From Perception to Simulation: Unveiling and Empowering Mental Spatial Dynamics in MLLMs

Anonymous ACL submission

Abstract

Spatial reasoning is a fundamental component of world understanding for Multimodal Large Language Models (MLLMs). While existing research has advanced the perception of spatial relations in static scenes, the domain of dynamic spatial simulation remains relatively unexplored. To bridge this gap, we introduce MentalGrid, a benchmark focused on spatial reasoning tasks that require models to simulate evolving spatial states without intermediate visual feedback. Extensive evaluation on our benchmark reveals a significant deficiency in the spatial simulation capabilities of existing models. Specifically, we find that performance is constrained by perceptual limitations in simple simulation scenarios and hindered by inadequate multi-step reasoning as complexity increases. To address these challenges, we propose a three-stage curriculum learning framework utilizing 16k specialized spatial reasoning samples. This approach yields SpatialWalker, which achieves a 33.8% average improvement over its base model, surpassing proprietary models like Gemini-3-Flash in spatial simulation. Furthermore, SpatialWalker demonstrates robust generalization, maintaining a 4.9% average gain on out-of-domain benchmarks, demonstrating the pivotal role of spatial simulation as a fundamental building block for broader spatial intelligence.

1 Introduction

Spatial reasoning is the ability of an agent to perceive, understand, and mentally manipulate spatial relationships between objects (Harris, 2023; Xiao et al., 2025; Liu et al., 2025a). It constitutes a fundamental component of how multimodal large language models (MLLMs) understand the physical world. Despite recent advances that enable models to achieve impressive performance on a wide range of visual reasoning tasks (Yue et al., 2024; Song et al., 2025; Zhu et al., 2025), their spatial reasoning capabilities remain limited (Yang et al., 2025a; Li et al., 2025b; Xiao et al., 2025). This gap has motivated growing interest in systematic investigations of spatial reasoning in MLLMs.

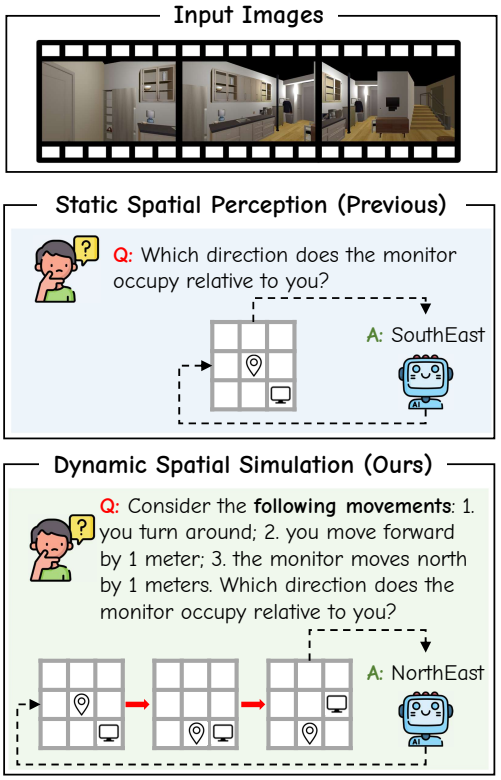


Figure 1: A comparison of two types of spatial reasoning capabilities. Arrows indicate the models' reasoning processes.

Recent work has primarily studied spatial reasoning by focusing on models' spatial perception capabilities (Wu et al., 2025a; Zhao et al., 2025; Brown et al., 2025). As illustrated by the first example in Figure 1, these works typically assume static environments, in which the spatial states of objects and agents remain fixed (Li et al., 2025a; Yang et al., 2025a,b). Under this setting, models can solve the task by performing a single-pass perception over the input images or videos to localize objects. However, real-world environments

are inherently dynamic, where the spatial positions of objects and agents continuously change over time. As shown in the second example of Figure 1, reasoning in such settings requires models to mentally simulate the evolving spatial state of the environment (e.g., the changing direction between the monitor and agent), a capability we refer to as **spatial simulation**. While existing studies have explored mental simulation, they predominantly focus on isolated objects (Jia et al., 2025; Li et al., 2025c) or restricted viewpoints (Li et al., 2025a; Yin et al., 2025), leaving this capability largely underexplored.

In cognitive neuroscience, spatial simulation is supported by internal representations like grid cells (Bellmund et al., 2016; Herweg and Kahana, 2018; Liang et al., 2024), which enable humans to mentally track their spatial state without continuous visual feedback. To investigate whether MLLMs exhibit a similar capacity, we introduce **MentalGrid**, a spatial reasoning benchmark featuring 3,356 question-answer pairs that evaluates models' ability to predict evolving spatial states without intermediate visual feedback. As illustrated in Figure 2, our benchmark goes beyond static spatial perception by integrating language-guided action sequences into spatial reasoning. It introduces final state prediction and dynamic process tracking tasks, requiring models to predict final spatial states such as distances and directions, as well as to track state changes throughout the movements.

Our extensive evaluation of 12 state-of-the-art MLLMs on MentalGrid reveals that existing models struggle with spatial simulation, while the best-performing Gemini-3.1-Pro achieves only around **58%** accuracy, remaining approximately **16%** behind human performance. Besides, for open-source models, spatial simulation performance is substantially worse than spatial perception. We further analyze the bottlenecks in the model's spatial simulation performance, finding that perceptual limitations constrain performance for simple simulation steps, while inadequate multi-step reasoning becomes the main limitation as complexity increases.

To mitigate these limitations, we design a three-stage curriculum learning framework based on 16k curated spatial reasoning samples. This framework starts with foundational perceptual grounding, advances to process-aware simulation training to internalize spatial transitions, and culminates in reinforcement learning-based optimization to refine fine-grained simulation trajectories. This approach yields **SpatialWalker**, a 3B model that achieves a **33.8%** average improvement over its base, surpassing proprietary systems like Gemini-3-Flash. Furthermore, SpatialWalker exhibits robust generalization with **4.9%** average performance gains on out-of-domain benchmarks, underscoring the role of spatial simulation as a fundamental building block for broader spatial intelligence.

Our contributions are as follows: **(1) MentalGrid Benchmark:** a novel benchmark for dynamic spatial simulation that requires tracking evolving states across action sequences without intermediate visual feedback. **(2) Curriculum Learning Framework:** a three-stage strategy, derived from our diagnostic findings, that evolves from perceptual grounding to spatial simulation trajectory optimization. **(3) SpatialWalker Model:** a 3B model that outperforms its base by **33.8%**, surpasses Gemini-3-Flash, and achieves an average **4.9%** out-of-domain gains.

2 MentalGrid Benchmark

2.1 Overview

We introduce MentalGrid, a multi-image spatial reasoning benchmark designed to evaluate the spatial simulation capabilities of MLLMs. MentalGrid comprises 3,356 questions over 1,100 simulated images. These images are generated from 349 distinct scenes with the Habitat-Sim simulator (Savva et al., 2019), representing a diverse set of real-world environments commonly encountered in daily life (e.g., offices, homes). As shown in Figure 2, each question in MentalGrid consists of image-based inputs specifying the initial spatial state, along with language-defined action sequences that update the environment state. Given these inputs, we require models to perform simulations over the dynamic spatial environment, including inter-object distances and relative directions, without access to intermediate visual observations. We report dataset statistics and comparisons with other benchmarks in the Appendix A.

2.2 Task Formulation

We formally define a spatial simulation trajectory $\mathcal{T}$ as the transition from an initial environment state $\mathcal{S}_0$ to a final state $\mathcal{S}_n$ following a sequence of actions $\mathcal{A}$:

$$\mathcal{T} = \mathcal{S}_0 \xrightarrow{\mathcal{A}} \mathcal{S}_f = \mathcal{S}_0 \xrightarrow{a_1} \mathcal{S}_1 \xrightarrow{a_2} \cdots \xrightarrow{a_n} \mathcal{S}_n \quad (1)$$

where $\mathcal{A} = \{a_1 \cdots a_n\}$, a_i denotes the action in step i, n is the length of $\mathcal{A}$ and $\mathcal{S}_i$ represents the in-

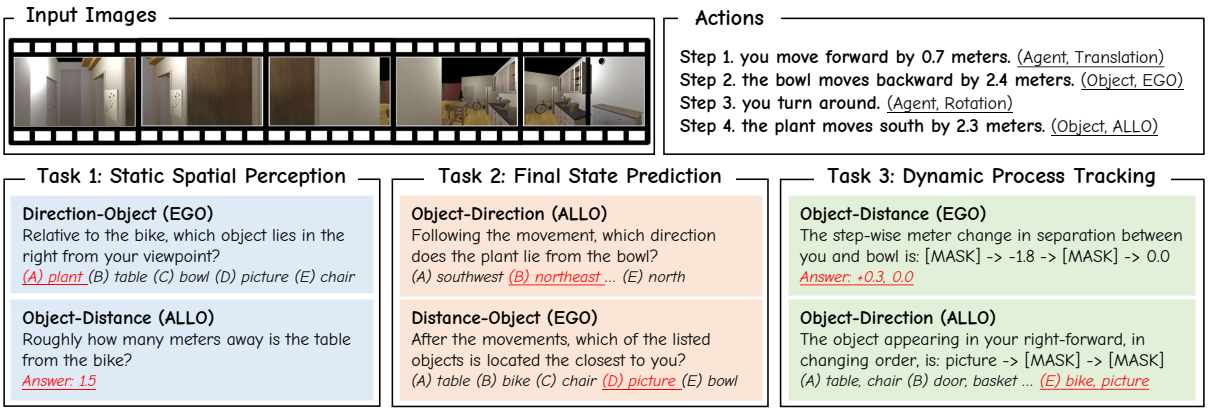


Figure 2: Examples in our MentalGrid benchmark.

termediate environment state after executing action a_i. Given $\mathcal{A}$ and the input images $\mathcal{I}$, the model is required to correctly predict any intermediate state $\mathcal{S}_i$ in $\mathcal{T}$. Based on the index i, we define three task categories: **(1) Static Spatial Perception (SSP):** When $i = 0$, the task degenerates into a static spatial perception setting, where the model only needs to interpret $\mathcal{S}_0$ from $\mathcal{I}$ without any action execution. This setting serves as a baseline to compare against the two spatial simulation tasks in identical scenes. **(2) Final State Prediction (FSP):** When $i = n$, the model is required to predict the spatial state $\mathcal{S}_n$ after the execution of the whole action sequence $\mathcal{A}$, evaluating its ability to simulate the final outcome of multi-step spatial transformation. **(3) Dynamic Process Tracking (DPT):** Beyond final outcome prediction, we also evaluate whether a model can continuously simulate the spatial dynamics throughout the action sequence. Specifically, models are required to track arbitrary intermediate states $\mathcal{S}_i$ for $0 < i < n$, as well as the transitions $\Delta(\mathcal{S}_i, \mathcal{S}_{i+1})$ between them.

