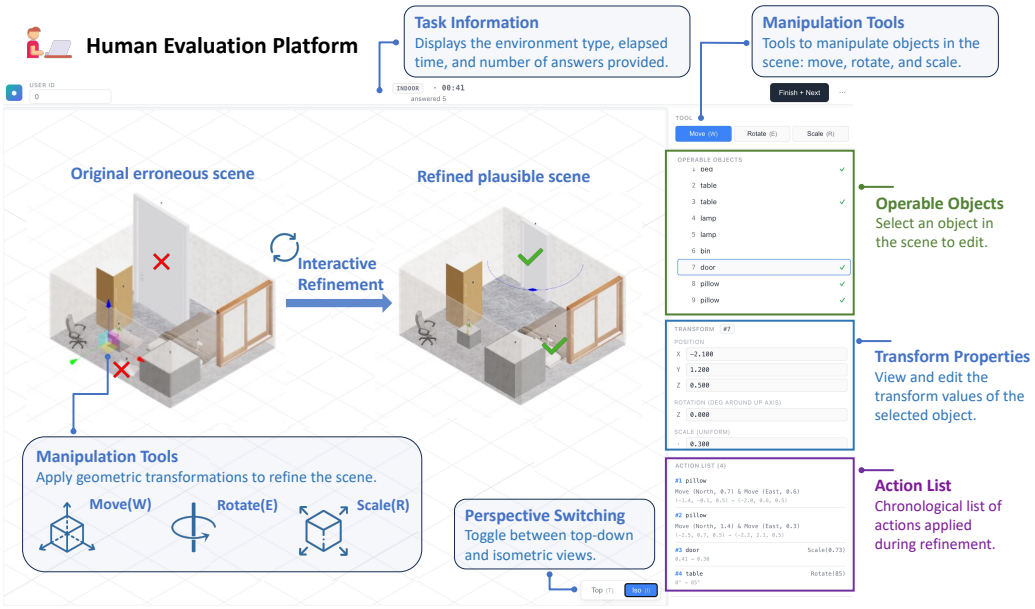


Figure 3: The web-based evaluation interface for interactive scene refinement.

2.2 Evaluation Environment Setup

To evaluate the reasoning-to-action capabilities of current VLMs, we design the Multi-turn Interactive Refinement task, where models are required to interact with the environment through iterative action-based feedback. Unlike static QA tasks, this setting requires the model to inspect the current scene, identify spatial errors, generate corrective actions, and refine its decisions based on the updated environment. In the following, we describe its implementation through the Multi-turn Iterative Rendering Workflow and the human Evaluation Platform.

Multi-turn Iterative Rendering Workflow In the Multi-turn Interactive Refinement task, VLMs are required to detect and correct abnormal spatial configurations through iterative action-based interaction. These error types are designed to reflect common failure modes in 3D layout construction [20, 4], where objects or buildings must avoid physical conflicts, stay within valid spatial boundaries, and maintain plausible orientations. For Urban Architectural scenes, the model is asked to identify three types of building-related errors: collision, road conflict, and orientation error. For Indoor Scene layouts, the model focuses on three types of object-related errors: collision, wall conflict, and orientation error. At each turn, the model can manipulate the current scene using three types of actions: `move(direction, number)`, `rotate(degree)`, and `scale(number)`. As illustrated in the rightmost part of Figure 2, each interaction starts by providing the VLM with the top-view and isometric-view renderings of an original erroneous scene. The model then determines which spatial errors exist and outputs corrective action commands. These commands are parsed by and executed in the simulator, after which the updated scene is rendered again from both the top and isometric views. In the next turn, the VLM receives the newly rendered views together with the original top-view and isometric-view images, the last-turn action command, and the accumulated previous action history. This iterative process continues until the VLM judges that all errors in the scene have been fixed or the predefined maximum number of turns is reached.

Human Evaluation Platform To enable human participants to perform the Multi-turn Interactive Refinement task, we develop a web-based evaluation interface. As illustrated in Figure 3, the platform provides an intuitive 3D environment equipped with a Transform Gizmo, allowing participants to perform free-form editing by directly manipulating objects in the scene. Users can select target entities from the Object Checklist and precisely adjust their spatial properties (position, rotation, and scale) using either the interactive handle or the Transform Properties panel. To ensure high-quality refinement, the interface also features perspective switching for multi-angle inspection and an Operation History log for tracking iterative modifications. This flexible workflow empowers human users to effectively rectify spatial inconsistencies, ensuring the final generated scenes are both structurally and functionally plausible.

Table 2: Main results on the SpatialAct. Rep. Rate denotes Repair Rate, Succ. Rate denotes Scene Success Rate, Eff. Turn denotes Effective Repair Turn Ratio, Pre. Stop denotes Premature Stop Ratio, and Tok./Scene denotes Average Completion Tokens per Scene.

