# SpatialBlock: Enhancing Spatial Intelligence in LVLMs via Synthetic Block-Stacking Problem

**Anonymous authors**
Paper under double-blind review

## Abstract

Large Vision-Language Models (LVLMs) have achieved strong performance on diverse visual tasks, yet their ability to reconstruct and reason about the 3D structure of the scene depicted in 2D images – referred to as *spatial intelligence* – remains limited. Existing approaches attempt to address this gap by using real-scene spatial question answering datasets that require dense geometric annotations. However, constructing such labels is costly, time-consuming, and often noisy due to reliance on external perception modules. In this work, we propose a novel paradigm inspired by human cognitive development: learning foundational spatial skills through structured block-manipulation tasks. We introduce **SpatialBlock-15k**, a synthetic dataset of 15,000 block-stacking problems covering 3D-to-2D projection, viewpoint transformation, and structural combination. The dataset further incorporates controlled color modulation as visual cues to encourage anchor-based reasoning in visually complex conditions. Experiments demonstrate that LVLMs trained on our dataset through either direct answering or reasoning-based prediction significantly outperform baselines and generalize to real-world spatial tasks, despite the dataset's synthetic and compact nature.

## 1 Introduction

Large Vision–Language Models (LVLMs) (Wang et al., 2024; Bai et al., 2025a; Zhu et al., 2025) have demonstrated remarkable performance across a wide range of 2D image understanding and reasoning tasks. Despite this progress, their ability to mentally reconstruct the 3D structure of a scene from 2D images – namely, *spatial intelligence* – remains limited. This shortcoming poses a fundamental bottleneck for deploying LVLMs in real-world applications that require robust spatial reasoning, such as autonomous driving (Tian et al., 2025) and robotics (Feng et al., 2025).

To address this limitation, prior work has extended model architectures by adding a specialized module for learning 3D features (Cheng et al., 2024; Wu et al., 2025) or designed tasks that directly require spatial reasoning (Yang et al., 2025; Yin et al., 2025) to encourage models to learn such capabilities. However, these approaches typically rely on real-world scene question answering datasets that require dense geometric and object-level annotations of the visual context. Constructing such dense labels is not only time-consuming and costly, but also noisy due to the frequent reliance on external modules (*e.g.*, segmentation or depth estimation models), thereby limiting their reliability and scalability.

In contrast to methods that depend on annotation-heavy real-scene data, we pursue a more efficient alternative inspired by the way humans develop spatial intelligence. Developmental psychology (Levine et al., 2012; Jirout & Newcombe, 2015) suggests that foundational spatial abilities are often cultivated through structured manipulative tools. Specifically, block play has been widely shown to enhance essential components of spatial cognition, including spatial integration and mental simulation (Caldera et al., 1999; Casey et al., 2008). Despite its simplicity, state-of-the-art LVLMs struggle with such problems, suggesting that they lack even fundamental levels of spatial intelligence (Fig. 1).