In concrete questions, we instantiate the state $\mathcal{S}_i$ as a set of object coordinates $\mathcal{P}_i$, the agent position $\mathbf{v}_i$, and the agent orientation vector $\mathbf{o}_i$:

$$\begin{aligned} \mathcal{S}_i &= (\mathcal{P}_i, \mathbf{v}_i, \mathbf{o}_i), \quad \mathbf{v}_i, \mathbf{o}_i \in \mathbb{R}^2 \\ \mathcal{P}_i &= \{\mathbf{p}_j\}_{j=1}^{M}, \quad \mathbf{p}_j \in \mathbb{R}^2 \end{aligned} \tag{2}$$

where M denotes the number of objects, and j indexes individual objects. We design different types of queries to evaluate models' understanding of these coordinates and orientations (see Figure 2 for examples): **(1) Distance-Object:** Identifying the object that is closest to or farthest from a target object or the agent. **(2) Object-Distance:** Measuring distance or distance changes between objects, or between an object and the agent. **(3) Direction-Object:** Identifying the object located in a specified direction relative to a target object or the agent. **(4) Object-Direction:** Measuring the relative direction or direction changes between objects, or between an object and the agent. In addition, for each question and action, we define two reference frames: allocentric (ALLO) and egocentric (EGO). The ALLO reference frame adopts a fixed global direction (e.g., North), while the EGO reference frame is defined by the agent's current orientation.

2.3 Data Curation Pipeline

We develop a sophisticated data curation pipeline to automatically generate high-quality question-answer pairs.

Environment Synthesis. Firstly, we synthesize environments using the Habitat-Sim simulator, which supports scalable simulation of diverse 3D scenes with accurate ground-truth spatial annotations and provides precise control over object and agent states. We collect 349 different scenes from the ReplicaCAD (Straub et al., 2019), ArchitecTHOR, iTHOR, ProcTHOR, and RoboTHOR (Kolve et al., 2017). Each scene contains a unique object layout that defines the initial state $\mathcal{S}_0$. To obtain visual inputs, we position the camera at a designated location and capture multiple images by rotating it within the scene. The camera orientation in the first image is used to define the fixed global north and the agent's initial forward orientation o_0.

Query Construction. For each scene, we randomly select either two objects or one object and the agent as query targets, and populate handcrafted templates to generate four types of queries across the three tasks. To accommodate different query formulations, Object-Distance queries are designed as open-ended questions, while the others adopt a five-choice multiple-choice format.

Model	SSP	FSP		DPT		Overall	
	Acc (%) ↑	Acc (%) ↑	Drop (%) ↓	Acc (%) ↑	Drop (%) ↓	Acc (%) ↑	Drop (%) ↓
Human Level†	71.80	76.95	-7.17	76.00	-5.85	74.73	-6.51
Proprietary Models							
Gemini-3.1-Pro†	**64.00**	**65.40**	**-2.19**	47.40	25.94	**58.93**	11.88
GPT-5†	53.60	38.40	28.36	**59.70**	**-11.38**	50.57	8.49
Gemini-3-Flash	45.28	44.23	2.32	46.27	-2.19	45.24	**0.06**
GPT-4o	38.84	32.26	16.94	20.49	47.23	30.85	32.09
Spatial Models							
Spatial-SSRL-7B	28.79	21.92	23.87	17.35	39.73	22.87	31.80
SpatialLadder-3B	21.29	18.34	13.87	18.15	14.73	19.30	14.30
SpatialThinker-3B	23.21	18.43	20.62	12.09	47.92	18.10	34.27
SpaceR-7B	32.01	24.15	24.57	18.41	42.51	25.07	33.54
Open-Source Models							
Qwen3-VL-32B	40.86	41.18	-0.78	25.61	37.31	36.19	18.26
InternVL3.5-38B	36.67	31.05	15.33	23.98	34.61	30.78	24.97
Gemma3-12B	31.02	29.62	4.52	22.43	27.68	27.85	16.10
Qwen2.5-VL-7B	25.46	24.99	1.84	23.22	8.80	24.59	5.32

Table 1: Evaluation on MentalGrid. The best result among all models is indicated in **bold** and the best result among open-source models is underlined. † indicates results on the MentalGrid-tiny set.

Action Specialization. After generating the queries, we randomly sample an action sequence $\mathcal{A}$ to construct complete FSP and DPT questions. Here, each $a_i \in \mathcal{A}$ is formulated as two types of function mapping f:

$$\begin{aligned} f_t(\mathbf{p}, \theta, d) &= \mathbf{p} + d \cdot \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} \\ f_r(\mathbf{o}_i, \theta) &= \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \mathbf{o}_i \end{aligned} \tag{3}$$

where f_t denotes the translation transformation and f_r denotes the rotation transformation. The variable $\mathbf{p}$ represents the position, d denotes the translation distance, and θ denotes the direction. We randomly sample action steps, transformation functions, distances, and directions to generate $\mathcal{A}$, while ensuring that most actions in $\mathcal{A}$ modify the states of queried targets.

Answer Generation. Finally, given a query and an action sequence, we compute the full state trajectory $\mathcal{T}$ from the initial state $\mathcal{S}_0$, from which the ground-truth answer to the query can be derived. We then generate negative options by randomly sampling incorrect directions or objects, completing the construction of each question-answer pair.

Quality Control. While the preceding steps ensure question validity, we implement a multi-stage refinement process to maximize data quality: **(1) Filtering Small Objects:** Objects occupying less than 0.1% of the image pixels are excluded to mitigate perceptual challenges such as occlusion and low resolution. **(2) Removing similar objects:** Since our tasks require each object to be unique, we remove scenes containing duplicated or semantically similar object categories. **(3) Resolving Spatial Overlap:** To prevent ambiguity caused by discarding height information in 2D coordinate projections, we exclude target objects whose spatial distances fall below a predefined threshold to avoid vertical overlap. We provide additional details and prompts in Appendix B.

3 Spatial Simulation Evaluation

3.1 Evaluation Setup

Evaluation Models. We conduct a comprehensive evaluation of 12 diverse MLLMs. For proprietary models, we consider Gemini-3.1-Pro (Google DeepMind, 2025b), GPT-5 (OpenAI, 2025), Gemini-3-Flash (Google DeepMind, 2025a) and GPT-4o (OpenAI, 2024). For open-source models, we evaluate Qwen3-VL-32B (Team, 2025), InternVL3.5-38B (Wang et al., 2025b), Gemma3-12B (Kamath et al., 2025) and Qwen2.5-VL-7B (Bai et al., 2025). For spatial models, we select SpatialLadder-3B (Li et al., 2025b), SpatialThinker-3B (Batra et al., 2025), Spatial-SSRL-7B (Liu et al., 2025b) and SpaceR-7B (Ouyang et al., 2025).

Human Performance. We extract 50 examples from each task to form a tiny test set. From it, we sample 30 questions at a time and assign them to one of five students, then calculate the average human performance.

Metrics. We use Accuracy (Acc) for multiple-

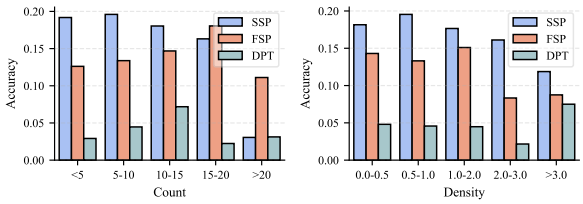
choice questions and Mean Relative Accuracy (MRA) for open-ended questions (Yang et al., 2025a). MRA averages performance across multiple confidence thresholds τ based on relative error:

$$\text{MRA} = \frac{1}{10}\sum_{\tau\in\mathcal{C}} \mathbb{I}\left(\frac{|\hat{y}-y|}{y} < 1-\tau\right) \tag{4}$$

where $\hat{y}$ denotes the model's prediction, y denotes the ground truth, and $\mathcal{C} = \{0.5, 0.55, ..., 0.95\}$. For brevity, we use "Accuracy" to refer to both metrics. To disentangle errors between spatial simulation and perception, we define a relative drop rate for each task $t \in \{\text{FSP}, \text{DPT}\}$ as $\text{Drop}_t = (\text{Acc}_{\text{SSP}} - \text{Acc}_t)/\text{Acc}_{\text{SSP}}$. A higher drop indicates greater difficulty in internal spatial simulation beyond static perception. Additional details and comprehensive results across a wider range of frontier models on both MentalGrid and MentalGrid-tiny are provided in Appendix C.

3.2 Main Results on MentalGrid

Table 1 shows overall model performance on MentalGrid. Our key observations are as follows: **(1) Current models leave significant room for improvement in spatial simulation.** Although advanced models like Gemini-3.1-Pro demonstrate relatively strong spatial simulation performance (58.93 % accuracy), Table 1 shows that they still lag behind human performance by **16%**. Besides, the majority of open-source models perform near the chance level, indicating significant limitations in their spatial simulation capabilities. **(2) Open-source models show weaker spatial simulation than spatial perception.** As shown in Table 1, most open-source models achieve their best performance on the SSP task, with large positive Drop values. In contrast, proprietary models such as Gemini-3-Flash exhibit relatively consistent performance across different tasks. This suggests that spatial reasoning in open-source models relies heavily on static perception and struggles to generalize when reasoning requires multi-step spatial state simulation. **(3) Process modeling is more challenging than outcome prediction.** In Table 1, most models exhibit larger performance degradation on DPT than on FSP. This observation indicates that accurate final-state prediction does not necessarily reflect a faithful understanding of the underlying intermediate spatial dynamics. Consequently, these results highlight the importance of strengthening the entire spatial simulation process, rather than optimizing predictions at isolated states alone. This finding serves as the foundation for our training framework in §4.1.



(a) Object Count (b) Spatial Density

Figure 3: Evaluation under different perception complexity on Qwen2.5-VL-3B.

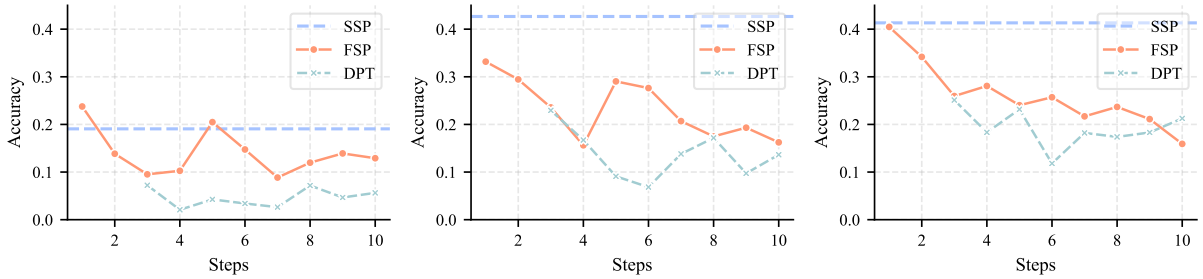
3.3 Impact of Perception Complexity

Spatial simulation first requires perceiving and localizing target objects in the initial scene. We therefore begin by analyzing how perception complexity influences model performance. Specifically, we group questions by object count and object density, and report accuracy within each group. Results for Qwen2.5-VL-3B are shown in Figure 3, with additional results provided in Appendix D.