Task	**Multi-turn Refinement**					**Single-step Edit**	**Object Meaning**	**Spatial Relation**	**Spatial Orientation**	**Mental Rotation**	**Spatial Visualization**
Metrics	Rep. Rate ↑	Succ. Rate ↑	Eff. Turn ↑	Pre. Stop ↓	Tok./ Scene	Acc. ↑	Acc. ↑	Acc. ↑	Acc. ↑	Acc. ↑	Acc. ↑
Proprietary Models											
GLM-5V-Turbo	-0.012	0.035	0.262	0.864	31101	0.452	0.695	0.738	0.617	0.500	0.665
GPT-5.4 mini	0.088	0.009	0.176	0.885	10060	0.595	0.720	0.810	0.616	0.570	0.813
GPT-5.4	0.208	0.038	0.228	0.791	23040	0.664	0.751	0.826	0.625	**0.690**	**0.850**
Gemini-3.1 Pro	**0.411**	**0.206**	**0.293**	**0.566**	22814	**0.721**	**0.770**	**0.835**	**0.720**	0.628	0.785
Open-source Models											
Kimi-K2.5	0.032	0.009	0.055	0.920	9211	0.368	0.615	0.670	0.393	0.328	0.473
Qwen3.6-27B	0.035	0.005	0.020	0.922	8565	0.343	0.600	0.668	0.446	0.155	0.581
Qwen3.6-35B-A3B	-0.099	0.016	0.252	0.920	196313	0.442	0.640	0.788	0.681	0.585	0.700

Metric Design For the first-level Basic Spatial Abilities tasks and the second-level Single-step Error Detection and Fix task, we use accuracy as the evaluation metric. For the Multi-turn Interactive Refinement task, we design five metrics targeting two aspects of model performance: repair accuracy and repair efficiency. We first introduce two metrics related to repair accuracy. First, we define the Repair Rate to measure the model's ability to reduce scene errors:

$$\text{Repair Rate} = \frac{E_{\text{before}} - E_{\text{after}}}{E_{\text{before}}}, \tag{1}$$

where E_{before} and E_{after} denote the number of errors in the scenes before and after model refinement, respectively. This metric directly reflects how effectively the model repairs abnormal spatial configurations.

Second, we define the Scene Success Rate to evaluate whether the model can completely resolve all errors at the scene level:

$$\text{Scene Success Rate} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \mathbb{I}(E^s_{\text{after}} = 0), \tag{2}$$

where E^s_{after} denotes the number of remaining errors in scene s after refinement. This metric measures the proportion of scenes that are fully corrected after model interaction, providing a scene-level evaluation of both error identification and correction.

Next, we introduce three efficiency-oriented metrics that measure how efficiently the model performs refinement during multi-turn interaction. First, we define the Effective Repair Turn Ratio to measure the proportion of interaction turns that actually reduce scene errors:

$$\text{Effective Repair Turn Ratio} = \frac{\sum_{s \in \mathcal{S}} \sum_{t=1}^{T_s} \mathbb{I}(E^s_t < E^s_{t-1})}{\sum_{s \in \mathcal{S}} T_s}, \tag{3}$$

where E^s_t denotes the number of errors remaining in scene s after the t-th refinement turn, and T_s denotes the total number of interaction turns performed for scene s. This metric reflects how frequently the model produces effective actions during refinement.

Second, we define the Premature Stop Rate to measure the proportion of scenes in which the model stops refinement while errors still remain:

$$\text{Premature Stop Rate} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \mathbb{I}(E^s_{T_s} > 0), \tag{4}$$

where $E^s_{T_s}$ denotes the number of remaining errors in scene s at the end of the interaction, either when the model decides to stop or when the maximum number of turns is reached. This metric captures whether the model can correctly judge the completion status of the scene.

Third, we define the Average Completion Tokens per Scene to measure the token cost:

$$\text{Avg. Completion Tokens} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} C_s \tag{5}$$

where C_s denotes the total number of completion tokens generated by the model across all interaction turns for scene s. This metric evaluates the reasoning cost of multi-turn refinement.

GLM-5V-Turbo

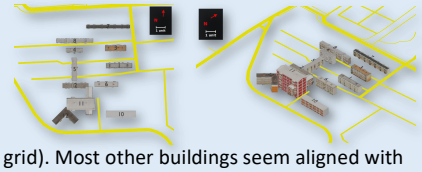
Erroneous understanding and reasoning

Reasoning:

…… Object 2 (stool): Circular object in the center-bottom area. Looks properly placed. Object 3 (curtain): The wavy/curved structure on the right side. This appears to be a curtain or fabric element. It seems to be within bounds. Object 4 (door): Located in the upper area, appears to be a door on the north wall. This looks like it's part of the wall structure. Object 5 (toilet): Located on the far right edge of the image. Looking at it carefully - this appears to be right at or possibly extending beyond the wall boundary on the east side. …… The toilet (object 5) sits very close to the eastern boundary and may actually extend past the wall line. The shelf (object 8) also appears positioned near that same edge. In the isometric view, both objects seem to be at or beyond where the transparent wall surface ends, suggesting they might have wall conflict issues. The door (object 4) appears integrated into the north wall as expected, so that placement looks correct. The curtain (object 3) is a soft, flexible element that doesn't create hard geometry conflicts. For the rectangular objects like the shelf, I need to verify whether their orientation aligns properly with the room's axes.