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

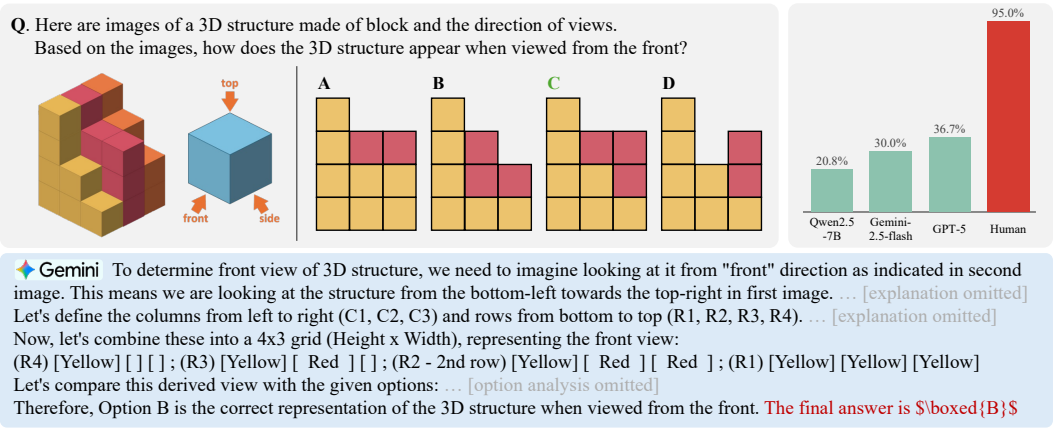



Figure 1: **Performance Gap Between Humans and LVLMs in Block-Stacking Problem**. The top row shows a block-stacking task that requires predicting a 2D projection from a given 3D structure. While humans achieve near-perfect accuracy, state-of-the-art LVLMs struggle even with basic spatial reasoning problems. Leading proprietary models exhibit flawed spatial logic, often producing incorrect interpretations of the structure.

Motivated by this observation, we propose **SpatialBlock-15k**, a novel synthetic dataset of block-stacking problems designed to improve spatial intelligence. The dataset comprises three categories of spatial reasoning tasks: (1) 3D-to-2D projection, (2) viewpoint transformation, and (3) structural combination. To better approximate real-world spatial reasoning scenarios, where humans interpret scenes by anchoring on a specific object and reasoning relationships relative to it, we introduce color modulation as a visual cue, encouraging models to identify task-relevant elements in visually complex scenes. Unlike costly real-scene datasets, our dataset can be generated synthetically, enabling clean, scalable, and cost-effective data construction.

Experimental results show that LVLMs trained on SpatialBlock-15k, with either direct answer prediction or reasoning-based prediction, significantly outperform baseline models, even though the training data are entirely synthetic and relatively small in scale. These findings suggest that targeted training on fundamental spatial reasoning tasks can effectively enhance the spatial intelligence of LVLMs and provide a scalable alternative to annotation-heavy real-scene supervision.

In summary, our contribution is three-fold:

- We introduce a new perspective on enhancing spatial intelligence in LVLMs by shifting from annotation-heavy real-scene supervision to foundational spatial skill learning through structured synthetic tasks.
- We present **SpatialBlock-15k**, a scalable synthetic dataset of block-stacking problems spanning 3D-to-2D projection, viewpoint transformation, and structural combination, enhanced with controlled color cues for task-relevant reasoning.
- We show that training LVLMs on SpatialBlock-15k significantly improves spatial reasoning performance and generalizes effectively to real-world scenes, despite relying solely on synthetic data.

## 2 Related Work

### 2.1 Visual Spatial Reasoning in LVLMs

Recent efforts to improve spatial intelligence in LVLMs focus on building a dedicated encoder or proposing a specialized training strategy. On the architectural side, some approaches modify the model by introducing spatial tokens into the vision encoder (Tong et al., 2024; Lou et al., 2025), incorporating depth-aware plugin modules (Cheng et al., 2024), or adding spatial encoders to learn 3D features (Wu et al., 2025). Other research focuses on the training process, designing hierarchical schemes that transition from spatial perception to

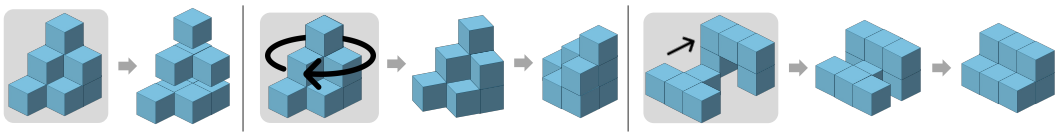



Figure 2: **Spatial Abilities Underlying Block-Stacking Reasoning**. Solving block-stacking tasks involves (a) inferring latent 3D composition, (b) reasoning under transformation, and (c) integrating multiple components into a coherent structure.

complex reasoning (Ma et al., 2025a; Li et al., 2026; Liu et al., 2026) or encouraging the model to output cognitive maps that encode object position and orientation (Yang et al., 2025; Yin et al., 2025). Despite their effectiveness, these approaches generally require real-scene annotations or pseudo-3D signals from external modules, which limits their scalability and robustness.