For the object count, Figure 3a shows that SSP performance exhibits a sharp decay once the object count in the scene exceeds a critical threshold (e.g., > 20), uncovering a clear perceptual bottleneck. However, this degradation is less pronounced for FSP or DPT, suggesting that while increased object scale hampers initial perceptual grounding, spatial simulation performance is relatively less affected by it. To further quantify environmental complexity, we define Spatial Density (ρ) via the Mean Nearest Neighbor Distance (Clark and Evans, 1954):

$$\rho = \frac{1}{M}\sum_{i=1}^{M} \min_{j\neq i} \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{5}$$

where (x_i, y_i) is the position of each object. While traditional counts represent global magnitude, a higher ρ captures the local clumping of the environment. As shown in Figure 3b, increased density significantly degrades SSP accuracy due to heightened perceptual interference, whereas spatial simulation tasks remain relatively robust to such local crowding. Overall, spatial simulation is markedly less sensitive to scene complexity than spatial perception. This decoupling suggests that: while perceptual grounding constitutes the primary bottleneck for static perception tasks, **spatial simulation is constrained by additional critical factors beyond perception alone.**



(a) Raw Image (b) 2D Map Image (c) 2D Map Text

Figure 4: Evaluation under different action steps on Qwen2.5-VL-3B.

3.4 Impact of Simulation Complexity

We investigate how the complexity of the simulation process affects model performance. Given the known limitations of current models in 3D perception (Li et al., 2025b), we introduce two simplified input formats to reduce perceptual confounds in our evaluation: **(1) 2D Map Image:** We aggregate the coordinates of different objects across multiple scene images to construct a top-down 2D grid map. This input allows models to perform our spatial reasoning tasks without relying on 3D perceptual capabilities such as depth estimation or object detection. **(2) 2D Map Text:** We further include a JSON-based format that specifies the precise coordinates of all objects in the scene. This input format does not require any visual perception from the model and conveys spatial information purely through language. We demonstrate examples for each format in Figure 15.

We consider the number of action steps and the entropy of action distributions as measures of simulation complexity, and compare model performance under different input formats in Figures 4 and 16. From these results (refer to Appendix E for extended experiments), we establish the following conclusions: **(Finding 1) Perceptual Dominance in Low-Complexity Simulations.** As shown in Figure 4b and 4c, providing structured spatial representations (such as 2D map images or text) substantially improves SSP accuracy and yields modest gains for FSP and DSP in small-step regimes. This indicates that under limited simulation complexity, perceptual grounding still constrains spatial simulation performance. **(Finding 2) Long-range Reasoning as the Primary Bottleneck.** However, as the number of action steps increases, FSP performance in Figures 4b and 4c exhibits a pronounced decline, eventually converging to the baseline established by raw inputs in Figure 4a. This suggests that while explicit spatial grounding can mitigate initial perceptual uncertainty, the compounding difficulty of multi-step reasoning remains the fundamental limiting factor in spatial simulation. **(Finding 3) Models Exhibit Minimal Capability for Process-Oriented Spatial Simulation.** Even under short action sequences and with explicit coordinate inputs, DPT accuracy remains consistently low, indicating that models struggle to reason about intermediate spatial state transitions. Such results highlight an intrinsic limitation: current open-source models lack process-oriented spatial simulation capabilities and instead rely on shallow or outcome-focused spatial reasoning strategies.

4 Spatial Simulation Training

4.1 Our Training Framework

Guided by the analysis in §3, in this section, we propose a targeted training strategy to address current bottlenecks and enhance spatial simulation capabilities.

Training Data Construction. We first construct the training data following the pipeline described in § 2.3. In particular, when generating training environments, we exclude the ArchitecTHOR and iTHOR datasets, which are reserved for out-of-domain evaluation. From the remaining datasets, we generate a total of 16k QA pairs, including 5,368 for the static spatial perception, 5,289 for the final state prediction, and 5,294 for the dynamic process tracking. For each question, we prompt GPT-5-mini to think step-by-step and deduce the final answer. Chains of thought (CoT) with accuracy above 0.9 are retained as training trajectories.

Curriculum Learning Framework. Building upon the above training data, we design a three-stage curriculum training framework that systematically improves the model's spatial simulation abilities. **(Stage 1) Perceptual Grounding on Static Spatial Perception:** Motivated by Finding 1, in the first stage, we perform supervised fine-tuning

Model	In-domain				Out-of-domain				Overall
	SSP	FSP	DPT	Avg.	SSP	FSP	DPT	Avg.	
Proprietary Models									
Gemini-3-Flash	43.48	42.44	45.37	43.73	**47.74**	**46.72**	47.58	**47.34**	45.24
GPT-4o	36.81	32.23	18.68	29.50	41.61	32.30	23.11	32.74	30.85
Qwen2.5-VL-7B Based Spatial Models									
Qwen2.5-VL-7B	23.81	24.41	21.65	23.33	27.70	25.78	25.48	26.36	24.59
Spatial-SSRL-7B	26.79	24.41	16.61	22.75	31.50	18.46	18.41	23.02	22.87
SpaceR-7B	30.09	24.10	17.37	24.02	34.63	24.22	19.90	26.54	25.07
Qwen2.5-VL-3B Based Spatial Models									
Qwen2.5-VL-3B	19.43	12.12	4.35	12.16	17.11	16.22	4.92	13.06	12.54
SpatialThinker-3B	22.24	16.55	12.8	17.31	24.53	21.02	11.07	19.19	18.10
SpatialLadder-3B	21.18	17.38	18.45	19.02	21.44	19.67	17.73	19.69	19.30
SpatialWalker-3B (Ours)	**46.51**	**51.08**	**45.65**	**47.79**	39.41	45.24	**49.01**	44.35	**46.35**
Δ *Improvement*	+27.08	+38.96	+41.30	+35.63	+22.30	+29.02	+44.09	+31.29	+33.81

Table 2: Evaluation results on our MentalGrid benchmark, the highest performance is indicated in **bold**.

Model	MindCube				OmniSpatial					STARE	Overall
	Among	Around	Rotation	Avg.	DR.	SI.	CL.	PT.	Avg.		
SpatialLadder-3B	46.33	58.00	27.00	**45.43**	39.76	43.33	23.41	39.22	37.57	31.75	38.25
SpatialThinker-3B	40.17	**62.00**	33.50	44.10	42.14	47.33	25.00	41.35	40.05	32.33	38.83
Spatial-SSRL-7B	29.83	35.20	**35.50**	32.19	44.52	48.33	**26.19**	**44.03**	41.06	**42.07**	38.77
Qwen2.5-VL-7B	27.83	38.80	33.50	31.52	30.48	**49.00**	22.62	39.75	36.20	38.53	35.42
Qwen2.5-VL-3B	36.83	48.40	33.00	38.86	37.14	42.00	25.00	39.39	36.92	33.71	36.50
SpatialWalker-3B	**46.50**	51.60	34.00	45.33	**55.95**	43.67	25.79	40.11	**42.79**	36.14	**41.42**
Δ *Improvement*	+9.67	+3.20	+1.00	+6.47	+18.81	+1.67	+0.79	+0.72	+5.87	+2.43	+4.92

Table 3: Evaluation results on other benchmarks. Best results are **bolded** and second best are underlined.

on the SSP task to establish reliable perceptual grounding. This stage focuses exclusively on improving object localization and spatial grounding, enabling the model to accurately understand the corresponding spatial relations within static scenes. **(Stage 2) Synergistic Learning of Outcome and Process for Spatial Simulation:** Motivated by Findings 2 and 3, Stage 2 fine-tunes the model on joint FSP and DPT trajectories to address the bottleneck in multi-step state simulation. While FSP incentivizes accurate terminal outcome prediction, DPT enforces rigorous reasoning over intermediate transitions. This synergy enables the model to transition from shallow, outcome-biased heuristics to robust, process-aware spatial simulation. **(Stage 3) Fine-Grained State Evolution via Reinforcement Learning:** To further address the process-oriented deficiencies identified in Finding 3, we introduce a third stage centered on Reinforcement Learning (RL) using the DPT dataset. While Stage 2 establishes a supervised foundation, RL further refines the model's performance on state-evolution trajectories. We employ GRPO (Shao et al., 2024) for stable policy optimization, utilizing a format reward that enforces a structured reasoning template and an accuracy reward validates the predicted spatial state against the ground truth. By optimizing for this joint objective, the model learns to bridge the gap between static perception and dynamic simulation, fostering a more robust internal representation of spatial state trajectories. We report more training details in Appendix F.

4.2 Evaluation on MentalGrid

Table 2 compares SpatialWalker with state-of-the-art baselines across in-domain and out-of-domain scenes on the MentalGrid benchmark. Our model achieves an overall accuracy of **46.35%**, delivering a remarkable **+33.81%** improvement over the Qwen2.5-VL-3B backbone. Notably, SpatialWalker not only outperforms existing open-source spatial models but also surpasses proprietary models like Gemini-3-Flash, demonstrating that our multi-stage training effectively resolves the bottleneck in spatial simulation. Furthermore, SpatialWalker exhibits robust generalization, maintaining a **44.35%** average accuracy on out-of-domain scenes and achieving the highest accuracy in the challenging DPT task (**49.01%**), underscoring its superior ability to generalize spatial transition dynamics to unseen environments. We provide additional comparisons with test-time world-model-

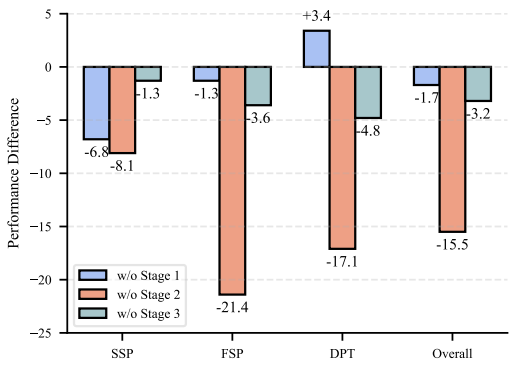


Figure 5: Ablation study results.

based approaches in Appendix G, where our model consistently achieves stronger performance.

4.3 Generalization to Other Benchmarks

To demonstrate that our model's performance extends beyond our benchmark, we evaluate its performance across a diverse range of tasks: MindCube (Yin et al., 2025) tests spatial mental modeling in real-world scenarios, OmniSpatial (Jia et al., 2025) evaluates comprehensive spatial reasoning abilities, and STARE (Li et al., 2025c) assesses multi-step visual simulation performance (see Appendix H for more benchmarks). As shown in Table 3, our SpatialWalker-3B achieves a superior overall score of **41.42%**, consistently outperforming other specialized spatial models. Most notably, our model excels in the dynamic spatial reasoning (DR.) task of OmniSpatial, reaching an **18.81%** gain over the base model. These results validate that our approach significantly enhances the model's transferable dynamic spatial reasoning capabilities across diverse scenarios, demonstrating strong generalization to spatial reasoning in real-world environments.