Actions:

Object 3: Move(West, 0.3)
Building 8: Move (Southwest, 0.4)



Erroneous action

Reasoning:

…… Building 12 appears to be rotated at an angle (not aligned with grid). Most other buildings seem aligned with the grid. Building 5: This vertical building appears to have a yellow road running through its lower portion. This is a RoadViolation. Building 5 sits directly on top of a horizontal yellow road, which is a clear violation. Building 12 is rotated approximately 45 degrees relative to the road grid, making it misaligned with the surrounding orientation.

Actions:

Building 5: Move(North, 0.8)
Building 12: Rotate (-45)

Figure 4: Representative error cases of GLM-5V-Turbo in SpatialAct.

3 Experiments

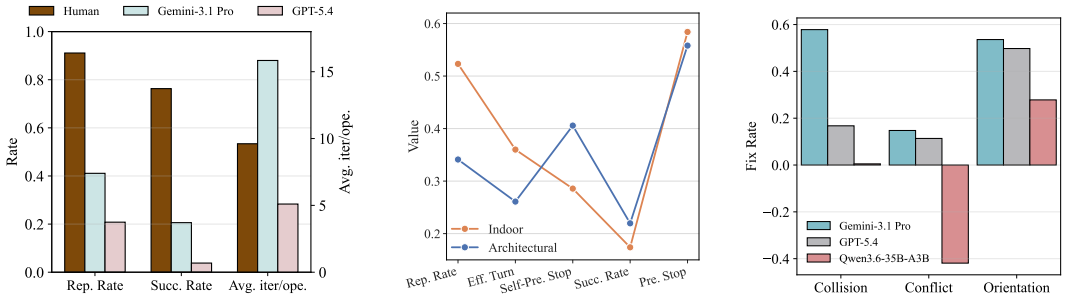
3.1 Benchmark VLM Agents

We comprehensively evaluate 7 strong VLMs, covering diverse model scales and model families. For proprietary models, we include GPT-5.4 mini, GPT-5.4 [15], GLM-5V-Turbo [6], and Gemini-3.1 Pro [2]. For open-source models, we evaluate Kimi-K2.5 [22], Qwen3.6-27B [16], and Qwen3.6-35B-A3B [17]. During evaluation, we adopt a unified inference setting across all models. For the Multi-turn Interactive Refinement task, we set the maximum number of iterative turns to 30. Since the evaluated models are reasoning models with a context window of approximately 200K tokens, we limit the maximum number of completion tokens for each single-turn response to 8,096.

3.2 Main Results

Overall Performance of Multi-turn Interactive Refinement As shown in Table 2, closed-source models substantially outperform other models on the Multi-turn Interactive Refinement task. In particular, Gemini-3.1 Pro achieve the leading performance, with Repair Rates of 0.411 and Scene Success Rates of 0.206, respectively. It also exhibits the highest refinement efficiency, indicating that stronger proprietary VLMs are not only more capable of reducing spatial errors, but also more effective in completing repairs with fewer redundant interactions and lower interaction cost. In contrast, open-source models, as well as GLM-5V-Turbo, perform poorly on this task. Their Repair Rates are close to zero or even negative, suggesting that these models often fail to correct existing errors and may introduce additional errors during interaction. To make this bottleneck concrete, Figure 4 shows two representative failure modes: the model can misdiagnose whether a spatial violation exists, and even when the diagnosis is correct, it can still issue an incorrect corrective action. Together, these cases indicate that the core weakness lies in both error diagnosis and reasoning-to-action execution. Overall, Multi-turn Interactive Refinement remains a highly challenging task for existing VLMs. Although the strongest closed-source models show encouraging progress, their scene-level success rates are still far from saturation, while most open-source models struggle to perform reliable spatial repair.

Results on Single-step and Basic Spatial Ability Tasks We further evaluate model results on the Single-step Detection and Fix task and basic spatial ability tasks. Compared with Multi-turn Interactive Refinement, models generally achieve higher accuracies in these settings, suggesting that current VLMs are more capable of solving isolated spatial reasoning problems than performing



(a) (b) (c)

Figure 5: (a) Comparison of human and model performance. 'Avg. iter./ope.' denotes average iteration/operation per scene. (b) Performance of the model across different layout types. 'Self-Pre. Stop' denotes the fraction of turns in which the model chooses to stop before reaching the maximum allowed iterations while errors remain; lower values indicate better performance. (c) Model sensitivity to different error types and correction performance.

long-horizon interactive repair. For Single-step Detection and Fix, proprietary models maintain a clear advantage. Gemini-3.1 Pro achieves the best accuracy of 0.721, followed by GPT-5.4 and GPT-5.4 mini with 0.664 and 0.595, respectively, while open-source models remain much lower, ranging from 0.343 to 0.442. For basic spatial ability tasks, Gemini-3.1 Pro and GPT-5.4 remain the leading models across all five tasks. Open-source models show uneven performance across basic abilities. For example, Kimi-K2.5 performs relatively well on Spatial Relation, but remains weaker on Mental Rotation. Overall, these results show that models exhibit diverse basic spatial capabilities, but these capabilities alone do not directly indicate their performance in multi-turn refinement.

Human Baseline Seven human testers performed the Multi-turn Interactive Refinement task using the evaluation platform described in Section 2, as shown in Figure 5a. These repair tasks are straightforward for humans, who achieve a Repair Rate of 0.911 and a Scene Success Rate of 0.763. The repair rate exceeds that of the best-performing model by 50 percentage points, highlighting the remaining gaps in current VLMs in reasoning about scene errors, planning corrective actions, and executing multi-turn interactions.

Key Findings: Multi-turn interactive spatial correction remains a major bottleneck.

- **State-tracking limitation:** Current VLMs can solve many static or single-step spatial tasks, but fail to maintain consistent spatial state across iterative action-feedback loops.
- **Failure modes:** In Multi-turn Interactive Refinement, current VLMs fail for two coupled reasons: *diagnosis errors* (understanding and reasoning) and *reasoning-to-action errors*.
- **Human-level gap:** Although iterative spatial correction is simple for humans, current VLMs remain about 55% lower in Repair Accuracy, revealing a substantial reliability gap.

3.3 Spatial Context and Error Sensitivity

Scene-wise Analysis of VLM Performance We analyze the performance of Gemini-3.1 Pro separately on indoor and urban architectural scenes, as shown in Figure 5b. Compared with architectural layouts, the model achieves higher Repair Rate and Effective Repair Turn Ratio in indoor scenes, and its self-selected Premature Stop Ratio is lower. This indicates that the model can more easily identify and correct errors in indoor scenes, likely reflecting greater exposure to such environments in its training data. Interestingly, architectural scenes show higher Scene Success Rate despite weaker per-turn repair effectiveness. A plausible interpretation is that, although architectural scenes are less familiar to the model, their layouts may be simpler and include fewer tightly coupled object dependencies, making full-scene completion easier once key errors are addressed. Overall, the model appears stronger in indoor repair behavior, while final scene success is also shaped by scene-level structural simplicity and error coupling.

Error-type Sensitivity in VLM Refinement Figure 5c shows a consistent error-type hierarchy across Gemini-3.1 Pro, GPT-5.4, and Qwen3.6-35B-A3B: orientation errors are most recoverable, while conflict errors, whether related to roads or walls, are the most challenging. This pattern suggests that current VLMs are relatively strong at attribute-level adjustments but weaker at constraint-level

reasoning that requires jointly modeling boundaries, topology, and multi-object relations. Importantly, conflict repair often requires coordinated updates rather than single-object edits, so one incorrect step can propagate new violations across turns. These findings imply that the bottleneck is not only perception, but constraint-aware planning and cross-turn consistency under coupled spatial dependencies.

Impact of Initial Scene Complexity on Model Performance As shown in Figure 6a, we group indoor scenes by initial complexity based on the number of errors: 1–3, 4–6, and 7 or more. Not surprisingly, as the initial scene complexity increases, both Repair Rate and Scene Success Rate gradually decrease. This pattern reflects the increasing difficulty of tracking multiple errors simultaneously, which often leads to missed detections or conflicting corrective actions. Multi-turn interactions in such complex scenes also demand stronger long-horizon consistency and error management capabilities. Current VLMs are prone to interference when handling multiple errors, particularly in scenarios involving spatial conflicts or complex dependency relationships, resulting in poor refinement performance as scene complexity rises.

Key Findings: Joint effects of scene context, structural constraints, and error complexity.

- **Scene-dependent behavior:** VLMs are more effective at repairing errors in indoor scenes.
- **Constraint-level bottleneck:** Across models, orientation errors are easier to repair than conflict errors, indicating that the main limitation lies in constraint-aware, multi-object coordination.
- **Complexity amplifies instability:** Current VLMs struggle to maintain cross-turn consistency under dense, coupled error conditions.