4.4 Ablation Study

Figure 5 reveals the critical contribution of each module within our framework. Stage 2 proves most essential, as its removal causes a 15.5% overall accuracy drop, validating simulation trajectory training as the cornerstone of dynamic spatial reasoning. Stage 3 contributes meaningfully with a 3.2% overall loss, confirming that policy refinement on sequential state transitions is vital for robust simulation. While excluding Stage 1 causes a 6.8% degradation in static perception, it leads to a 3.4% gain in DPT, suggesting a potential trade-off between static perceptual priors and the fluid reasoning required to track evolving process changes. Appendix I provides additional ablation results.

5 Related Work

5.1 Spatial Reasoning in MLLMs

Spatial reasoning in MLLMs centers on the capacity to perceive, understand, and mentally manipulate spatial relationships between objects. For evaluation, most studies represent object layouts using images or videos to assess models' understanding of geometric properties and spatial arrangements (Yang et al., 2025a; Li et al., 2025a; Yang et al., 2025b; Jia et al., 2025). For methods, existing works seek to bolster spatial reasoning by enhancing 3D visual perception, primarily focusing on improving perceptual capabilities through the integration of extra visual modules (Zhao et al., 2025; Wu et al., 2025a), tool augmentation (Wu et al., 2025b; Lee et al., 2025), or advanced training techniques (Li et al., 2025b; Batra et al., 2025). Our work shifts the focus from static perception toward dynamic spatial simulation.

5.2 Mental Simulation Evaluation

Existing research has extensively evaluated the mental simulation capabilities of models in spatial reasoning. Prior studies have investigated geometric manipulations, such as mental rotation and folding (Li et al., 2025c; Wang et al., 2025a), while others focus on scene perception across varying viewpoints (Li et al., 2025a; Yin et al., 2025). Furthermore, several works evaluate model performance in sequential planning and spatial navigation when provided with explicit start and end points (Yang et al., 2025a; Wang et al., 2024). However, these studies primarily operate within static environments, leaving a gap in understanding long-horizon transitions and evolving spatial states.

6 Conclusion

In this paper, we introduce MentalGrid, a benchmark evaluating the prediction of evolving spatial states without intermediate visual feedback. Our analysis reveals that existing models primarily struggle with spatial simulation. To address this, we develop a three-stage curriculum framework evolving from perceptual grounding to RL-based trajectory optimization. This approach yields SpatialWalker, a 3B model that achieves a **33.8%** improvement over its base, surpasses Gemini-3-Flash, and demonstrates robust generalization across out-of-domain benchmarks. Our findings underscore spatial simulation as a fundamental building block for broader spatial intelligence.

Limitations

Although our work conducts an in-depth evaluation and proposes training framework for improving MLLMs' spatial simulation abilities, it has several limitations. First, real-world spatial simulation may involve complex physical factors such as object pose, materials and collisions. Since existing open-source models lack such prior knowledge, and to reduce the noise introduced by these additional factors in the evaluation of pure spatial reasoning ability, our task, consistent with prior work on spatial reasoning, does not take these factors into account. Second, since our benchmark requires arbitrary rotation and translation operations on both objects and agents, collecting such data from real-world environments is both challenging and prohibitively expensive. Therefore, all of our scenes are generated from diverse simulated environments. Nevertheless, our experimental results demonstrate that the learned models generalize well to real-world scenarios. Future work may further explore and improve these two directions.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning through both a novel methodology and a new evaluation benchmark. Regarding our method, there are many potential societal consequences of our work, none which we feel must be specifically highlighted here. In constructing our benchmark, we ensure strict adherence to copyright and licensing regulations, explicitly avoiding data from sources that prohibit copying or redistribution. Furthermore, we have manually filtered the dataset to exclude any images containing private information or harmful content. The data within our benchmark are not intended to, nor are they capable of, replacing original data sources. Finally, we did not employ external crowdsourcing or paid annotation platforms. All participants contributed voluntarily with a full understanding of the research goals, procedures, and intended data use.

The Use of Large Language Models

In this study, a large language model (LLM) was employed as a tool to assist in the refinement and enhancement of the manuscript's language. The specific usages of the LLM include:

- **Grammar and Syntax Improvement:** The LLM helped to correct grammatical errors and improve sentence structures, contributing to greater clarity and fluency in the writing.
- **Conciseness and Precision:** It provided suggestions for more concise and precise wording, aiding in the refinement of certain sections without altering their meaning.

It is important to note that while the LLM contributed to the refinement of the manuscript's language, the research ideas, data analysis, and conclusions were independently conceived and developed by the authors. The LLM's contributions were exclusively related to text refinement and did not extend to the conceptual aspects of the study.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923.

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. 2022. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *Preprint*, arXiv:2111.08897.

Hunar Batra, Haoqin Tu, Hardy Chen, Yuanze Lin, Cihang Xie, and Ronald Clark. 2025. Spatialthinker: Reinforcing 3d reasoning in multimodal llms via spatial rewards. *arXiv preprint arXiv:2511.07403*.

Jacob LS Bellmund, Lorena Deuker, Tobias Navarro Schröder, and Christian F Doeller. 2016. Grid-cell representations in mental simulation. *Elife*, 5:e17089.

Ellis Brown, Arijit Ray, Ranjay Krishna, Ross B. Girshick, Rob Fergus, and Saining Xie. 2025. SIMS-V: simulated instruction-tuning for spatial video understanding. *CoRR*, abs/2511.04668.

Philip J Clark and Francis C Evans. 1954. Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology*, 35(4):445–453.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2432–2443. IEEE Computer Society.

Google DeepMind. 2025a. Gemini 3. Accessed: 2026-01-29.

Google DeepMind. 2025b. Gemini 3.1 pro. Accessed: 2026-03-16.

Danielle Harris. 2023. Spatial reasoning in context: bridging cognitive and educational perspectives of spatial-mathematics relations. In *Frontiers in Education*, volume 8, page 1302099. Frontiers Media SA.

Nora A Herweg and Michael J Kahana. 2018. Spatial representations in the human brain. *Frontiers in human neuroscience*, 12:297.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. 2025. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. *CoRR*, abs/2506.03135.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, and 191 others. 2025. Gemma 3 technical report. *CoRR*, abs/2503.19786.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, and 1 others. 2017. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Phillip Y Lee, Jihyeon Je, Chanho Park, Mikaela Angelina Uy, Leonidas Guibas, and Minhyuk Sung. 2025. Perspective-aware reasoning in vision-language models via mental imagery simulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

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

Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. 2025b. Spatialladder: Progressive training for spatial reasoning in vision-language models. *arXiv preprint arXiv:2510.08531*.

Linjie Li, Mahtab Bigverdi, Jiawei Gu, Zixian Ma, Yinuo Yang, Ziang Li, Yejin Choi, and Ranjay Krishna. 2025c. Unfolding spatial cognition: Evaluating multimodal models on visual simulations. *CoRR*, abs/2506.04633.

Zilu Liang, Simeng Wu, Jie Wu, Wen-Xu Wang, Shaozheng Qin, and Chao Liu. 2024. Distance and grid-like codes support the navigation of abstract social space in the human brain. *Elife*, 12:RP89025.

Weichen Liu, Qiyao Xue, Haoming Wang, Xiangyu Yin, Boyuan Yang, and Wei Gao. 2025a. Spatial reasoning in multimodal large language models: A survey of tasks, benchmarks and methods. *CoRR*, abs/2511.15722.

Yuhong Liu, Beichen Zhang, Yuhang Zang, Yuhang Cao, Long Xing, Xiaoyi Dong, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025b. Spatial-ssrl: Enhancing spatial understanding via self-supervised reinforcement learning. *arXiv preprint arXiv:2510.27606*.

OpenAI. 2024. Hello gpt-4o. Accessed: 2026-01-29.

OpenAI. 2025. Introducing gpt-5. Accessed: 2026-01-29.

Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. 2025. Spacer: Reinforcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, and 1 others. 2019. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*.

Yueqi Song, Tianyue Ou, Yibo Kong, Zecheng Li, Graham Neubig, and Xiang Yue. 2025. Visualpuzzles: Decoupling multimodal reasoning evaluation from domain knowledge. *CoRR*, abs/2504.10342.

Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, and 1 others. 2019. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*.

Qwen Team. 2025. Qwen3-vl technical report. *CoRR*, abs/2511.21631.

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. 2024. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. In *The Thirty-Eighth Annual Conference on Neural Information Processing Systems*.

Siting Wang, Minnan Pei, Luoyang Sun, Cheng Deng, Kun Shao, Zheng Tian, Haifeng Zhang, and Jun Wang. 2025a. Spatialviz-bench: An mllm benchmark for spatial visualization. *arXiv preprint arXiv:2507.07610*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025b. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *Preprint*, arXiv:2508.18265.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. 2025a. Spatial-mllm: Boosting MLLM capabilities in visual-based spatial intelligence. *CoRR*, abs/2505.23747.

Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. 2025b. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. *arXiv preprint arXiv:2506.09965*.

Yuxi Xiao, Longfei Li, Shen Yan, Xinhang Liu, Sida Peng, Yunchao Wei, Xiaowei Zhou, and Bingyi Kang. 2025. Spatialtree: How spatial abilities branch out in mllms. *arXiv preprint arXiv:2512.20617*.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2025a. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 10632–10643. Computer Vision Foundation / IEEE.

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. 2025b. Mmsi-bench: A benchmark for multi-image spatial intelligence. *CoRR*, abs/2505.23764.

Yuncong Yang, Jiageng Liu, Zheyuan Zhang, Siyuan Zhou, Reuben Tan, Jianwei Yang, Yilun Du, and Chuang Gan. 2025c. Mindjourney: Test-time scaling with world models for spatial reasoning. *CoRR*, abs/2507.12508.

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

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 9556–9567. IEEE.

Wanyue Zhang, Yibin Huang, Yangbin Xu, JingJing Huang, Helu Zhi, Shuo Ren, Wang Xu, and Jiajun Zhang. 2025a. Why do mllms struggle with spatial understanding? A systematic analysis from data to architecture. *CoRR*, abs/2509.02359.

Yuyou Zhang, Radu Corcodel, Chiori Hori, Anoop Cherian, and Ding Zhao. 2025b. Spinbench: Perspective and rotation as a lens on spatial reasoning in vlms. *CoRR*, abs/2509.25390.