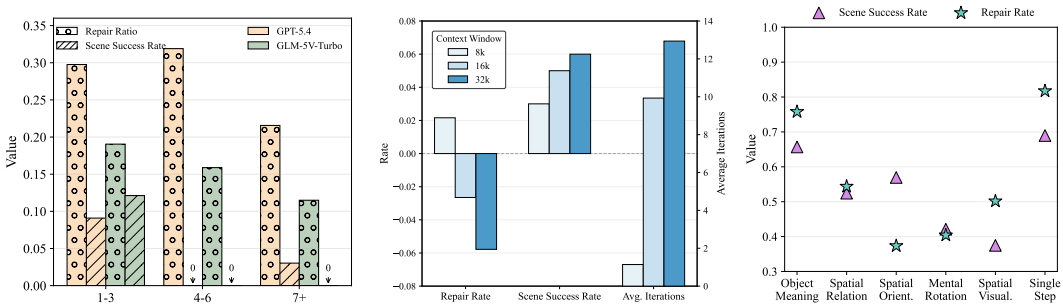
3.4 Influence of Context and Task Relationship on VLM Behavior

Effect of Context Window Size As illustrated in Figure 6b, we analyze context-window effects on Kimi-K2.5, a model that tends to produce long reasoning traces, using a random subset of 100 test examples. As the context window increases, the model produces more environment-applicable content and engages in more interaction turns, indicating that additional context is converted into more explicit deliberation and action attempts. However, Repair Rate and Scene Success Rate remain nearly unchanged. This decoupling between *reasoning volume* and *repair outcome* suggests a diminishing-return regime in which longer contexts primarily expand verbose or repetitive reasoning rather than improving correction quality. A likely explanation is that the bottleneck is not token budget itself, but cross-turn control quality, including state tracking, error prioritization, and action reliability under feedback. Under this view, larger windows increase opportunity for exploration but do not strengthen the policy that selects effective repairs. This also supports the context setting used in Table 2: once a sufficient reasoning budget is reached, further context expansion adds computational cost with limited performance gain.

Correlation between Simple and Complex Task Performance Figure 6c relates performance on six foundational tasks (five Basic Spatial Abilities plus Single-step Error Detection and Fix) to Multi-turn Interactive Refinement. Positive Pearson correlations across all six tasks indicate structural continuity between task levels, suggesting that foundational abilities provide the operational components required in complex multi-turn repair. Single-step Error Detection and Fix shows the strongest association with both Repair Rate ($r = 0.817$) and Scene Success Rate ($r = 0.690$), indicating that local detect-and-correct operations serve as the fundamental building blocks repeatedly composed in iterative refinement. Object Meaning is the second strongest correlate, suggesting that object-centric grounding, including identity, position, and relative spatial anchoring, is critical for propagating local edits to scene-level consistency. At the same time, the correlation pattern shows that foundational competence alone does not guarantee robust multi-turn success, since iterative repair additionally depends on cross-turn memory, conflict-aware planning, and stable action sequencing. Overall, complex refinement can be understood as a hierarchical integration of foundational spatial skills with higher-order coordination over feedback loops.

Key Findings: Control quality and hierarchical skill integration over context length.

- **Diminishing returns from context scaling:** More tokens and turns, but limited improvement in Repair Rate and Scene Success Rate.
- **Hierarchical but non-automatic skill transfer:** Foundational abilities align with multi-turn demands, yet robust performance still requires higher-order cross-turn coordination.



(a) (b) (c)

Figure 6: (a) Comparison of model performance under different initial scene complexities. (b) Model performance under different context window settings. (c) Relationship between performance on basic spatial tasks and complex interactive tasks, quantified by Pearson correlation coefficients (r).

4 Related Work

3D Spatial Intelligence Evaluation With the rapid development of VLMs, increasing attention has been paid to whether these models can perceive, reason about, and act upon spatial information in a human-like manner. This has led to a growing body of benchmarks for evaluating spatial intelligence from different perspectives. Some benchmarks [27, 38, 26, 19, 9, 13, 35, 29, 18], such as Spatial-DISE [8] and BSA [31], construct hierarchical suites of classical spatial tasks to assess fundamental spatial abilities in a systematic way. Recent work has further expanded the scope of spatial evaluation to more diverse and realistic scenarios [23, 12]. For example, SpaceVista [21] investigates spatial understanding across all-scale scenes ranging from millimeters to kilometers, while other benchmarks extend object-centric reasoning from single-step judgement to multi-step interaction with the environment [37]. Meanwhile, model inputs have also evolved from single-view observations to richer multi-view settings [32, 34, 39, 10]. CityCube [30], for instance, collects images from different viewing positions and orientations to support more comprehensive spatial perception. In parallel, the objective of spatial evaluation has gradually shifted from passive spatial understanding to action-oriented spatial interaction [25, 36, 11], where models are expected not only to recognize spatial relations but also to modify the environment through actions. Our SpatialAct follows this direction and provides a comprehensive benchmark for evaluating agentic VLMs in 3D environments from reasoning to action.

3D Layout Generation and Understanding Recent advances in AI for 3D layout have explored both layout generation [4] and layout understanding [7, 1, 3]. On the generation side, existing methods increasingly leverage foundation models to inject semantic and spatial commonsense into 3D layout synthesis. For example, LayoutGPT [5] directly uses LLMs to produce structured layout representations, while Holodeck [33] and LayoutVLM [20] further combine VLM reasoning with spatial constraints or differentiable optimization to improve semantic coherence and physical plausibility. On the understanding side, recent work has begun to move beyond passive spatial perception toward layout editing and manipulation. [28] studies embodied agents that restore shuffled indoor scenes through observation and interaction, while 3D-Layout-R1 [40] formulates language-guided layout editing as structured scene-graph reasoning. However, most existing 3D layout works still leave open the question of whether vision-language models can directly complete layout tasks through reasoning and action. Since 3D layout quality critically depends on physical validity [20, 4], such as avoiding collisions, out-of-bound placements, and spatial misalignment, We introduce SpatialAct to evaluate whether agentic VLMs can reason over interactive 3D environments and generate action commands to iteratively repair or modify spatial layouts.

5 Conclusion

We present SpatialAct, a simulator-grounded hierarchical benchmark that probes whether VLMs can translate spatial reasoning into reliable action under dynamic 3D feedback. By jointly evaluating Multi-turn Interactive Refinement, Single-step Error Detection and Fix, and basic spatial abilities, SpatialAct reveals not only absolute performance gaps but also the capability structure behind them. Across models, strong results on basic tasks do not consistently transfer to robust multi-turn repair, indicating that the core bottleneck lies in cross-turn state maintenance, constraint-aware planning, and stable reasoning-to-action execution rather than perception alone. Proprietary models remain

clearly ahead of open-source models, but all current systems still fall far short of human reliability on iterative spatial correction. We further find that performance is systematically shaped by scene context, structural constraints, and error complexity, with models generally handling indoor scenes and orientation-related corrections more effectively. Beyond benchmarking, SpatialAct provides an actionable diagnostic framework for developing VLM agents with stronger spatial state tracking, coordination across feedback loops, and dependable context-grounded action generation.

References

[1] Ahmed Abdelreheem, Filippo Aleotti, Jamie Watson, Zawar Qureshi, Abdelrahman Eldesokey, Peter Wonka, Gabriel Brostow, Sara Vicente, and Guillermo Garcia-Hernando. Placeit3d: Language-guided object placement in real 3d scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6645–6655, 2025.

[2] Google DeepMind. Gemini-3.1 pro. `https://deepmind.google/models/model-cards/gemini-3-1-pro/`, 2026.

[3] Mohamed El Amine Boudjoghra, Ivan Laptev, and Angela Dai. Scanedit: Hierarchically-guided functional 3d scan editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 27105–27115, 2025.

[4] Haoran Feng, Yifan Niu, Zehuan Huang, Yang-Tian Sun, Chunchao Guo, Yuxin Peng, and Lu Sheng. Repurposing 3d generative model for autoregressive layout generation. *arXiv preprint arXiv:2604.16299*, 2026.

[5] Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems*, 36:18225–18250, 2023.

[6] Wenyi Hong, Xiaotao Gu, Ziyang Pan, Zhen Yang, Yuting Wang, Yue Wang, Yuanchang Yue, Yu Wang, Yanling Wang, Yan Wang, et al. Glm-5v-turbo: Toward a native foundation model for multimodal agents. *arXiv preprint arXiv:2604.26752*, 2026.

[7] Ian Huang, Yanan Bao, Karen Truong, Howard Zhou, Cordelia Schmid, Leonidas Guibas, and Alireza Fathi. Fireplace: Geometric refinements of llm common sense reasoning for 3d object placement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13466–13476, 2025.

[8] Xinmiao Huang, Qisong He, Zhenglin Huang, Boxuan Wang, Zhuoyun Li, Guangliang Cheng, Yi Dong, and Xiaowei Huang. Spatial-dise: A unified benchmark for evaluating spatial reasoning in vision-language models. *arXiv preprint arXiv:2510.13394*, 2025.

[9] Aditya Sanjiv Kanade and Tanuja Ganu. Do you see me: A multidimensional benchmark for evaluating visual perception in multimodal llms. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7285–7326, 2026.

[10] Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, et al. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. *arXiv preprint arXiv:2505.21500*, 2025.

[11] Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Li Li, Ruohan Zhang, Weiyu Liu, Percy Liang, Li Fei-Fei, Jiayuan Mao, and Jiajun Wu. Embodied agent interface: Benchmarking llms for embodied decision making. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 100428–100534. Curran Associates, Inc., 2024.