Baining Zhao, Ziyou Wang, Jianjie Fang, Chen Gao, Fanhang Man, Jinqiang Cui, Xin Wang, Xinlei Chen, Yong Li, and Wenwu Zhu. 2025. Embodied-r: Collaborative framework for activating embodied spatial reasoning in foundation models via reinforcement learning. *CoRR*, abs/2504.12680.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Kejian Zhu, Zhuoran Jin, Hongbang Yuan, Jiachun Li, Shangqing Tu, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2025. MMR-V: what's left unsaid? A benchmark for multimodal deep reasoning in videos. *CoRR*, abs/2506.04141.

A Details of Our MentalGrid Benchmark

A.1 Dataset Statistics

Table 4 reports the key statistics of our MentalGrid benchmark. We calculate the average response length using the correct answers from GPT-5-mini.

Statistics	Number
Total Questions	3,356
Total Query Types	24
Total Images	1,100
Total Secnes	349
Task Type	
- Static Spatial Perception	1162 (34.62%)
- Final State Prediction	1143 (34.06%)
- Dynamic Process Tracking	1051 (31.32%)
Multi-choice Questions	2310 (68.83%)
Open-Ended Questions	1046 (31.17%)
Average Question Length	449.95
Average Response Length	1156.04

Table 4: Key statistics of MentalGrid.

A.2 Comparison with Other Benchmarks

To further distinguish MentalGrid from existing spatial reasoning benchmarks, we provide a detailed comparison in Table 5, highlighting its advantages in three key aspects: **(1) Data Integrity and Novelty:** Unlike many prior datasets that synthesize data from common sources like ScanNet (Dai et al., 2017) or ARKitScenes (Baruch et al., 2022), which carry a high risk of data contamination, our dataset generates scenes and objects within a custom simulation environment, ensuring spatial novelty and mitigating performance inflation. **(2) Dynamic Agent-Environment Interaction:** While current benchmarks are primarily restricted to limited viewpoint switching within static environments, MentalGrid features arbitrary agent perspectives and dynamic environmental state changes, necessitating that models move beyond static spatial perception to actively track evolving states. **(3) Long-Horizon Simulation Complexity:** Our simulation trajectories are designed with significantly greater depth, demanding sophisticated long-horizon spatial simulation capabilities to resolve complex tasks.

To further demonstrate the simulation complexity of our tasks, we compare the average reasoning path lengths of various models across MentalGrid and other existing mental simulation datasets in Figure 6. The results indicate that models consistently generate significantly longer trajectories on our dataset, which directly validates the higher simulation complexity of MentalGrid.

A.3 More Examples in MentalGrid

Figures 7, 8 and 9 present further examples from MentalGrid, spanning a wide range of environments and task types.

B Details of MentalGrid Data Curation Pipeline

The construction of the MentalGrid benchmark follows a rigorous four-stage pipeline designed to produce high-quality spatial simulation tasks.

B.1 Details of Environment Synthesis

The pipeline begins by loading realistic 3D environments and populating them with navigable agents and interactable objects. For each scene, the system identifies a valid traversable position by sampling random coordinates and verifying them through a pathfinding module. To ensure spatial diversity and avoid overlapping scenes, a minimum distance constraint of 2.0 m is enforced between consecutive sampling points. Once a location is secured, the agent captures a series of panoramic observations by rotating its camera across multiple viewpoints. The number of views and the degree of overlap between adjacent frames are randomized to simulate varied perceptual conditions. A critical step in this stage is scene purification, where any objects not visible from the initial set of viewpoints are programmatically removed from the environment. This ensures that the resulting scene metadata remains focused on the entities that the agent can perceive and reason about, maintaining a minimum object count for task complexity.

B.2 Details of Query Construction

Following environment synthesis, the pipeline establishes the objective of the spatial task. Two distinct entities are selected from the curated scene to serve as the subject and the reference point. To prevent geometric ambiguity and ensure the task is resolvable, the system enforces a minimum physical separation between these entities. Based on the task category, such as direction estimation or distance comparison, the pipeline selects a linguistic template from a library of natural language prompts.

Dataset	Size	Source	Image	View	Env.	Sim. Complexity
SpatialEval (Wang et al., 2024)	4.6K	S+E	Mixed	Static	Static	None
VSI-Bench (Yang et al., 2025a)	5K	E	Real-life	Fixed	Static	Low
ViewSpatial-Bench (Li et al., 2025a)	5.7K	E	Real-life	Fixed	Static	Low
MMSI-Bench (Yang et al., 2025b)	1K	E	Real-life	Fixed	Static	Low
MindCube (Yin et al., 2025)	21.1K	E	Real-life	Fixed	Static	Medium
OminiSpatial (Jia et al., 2025)	8.4K	E	Mixed	Fixed	Static	Medium
MentalGrid (Ours)	**3.5K**	**S**	**Simulation**	**Arbitrary**	**Dynamic**	**High**

Table 5: The comparison between MentalGrid and other existing benchmarks. For Data Source, S denotes synthetic data, while E represents data derived from existing sources. View and Env. indicate whether the agent's viewpoint and the environment state are static or dynamic. Sim. Complexity assesses the overall difficulty of the mental simulation required for each task.

This stage also determines the reference frame by choosing between an egocentric perspective relative to the agent or an allocentric perspective relative to the global coordinate system. The question templates for the generation of each task are listed in Table 6.

B.3 Details of Action Specialization

Once the query is established, the pipeline simulates environmental changes to transform the static scene into a dynamic trajectory. A sequence of actions ranging from one to ten steps is generated, with a minimum of three steps enforced for tasks involving process tracking. To ensure the actions are relevant to the subsequent query, the simulation prioritizes the movement of the target entities involved in the task. This creates a direct causal link between the action sequence and the evolving spatial states. Both the agent movements and the object displacements are interleaved and then synthesized into a chronological narrative utilizing varied instructional prefixes to simulate diverse natural language commands.

B.4 Details of Answer Generation

The final stage computes the ground truth by tracking the state of the environment throughout the simulation. For orientation tasks, continuous angular values are mapped onto a set of discrete categories such as 15° or 90° turns. In process tracking scenarios, the pipeline records the state at every intermediate step. To evaluate simulation integrity, specific states within the sequence are masked, requiring the model to infer these hidden snapshots from the overall action flow. To facilitate robust evaluation, the pipeline generates hard negative options sampled from neighboring spatial categories or by reversing the direction of movement. This ensures that a correct answer reflects true spatial simulation rather than simple elimination or heuristic-based guessing.

B.5 Details of Quality Control

The reliability and rigor of the MentalGrid benchmark are sustained through a multi-layered quality control framework integrated into the data curation pipeline. During the environment synthesis phase, spatial diversity is guaranteed by enforcing a strict minimum distance constraint between sampling points, which ensures that each scene represents a unique sub-region and prevents data redundancy. Furthermore, the scene purification process programmatically eliminates occluded or distant objects that are absent from the agent's initial set of viewpoints, thereby ensuring that all entities involved in the subsequent simulation are properly grounded in the initial observation. To maintain the reasoning complexity of the tasks, samples that fail to meet a minimum visible object count are automatically discarded. In the action generation stage, the pipeline prioritizes the movement of queried objects through a targeted sampling ratio, which creates a robust causal link between the observed dynamics and the final spatial question. For task evaluation, the generation of hard negatives involves strategic sampling from neighboring spatial categories and direction reversals, which effectively prevents the model from relying on simple elimination or language-based heuristics. Finally, the use of diverse instructional prefixes and randomized reference frames ensures that the benchmark remains resilient to template-level over-fitting while maintaining high linguistic variability.

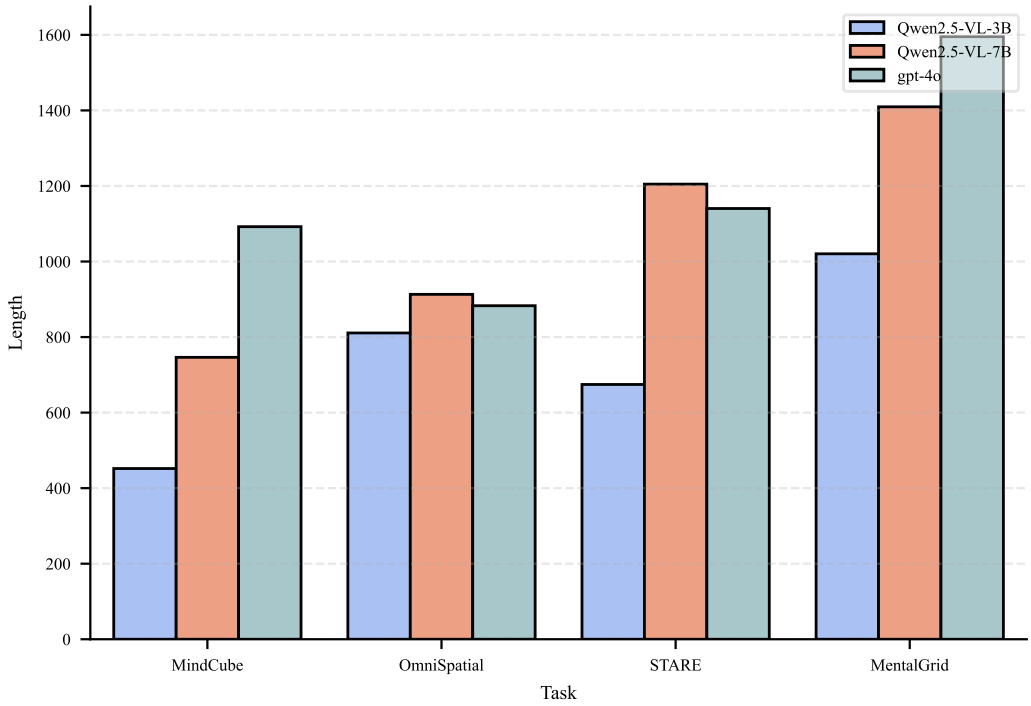


Figure 6: Response length comparison on different benchmarks.

C Details of Main Evaluation Experiments

C.1 Experimental Setup

We utilize the vLLM framework (Kwon et al., 2023) deployed on 8 NVIDIA RTX 3090 GPUs for the inference of all open-source models. To ensure reproducibility, we adopt a greedy decoding strategy for all models unless otherwise specified. Specifically, the temperature is set to 0 and both top-p and top-k are set to 1. The specific system prompts used are detailed in Figure 10.

C.2 More Evaluation Results

In this section, we provide a comprehensive set of evaluation results on our benchmark. This includes the full evaluation results across all question types as presented in Table 7. Furthermore, we report the complete results on MentalGrid-tiny in Table 8 to facilitate a direct comparison with human performance.

D Details of Perception Complexity Experiments

Here, we introduce the image count as a new factor to reflect perceptual complexity. As illustrated in Figure 11, when fewer input images are provided (e.g., a single image), models achieve noticeably higher performance on static spatial perception. Conversely, performance on spatial simulation tasks (FSP and DPT) does not exhibit a corresponding improvement as the number of input images decreases. This suggests that the difficulty of mental simulation is decoupled from the density of the visual input stream. Moreover, we extended the analytical experiments from §3.3 to Qwen2.5-VL-7B, Gemma3-12B, and GPT-4o. The results, reported in Figure 12, 13, and 14, provide further empirical support for the conclusions presented in the main text.

E Details of Simulation Complexity Experiments

In addition to the number of action steps, we introduce Action Entropy as a metric to quantify the complexity and diversity of the simulation. For a given sequence of actions $\mathcal{A}$, we define each unique operator based on the combination o of its reference frame and movement direction. Let $\mathcal{O}$ be the set of unique operators occurring in a trajectory of length N. The probability of an operator $o \in \mathcal{O}$ is calculated as $p(o) = \frac{count(o)}{L}$. The action entropy

is then defined as:

$$H(A) = -\sum_{o \in \mathcal{O}} p(o) \log_2 p(o)$$

A higher entropy indicates a more varied and less predictable sequence of spatial transitions, representing a more challenging simulation scenario. We report the accuracy of Qwen2.5-VL-3B under different input configurations as a function of action entropy. As illustrated in Figure 16, when coordinate-based inputs are introduced, models achieve strong spatial simulation performance at low action entropy, but exhibit a pronounced degradation as entropy increases. This trend is consistent with the results reported in §3.4, further supporting our findings. We repeat this experiment on Qwen2.5-VL-7B to further demonstrate the generalizability of our findings (see Figure 17, 18).

F Details of Spatial Simulation Training

F.1 Details of Training Data Construction

The specific system prompts used for CoT generation are detailed in Figure 19.

F.2 Details of Training Implementation

For Stage 1 and Stage 2, we implement SFT using the LLaMA-Factory framework (Zheng et al., 2024). The models are optimized with the hyperparameter configurations in Table 9. For Stage 3, we implement GRPO using the Verl framework (Sheng et al., 2024). The corresponding reinforcement learning configurations are provided in Table 10. All experiments are conducted on 4 NVIDIA A100 GPUs.

G Additional Spatial Simulation Comparison

We reproduce the MindJourney-SVC and MindJourney-SWM methods as inference-time world-model augmentation approaches on MentalGrid-tiny, using GPT-4o as the backbone model (Yang et al., 2025c). The results in Table 11 indicate that, although methods like MindJourney incur significantly lower computational cost compared to our training approach, their performance falls far short of what is achieved with our curriculum training. Based on comparisons with direct CoT reasoning using GPT-4o, we conclude that in our more complex spatial reasoning tasks, the quality of images generated by existing world models is insufficient, introducing substantial noise that hinders model reasoning.

H Additional Out-of-domain Benchmark Evaluation

We further compare SpatialWalker with its base model, Qwen2.5-VL-3B, on two other spatial reasoning benchmarks: MulSeT(Zhang et al., 2025a) and SpinBench (Zhang et al., 2025b). The results are demonstrated in Table 12, which show that our method maintains strong generalization across a broader range of benchmarks.

I Additional Ablation Study

To further validate the effectiveness of our proposed multi-stage training strategy, we conduct two additional ablation studies: **(1) Reverse:** We swapped the execution order of Stage 1 and Stage 2 to investigate the necessity of our sequential training curriculum. **(2) Mix:** We integrated the training objectives of both Stage 1 and Stage 2 into a single-stage joint training process to compare against our decoupled design. The experimental results in Figure 20 indicate that performance declines across all tasks under both ablation settings, demonstrating the necessity of our progressive design.

Static Spatial Perception

Object-Distance (ALLO)

Provide a numerical estimate (m) of the distance from doorway to countertop.

Answer: 1.4

Object-Direction (EGO)

From your viewpoint, where is the countertop located in terms of direction?

(A) backward *(B) right-forward* (C) left (D) left-backward (E) forward

Distance-Object (EGO)

Which of the listed objects is located the closest to you?

(A) countertop (B) soapbottle *(C) toilet* (D) faucet (E) mirror

Direction-Object (ALLO)

Which of the listed objects is located in the north of faucet?

(A) mirror (B) countertop (C) doorway (D) faucet (E) soapbottle

Figure 7: Examples in MentalGrid (Part 1).

Final State Prediction

Object-Distance (EGO)

Starting from the reference view and performing: At step 1, the indoor_plant moves backward by 1.7 meters. At step 2, the indoor_plant moves forward by 1.8 meters. At step 3, the indoor_plant moves north by 0.7 meters. At step 4, the indoor_plant moves south by 1.4 meters. At step 5, the indoor_plant moves west by 0.4 meters. At step 6, you move forward by 0.5 meters. \nPlease estimate the final distance between you and the indoor_plant (in meters).

Answer: 3.9

Distance-Object (EGO)

Starting from the reference view and performing: At step 1, you turn left. At step 2, the towel moves west by 1.8 meters. At step 3, you move forward by 1.2 meters. At step 4, you turn left. At step 5, the bike moves to your left by 2.3 meters. At step 6, you move forward by 1.5 meters. At step 7, the umbrella moves west by 1.8 meters. At step 8, you turn left. At step 9, the bike moves to your left by 0.8 meters. \nAfter the movements, which object ends up closest to you?

(A) table *(B) pan* (C) umbrella (D) bin (E) indoor_plant

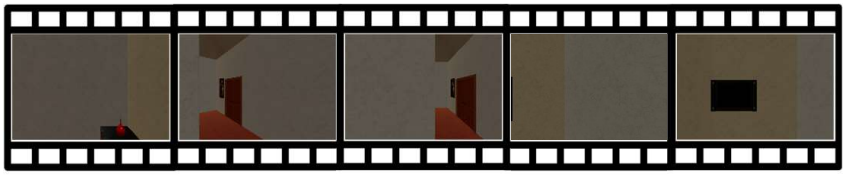
Direction-Object (ALLO)

Given the following movement sequence from the reference view: At step 1, the towel moves south by 1.2 meters. At step 2, the towel moves west by 0.4 meters. At step 3, the pan moves west by 0.7 meters. At step 4, the towel moves forward by 1.4 meters. At step 5, the towel moves to your left by 1.9 meters. At step 6, the towel moves north by 0.4 meters. At step 7, the towel moves to your right by 1.8 meters. At step 8, the towel moves to your right by 1.9 meters. \nOnce the actions conclude, which object occupies the northwest region with respect to the towel?

(A) bin (B) table (C) shelf (D) indoor_plant (E) bowl

Figure 8: Examples in MentalGrid (Part 2).

Dynamic Process Tracking



Object-Direction (ALLO)

From the reference view, suppose the following movements occur: At step 1, the painting moves west by 1.4 meters. At step 2, the painting moves south by 1.1 meters. At step 3, the painting moves to your left by 2.2 meters. At step 4, the painting moves north by 1.1 meters. At step 5, the doorway moves forward by 1.9 meters. Throughout the motion, the relative directional angle from the doorway toward the painting evolves step by step as: [MASK] -> Counter-clockwise rotation by about 15° -> Clockwise rotation by about 5° -> Clockwise rotation by about 15° -> [MASK]\nChoose the correct option for the [MASK] entries.

(A) Counter-Clockwise rotation by about 15° , No noticeable rotation (about 0°) *(B) No noticeable rotation (about 0°), Counter-clockwise rotation by about 15°* (C) No noticeable rotation (about 0°), No noticeable rotation (about 0°) (D) No noticeable rotation (about 0°), Counter-clockwise rotation by about 45° (E) Counter-Clockwise rotation by about 15° , Counter-clockwise rotation by about 45°

Object-Distance (ALLO)

Given the following movement sequence from the reference view: At step 1, the painting moves forward by 2.4 meters. At step 2, the dresser moves east by 1.8 meters. At step 3, the dresser moves forward by 0.8 meters. At step 4, the painting moves to your right by 0.9 meters. At step 5, the painting moves to your left by 2.3 meters. At step 6, the dresser moves to your right by 1.4 meters. \nDuring the dynamic process, the incremental change in distance (m) between dresser and painting per step is:\n+0.8 -> +1.6 -> [MASK] -> -0.9 -> +2.2 -> [MASK]\nFill each [MASK] slot with the proper value.

Answer: -0.2, +1.4

Figure 9: Examples in MentalGrid (Part 3).

Task	QueryType	Reference	Question Template
SSP	Object-Distance	EGO	*How far are you from the {target_obj} in meters?*
		ALLO	*How far is the {target_obj} from the {view_obj} (in meters)?*
	Distance-Object	EGO	*Which object lies farthest from your position? {options}*
		ALLO	*Which object lies farthest from the {target_obj}? {options}*
	Object-Direction	EGO	*In which direction is the {target_obj} positioned relative to you? {options}*
		ALLO	*In which direction is the {target_obj} located relative to the {view_obj}? {options}*
	Direction-Object	EGO	*Which object lies in your {direction}? {options}*
		ALLO	*Which object is located in the {direction} of the {target_obj}? {options}*
FSP	Object-Distance	EGO	*{action_text} What is the final distance (meters) between you and the {target_obj}?*
		ALLO	*{action_text} What is the final distance (meters) between the {target_obj} and the {view_obj}?*
	Distance-Object	EGO	*{action_text} After the movements, which object ends up farthest from you? {options}*
		ALLO	*{action_text} After the movements, which object ends up farthest from the {target_obj}? {options}*
	Object-Direction	EGO	*{action_text} After the movements, in which direction does the {target_obj} end up relative to you? {options}*
		ALLO	*{action_text} After the movements, in which direction does the {target_obj} end up relative to the {view_obj}? {options}*
	Direction-Object	EGO	*{action_text} After the movements, which object ends up in your {direction}? {options}*
		ALLO	*{action_text} After the movements, which object ends up in the {direction} of the {target_obj}? {options}*
DPT	Object-Distance	EGO	*{action_text} During the movement sequence, the per-step change in distance (meters) between you and {target_obj} is: {masked_seq}*
		ALLO	*{action_text} During the movement sequence, the per-step change in distance (meters) between {target_obj} and {view_obj} is: {masked_seq}*
	Distance-Object	EGO	*{action_text} During the movement, the objects furthest from you, in sequence, are: {masked_seq} {options}*
		ALLO	*{action_text} During the movement, the objects furthest from the {target_obj}, in sequence, are: {masked_seq} {options}*
	Object-Direction	EGO	*{action_text} During the movement process, the changes in the relative direction angle from you to the {target_obj} at each step are: {masked_seq}*
		ALLO	*{action_text} During the movement process, from your perspective, the changes in the relative direction angle from the {view_obj} to the {target_obj} at each step are: {masked_seq}*
	Direction-Object	EGO	*{action_text} Over the course of the movement, the object that falls within your {direction}, in sequence, is: {masked_seq}*
		ALLO	*{action_text} During the movement process, the object occupying the {direction} of the {target_obj}, in sequence, is: {masked_seq}*

Table 6: Question Templates for tasks in MentalGrid.

System Prompt for Model Inference

You are a spatial-reasoning assistant.