[12] Yijiang Li, Qingying Gao, Tianwei Zhao, Bingyang Wang, Haoran Sun, Haiyun Lyu, Robert D Hawkins, Nuno Vasconcelos, Tal Golan, Dezhi Luo, et al. Core knowledge deficits in multi-modal language models. *arXiv preprint arXiv:2410.10855*, 2024.

[13] Weichen Liu, Qiyao Xue, Haoming Wang, Xiangyu Yin, Boyuan Yang, and Wei Gao. Spatial reasoning in multimodal large language models: A survey of tasks, benchmarks and methods. *arXiv preprint arXiv:2511.15722*, 2025.

[14] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16488–16498, 2024.

[15] OpenAI. Gpt-5.4. https://openai.com/index/introducing-gpt-5-4/, 2026.

[16] Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026.

[17] Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026.

[18] Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind: Failing to translate detailed visual features into words. *arXiv preprint arXiv:2407.06581*, 2024.

[19] Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp Kraehenbuehl, and Vladlen Koltun. Does spatial cognition emerge in frontier models? *arXiv preprint arXiv:2410.06468*, 2024.

[20] Fan-Yun Sun, Weiyu Liu, Siyi Gu, Dylan Lim, Goutam Bhat, Federico Tombari, Manling Li, Nick Haber, and Jiajun Wu. Layoutvlm: Differentiable optimization of 3d layout via vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29469–29478, 2025.

[21] Peiwen Sun, Shiqiang Lang, Dongming Wu, Yi Ding, Kaituo Feng, Huadai Liu, Zhen Ye, Rui Liu, Yun-Hui Liu, Jianan Wang, et al. Spacevista: All-scale visual spatial reasoning from mm to km. *arXiv preprint arXiv:2510.09606*, 2025.

[22] Kimi Team and et al. Kimi k2.5: Visual agentic intelligence, 2026.

[23] Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *Advances in Neural Information Processing Systems*, 37:75392–75421, 2024.

[24] Shengyuan Wang, Zhiheng Zheng, Yu Shang, Lixuan He, Yangcheng Yu, Fan Hangyu, Jie Feng, Qingmin Liao, and Yong Li. Raisecity: A multimodal agent framework for reality-aligned 3d world generation at city-scale. *arXiv preprint arXiv:2511.18005*, 2025.

[25] Weizhen Wang, Chenda Duan, Zhenghao Peng, Yuxin Liu, and Bolei Zhou. Embodied scene understanding for vision language models via metavqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22453–22464, 2025.

[26] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9058–9069, 2025.

[27] Xingrui Wang, Wufei Ma, Tiezheng Zhang, Celso M de Melo, Jieneng Chen, and Alan Yuille. Spatial457: A diagnostic benchmark for 6d spatial reasoning of large mutimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24669–24679, 2025.

[28] Luca Weihs, Matt Deitke, Aniruddha Kembhavi, and Roozbeh Mottaghi. Visual room rearrangement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5922–5931, 2021.

[29] Yuxi Xiao, Longfei Li, Shen Yan, Xinhang Liu, Sida Peng, Yunchao Wei, Xiaowei Zhou, and Bingyi Kang. Spatialtree : How spatial abilities branch out in MLLMs. In *The First Workshop on Efficient Spatial Reasoning*, 2026.

[30] Haotian Xu, Yue Hu, Zhengqiu Zhu, Chen Gao, Ziyou Wang, Junreng Rao, Wenhao Lu, Weishi Li, Quanjun Yin, and Yong Li. Citycube: Benchmarking cross-view spatial reasoning on vision-language models in urban environments. *arXiv preprint arXiv:2601.14339*, 2026.

[31] Wenrui Xu, Dalin Lyu, Weihang Wang, Jie Feng, Chen Gao, and Yong Li. Defining and evaluating visual language models' basic spatial abilities: A perspective from psychometrics. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11571–11590, 2025.

[32] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643, 2025.

[33] Yue Yang, Fan-Yun Sun, Luca Weihs, Eli VanderBilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, et al. Holodeck: Language guided generation of 3d embodied ai environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16227–16237, 2024.

[34] Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. In *Structural Priors for Vision Workshop at ICCV'25*, 2025.

[35] Songsong Yu, Yuxin Chen, Hao Ju, Lianjie Jia, Fuxi Zhang, Shaofei Huang, Yuhan Wu, Rundi Cui, Binghao Ran, Zaibin Zhang, et al. How far are vlms from visual spatial intelligence? a benchmark-driven perspective. *arXiv preprint arXiv:2509.18905*, 2025.

[36] Lingfeng Zhang, Yuening Wang, Hongjian Gu, Atia Hamidizadeh, Zhanguang Zhang, Yuecheng Liu, Yutong Wang, David Gamaliel Arcos Bravo, Junyi Dong, Shunbo Zhou, et al. Et-plan-bench: Embodied task-level planning benchmark towards spatial-temporal cognition with foundation models. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 21566–21573. IEEE, 2025.