<RULE>
1. You are an embodied observer represented by the camera.
2. You may receive EITHER:
   - a single image, OR
   - multiple images captured at the SAME time step from different viewpoints of the SAME environment.
3. If only ONE image is provided:
   - Your position corresponds to the camera viewpoint of that image.
4. If MULTIPLE images are provided:
   - Image 1 defines the reference frame.
   - All other images provide complementary viewpoints and depict the same static moment.
5. At the initial reference frame, your facing direction equals the camera viewing direction of the reference image (Image 1) and is defined as North (+Y-axis) in a world-fixed coordinate system.
6. Relative directions (forward, backward, left, right) are always defined with respect to your CURRENT facing direction and rotate together with it.
</RULE>

<TASK>
You will receive:
1. **Image(s)** - one or more RGB frames depicting the same scene at a single time step.
2. **Actions (Optional)** - a natural language sequence of actions describing the movements of you (i.e., the camera) and the objects. We ignore any blockages while moving.
3. **Question** - a natural-language query about spatial relationships between objects in the image.
4. **Options (Optional)** - 5 answer candidates, each tagged by a capital letter (A, B, C, D, E).
</TASK>

<REQUIREMENTS>
- Always ground your answer in the visual evidence; do not hallucinate unseen objects.
- If uncertain, pick the most plausible option—never refuse or reply "insufficient information."
</REQUIREMENTS>

Figure 10: The prompt used for model inference.

Model	Static Spatial Perception				Final State Prediction				Dynamic Process Tracking				Overall
	dir-obj	dis-obj	obj-dir	obj-dis	dir-obj	dis-obj	obj-dir	obj-dis	dir-obj	dis-obj	obj-dir	obj-dis	
GPT-5-mini	42.57	54.20	42.98	38.54	41.53	43.51	55.87	43.55	27.02	42.24	66.08	41.39	45.70
Gemini-3-Flash	42.08	49.24	47.85	41.60	33.88	46.18	50.72	41.72	22.18	38.79	61.95	50.68	45.24
GPT-4o	43.56	47.71	32.66	35.62	25.68	35.50	38.11	27.42	19.35	29.31	16.52	22.24	30.85
Gemma3-12B	27.72	33.59	28.37	33.64	28.96	27.48	31.81	29.37	33.06	26.72	21.83	14.01	27.85
InternVL3.5-8B	29.70	39.69	36.39	30.97	27.87	22.90	34.38	26.45	20.56	25.86	21.24	16.93	27.84
SpaceR-7B	31.68	32.44	24.93	38.97	24.59	23.66	20.63	27.79	26.21	22.41	23.60	6.45	25.07
Qwen2.5-VL-7B	25.74	31.30	29.80	16.56	27.87	26.34	27.79	19.66	31.05	23.28	28.91	12.07	24.59
Spatial-SSRL-7B	28.71	35.50	26.36	26.22	17.49	24.43	26.65	17.62	21.77	22.41	22.42	7.57	22.87
SpatialLadder-3B	20.79	25.57	21.78	17.88	14.75	18.70	15.19	23.09	14.92	31.03	26.25	8.28	19.30
SpatialThinker-3B	28.22	26.34	26.07	15.10	25.68	23.28	26.36	3.04	18.15	23.28	12.68	3.46	18.10
Qwen2.5-VL-3B	18.32	28.63	18.62	10.72	13.11	18.32	22.35	2.35	6.85	7.76	3.54	2.93	12.54

Table 7: Evaluation results on each query type of MentalGrid.

Model	SSP	FSP		DPT		Overall	
	Acc (%) ↑	Acc (%) ↑	Drop (%) ↓	Acc (%) ↑	Drop (%) ↓	Acc (%) ↑	Drop (%) ↓
Human Level	71.80	76.95	-7.17	76.00	-5.85	74.73	-6.51
Gemini-3.1-Pro	64.00	65.40	-2.19	47.40	25.94	58.93	11.88
Gemini-3-Flash	60.00	52.00	13.33	51.50	14.17	54.50	13.75
GPT-5	53.60	38.40	28.36	59.70	-11.38	50.57	8.49
GPT-4o	50.00	35.00	30.00	18.70	62.60	34.57	46.30
Gemma3-12B	38.40	34.40	10.42	21.60	43.75	31.47	27.08
InternVL3.5-8B	50.80	27.80	45.28	15.40	69.69	31.33	57.48
SpaceR-7B	35.40	21.20	40.11	20.10	43.22	25.57	41.67
Qwen2.5-VL-7B	30.20	15.00	50.33	26.80	11.26	24.00	30.79
Spatial-SSRL-7B	29.20	14.40	50.68	20.00	31.51	21.20	41.10
SpatialLadder-3B	21.60	16.60	23.15	13.80	36.11	17.33	29.63
SpatialThinker-3B	17.00	14.40	15.29	14.80	12.94	15.40	14.12
Qwen2.5-VL-3B	15.80	16.40	-3.80	4.00	74.68	12.07	35.44

Table 8: More evaluation results on MentalGrid-tiny.

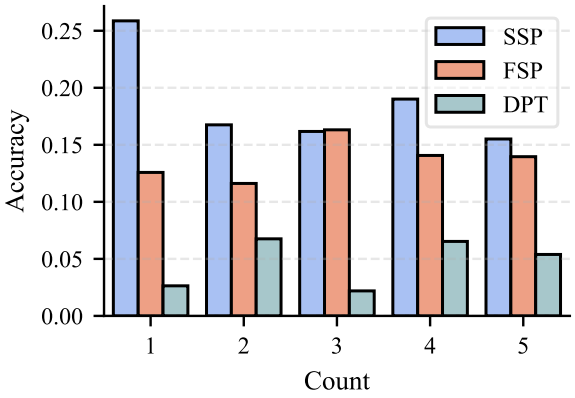


Figure 11: Evaluation under different image count on Qwen2.5-VL-3B.

Hyperparameter	**Setting**
Optimizer	AdamW
Learning rate	1×10^{-5}
LR scheduler	Cosine
Max sequence length	8192
Batch size per device	4
Gradient accumulation steps	4
Training epochs	2.0
Warmup ratio	0.1
Mixed precision	BF16
DeepSpeed strategy	ZeRO-3
Trainable Modules	**Status**
Vision Tower	Frozen
Multi-modal Projector	Frozen
Language Model	Full Fine-tuning

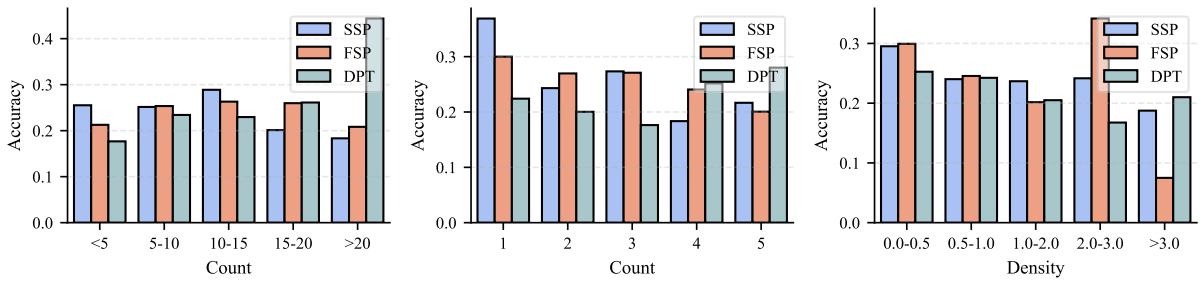
Table 9: Hyperparameters for Stage 1 and Stage 2 (SFT).

Hyperparameter	**Setting**
Algorithm	GRPO
Learning rate	1×10^{-6}
Group size	4
KL divergence coefficient	0.01
KL loss type	Low-variable KL
Entropy coefficient	0
Max prompt length	8192
Max response length	1024
Training batch size	64
PPO mini-batch size	16
Micro-batch size per GPU	1
Total epochs	2.0
Mixed precision	BF16
Parallelism strategy	FSDP + TP ($size = 2$)

Table 10: Hyperparameters for Stage 3 (RL).

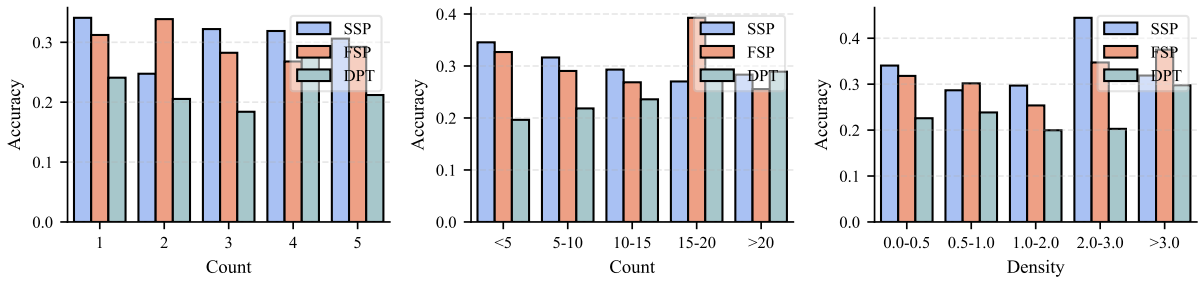
Model	**SSP**	**FPT**	**DPT**	**All**
MindJourney-SVC	36.60	30.80	22.80	30.10
MindJourney-SWM	37.20	29.40	16.00	27.50
GPT-4o	<u>50.00</u>	<u>35.00</u>	<u>18.70</u>	<u>34.60</u>
SpatialWalker	**52.00**	**53.20**	**41.70**	**49.00**

Table 11: Evaluation results on MentalGrid-tiny. Best results are **bolded** and second best are <u>underlined</u>.



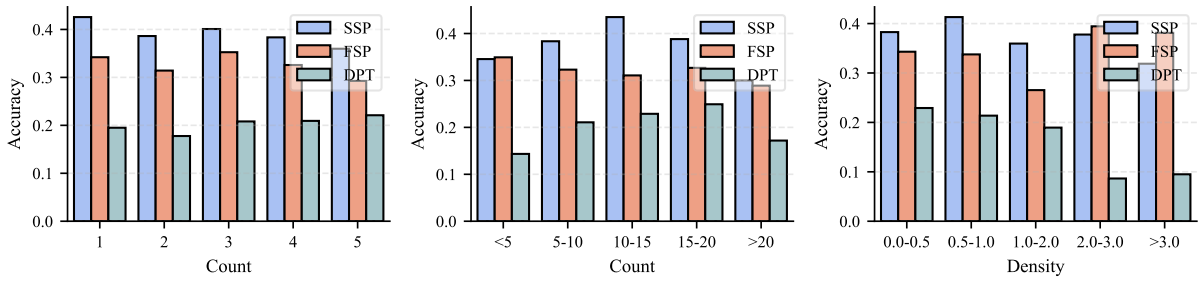
(a) Image Count (b) Object Count (c) Spatial Density

Figure 12: Comparison of methods under different perception complexity on Qwen2.5-VL-7B.