[37] Pingyue Zhang, Zihan Huang, Yue Wang, Jieyu Zhang, Letian Xue, Zihan Wang, Qineng Wang, Keshigeyan Chandrasegaran, Ruohan Zhang, Yejin Choi, et al. Theory of space: Can foundation models construct spatial beliefs through active exploration? In *The Fourteenth International Conference on Learning Representations*, 2026.

[38] Wenyu Zhang, Wei En Ng, Lixin Ma, Yuwen Wang, Junqi Zhao, Allison Koenecke, Boyang Li, and Lu Wang. SPHERE: Unveiling spatial blind spots in vision-language models through hierarchical evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11591–11609, Vienna, Austria, July 2025. Association for Computational Linguistics.

[39] Yong Zhao, Kai Xu, Zhengqiu Zhu, Yue Hu, Zhiheng Zheng, Yingfeng Chen, Yatai Ji, Chen Gao, Yong Li, and Jincai Huang. Cityeqa: A hierarchical llm agent on embodied question answering benchmark in city space. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12476–12491, 2025.

[40] Haoyu Zhen, Xiaolong Li, Yilin Zhao, Han Zhang, Sifei Liu, Kaichun Mo, Chuang Gan, and Subhashree Radhakrishnan. 3d-layout-r1: Structured reasoning for language-instructed spatial editing. *arXiv preprint arXiv:2603.22279*, 2026.

[41] Weipeng Zhong, Peizhou Cao, Yichen Jin, Li Luo, Wenzhe Cai, Jingli Lin, Hanqing Wang, Zhaoyang Lyu, Tai Wang, Bo Dai, et al. Internscenes: A large-scale simulatable indoor scene dataset with realistic layouts. *arXiv preprint arXiv:2509.10813*, 2025.

A Appendix

A.1 Discussion and Future Work

Our results reveal persistent challenges in multi-turn reasoning and action planning, with models showing systematic biases across scenes and error types. A key limitation is that all evaluations are conducted in simulated scenes, and it remains unclear how performance translates to real-world environments. Additionally, we did not explore methods to enhance model capabilities for this task. Future work includes extending evaluation to real-world scenes, exploring methods to improve multi-turn reasoning and action execution, and developing more robust, human-like spatial reasoning-to-action capabilities.

A.2 Task Subcategories and Prompt Design

To provide a fine-grained diagnosis of model performance, we further divide the Basic Spatial Ability tasks and the Single-step Error Detection and Fix task into multiple subcategories. As shown in Table 3, the five Basic Spatial Ability tasks cover Object Meaning, Spatial Relation, Spatial Orientation, Mental Rotation, and Spatial Visualization, with each category containing several subtypes that emphasize different aspects of 3D spatial understanding and reasoning. In addition, the Single-step Error Detection and Fix task is divided according to different error patterns, enabling targeted evaluation of whether models can identify abnormal spatial configurations and select the corresponding correction. Representative prompts for each subcategory are provided in the table.

Table 3: Subcategories and representative prompts for Basic Spatial Ability tasks and Single-step Error Detection and Fix tasks.

Task	**Prompt**
Object Meaning (7)	How many bins are in the room? What is object 1? Which numbered object is a commode? Which side of the box is the frame on? Which numbered object is east of the commode? Which numbered object is closest to the cabinet? Which numbered object is farthest from the cabinet?
Spatial Relation (4)	Which side is building 1 relative to building 5? Which two buildings are closest to each other? Which building is farthest from building 2? How many buildings are on the east side of building 3?
Spatial Orientation (3)	Consider the southernmost building in the isometric image. Relative to the scene center, which direction is that building located in the top-view image? An isometric image where only one building is labeled 'A', and a top-view image with numbered buildings. Which numbered building is 'A' in the top-view image? Identify the westernmost building in the image. Now imagine that the camera rotates CLOCKWISE by 90 degrees around the scene center.When the camera rotates, the cardinal directions (north, east, south, west) rotate together with the view. After this rotation, where will that building be located relative to the scene center in the rotated coordinate frame?
Mental Rotation (4)	If building 1 is rotated counterclockwise by 30 degrees around its own center, will it collide with building 4? Rotate building 1 counterclockwise by one of the angles below around its own center. Which rotation angle avoids collision with building 4? Building 1 is rotated clockwise by 300 degrees.Which image matches this rotation? In the original image, Building 6 is rotated clockwise by 270 degrees around the center of the region. Which option shows the correct result after this rotation?
Spatial Visualization (2)	If building 1 is moved West by 6 meters, will it collide with building 6? If we swap the positions of building 4 and building 5, will there be any collision in the scene AFTER the swap?
Single-step Error Detection and Fix (2)	Examine the scene carefully and identify what problem exists. Known issue: Building 1 has an angle anomaly near a road intersection. Choose ONE action to fix it in ONE step. The fix should resolve the issue without introducing new problems.