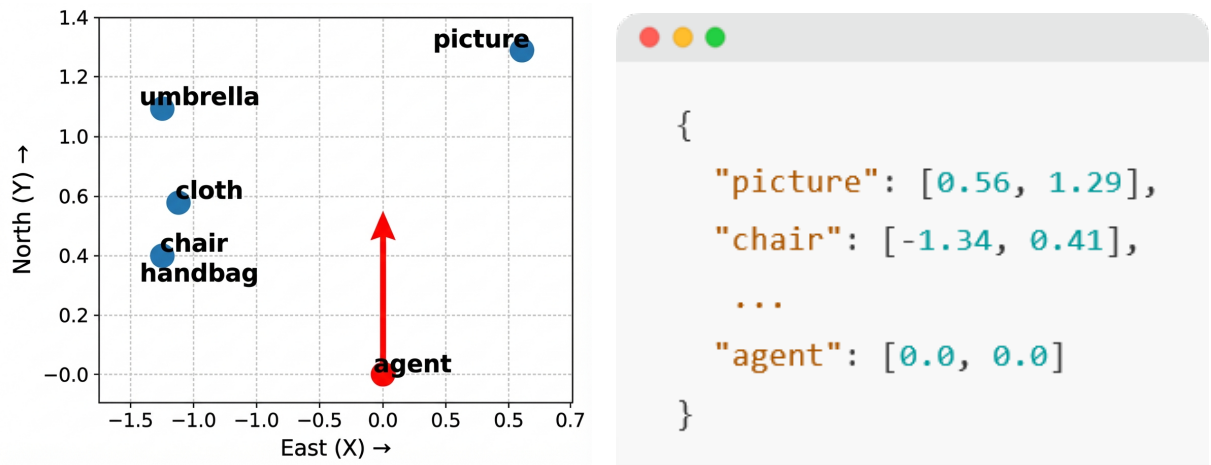
(a) Image Count (b) Object Count (c) Spatial Density

Figure 13: Comparison of methods under different perception complexity on Gemma3-12B.



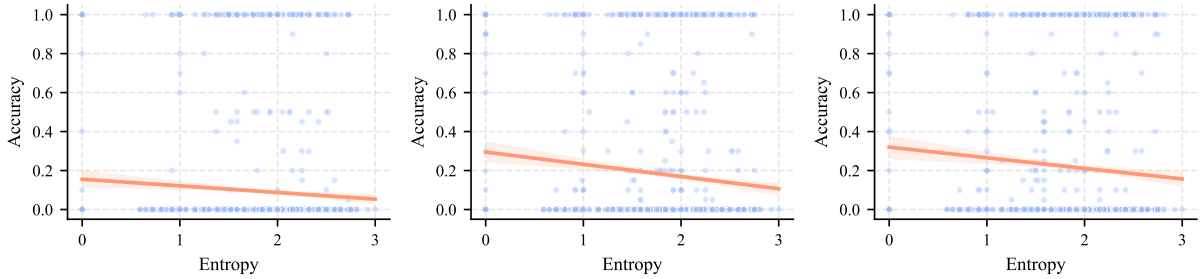
(a) Image Count (b) Object Count (c) Spatial Density

Figure 14: Comparison of methods under different perception complexity on GPT-4o.



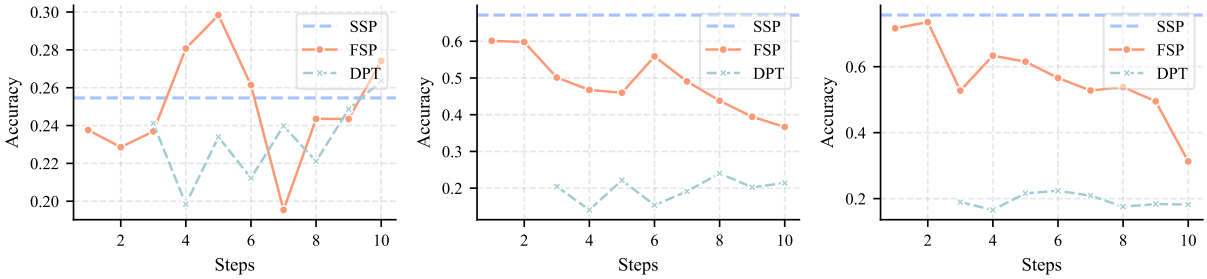
(a) 2D Map Image (b) 2D Map Text

Figure 15: Two simplified input formats for perception.



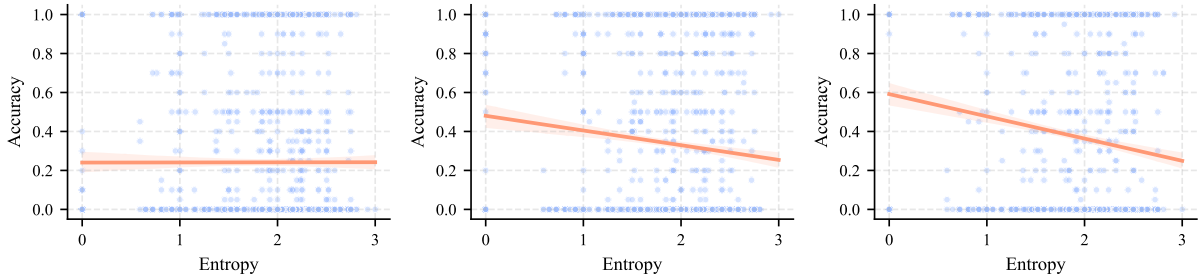
(a) Raw Image (b) 2D Map Image (c) 2D Map Text

Figure 16: Evaluation under different action entropy on Qwen2.5-VL-3B.



(a) Raw Image (b) 2D Map Image (c) 2D Map Text

Figure 17: Evaluation under different action steps on Qwen2.5-VL-7B.



(a) Raw Image (b) 2D Map Image (c) 2D Map Text

Figure 18: Evaluation under different action entropy on Qwen2.5-VL-7B.

Model	**MulSeT**			**SpinBench**							
	OR.	DC.	AT.	CVS.	IM.	DR.	PT.	ORG.	DT.	MR.	All
Qwen2.5-VL-3B	10.02	27.00	3.29	34.36	27.41	**42.78**	43.23	49.06	50.00	19.23	40.59
SpatialWalker	**26.47**	**42.15**	**33.72**	**46.65**	**46.91**	26.91	**46.98**	**82.55**	**52.56**	19.23	**52.40**
Δ ***Improvement***	+16.45	+15.15	+30.43	+12.29	+19.50	-15.87	+3.75	+33.49	+2.56	+0.00	+11.81

Table 12: Evaluation results on MulSeT and SpinBench. Best results are **bolded**.

System Prompt for CoT Generation

You are a spatial-reasoning assistant.

<RULE>
1. You are an embodied observer represented by the camera.
2. You may receive EITHER:
   - a single image, OR
   - multiple images captured at the SAME time step from different viewpoints of the SAME environment.
3. If only ONE image is provided:
   - Your position corresponds to the camera viewpoint of that image.
4. If MULTIPLE images are provided:
   - Image 1 defines the reference frame.
   - All other images provide complementary viewpoints and depict the same static moment.
5. At the initial reference frame, your facing direction equals the camera viewing direction of the reference image (Image 1) and is defined as North (+Y-axis) in a world-fixed coordinate system.
6. Relative directions (forward, backward, left, right) are always defined with respect to your CURRENT facing direction and rotate together with it.
</RULE>

<TASK>
You will receive:
1. **Image(s)** - one or more RGB frames depicting the same scene at a single time step.
2. **Actions (Optional)** - a natural language sequence of actions describing the movements of you (i.e., the camera) and the objects. We ignore any blockages while moving.
3. **Question** - a natural-language query about spatial relationships between objects in the image.
4. **Options (Optional)** - 5 answer candidates, each tagged by a capital letter (A, B, C, D, E).
5. **Hint** - Initial object coordinates (Origin: camera in reference image; +Y: viewing direction; +X: camera's right).
</TASK>

<REQUIREMENTS>
- You SHOULD fully utilize the information in hint to form accurate internal spatial representations and improve reasoning quality.
- During reasoning and explanation, you MUST behave as if no such hints were provided. Do NOT mention, reference, or allude to hints.
- All reasoning must be expressed as if it were derived naturally from visual perception and embodied spatial understanding.
- Always ground your answer in the visual evidence; do not hallucinate unseen objects.
</REQUIREMENTS>

Figure 19: The prompt used for CoT generation.

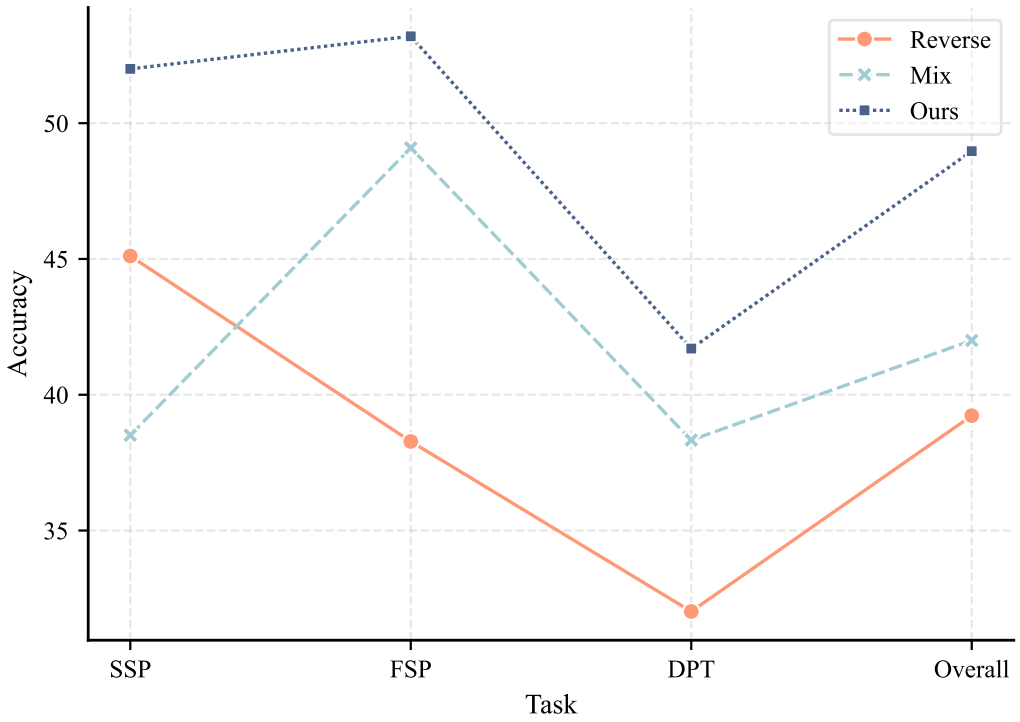


Figure 20: Additional ablation study results.