### 2.2 Training Datasets for Spatial Intelligence

Various training datasets have been proposed to enhance the spatial intelligence of LVLMs, ranging from simple spatial relations to complex multi-step reasoning. Early benchmarks (Chen et al., 2024; Ma et al., 2025b) leverage external models (*e.g.*, object detection, segmentation, depth, or pose estimation) to extract 3D information and generate low-level QA pairs such as distance, orientation, and spatial relations. To foster higher-level logic, recent works (Yang et al., 2025; Ouyang et al., 2025) use 3D annotated video data to construct relative relation and direction reasoning, or more sophisticated tasks such as route planning or spatio-temporal appearance ordering through human annotation. MindCube (Yin et al., 2025) focuses on spatial mental modeling, which requires synthesizing a full scene from partial views or inferring arrangements under perspective shifts. Despite these advancements, a significant bottleneck remains: these approaches rely on dense, real-scene 3D annotations, which are labor-intensive and inherently noisy, necessitating costly human annotation.

## 3 SpatialBlock-15k

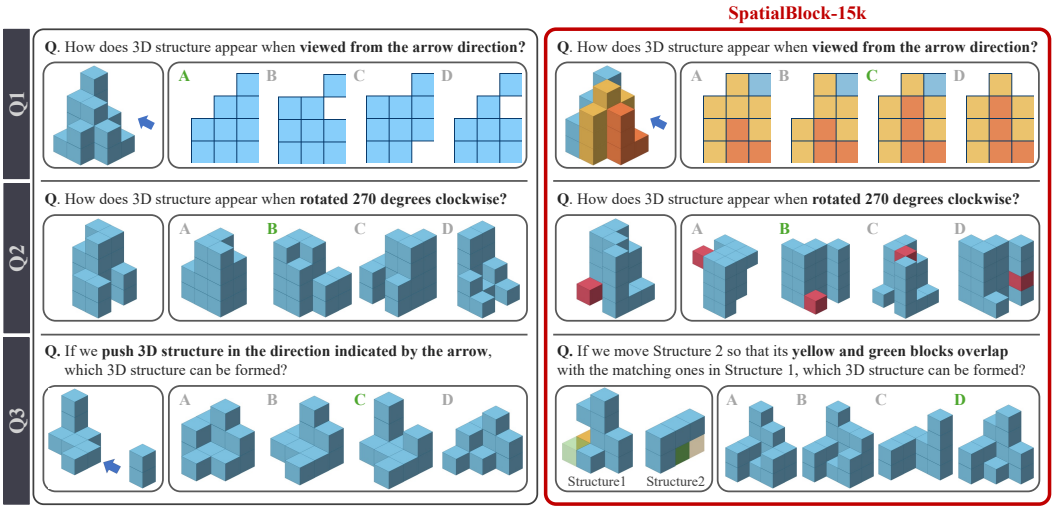
We introduce **SpatialBlock-15k**, a scalable synthetic dataset designed to systematically probe and enhance foundational spatial abilities, moving away from the conventional annotation-intensive paradigm that relies on dense real-scene 3D labels. Specifically, we utilize block-stacking problems, which play a pivotal role in the early stages of human spatial cognitive development (Sec. 3.1). Furthermore, we incorporate controlled color variations as visual cues to better reflect real-world spatial reasoning conditions (Sec. 3.2).

### 3.1 Block-Stacking Problems for Spatial Intelligence

Activities involving block-stacking require structural spatial understanding that extends beyond superficial visual pattern recognition. As illustrated in Fig. 2, solving such problems requires (a) spatial composition, the ability to reconstruct the complete 3D structure by deconstructing the visible configuration and inferring occluded blocks that must exist to support it; (b) mental simulation, the capacity to simulate the outcome of transformations applied to a 3D structure; and (c) spatial integration, the ability to anticipate the resulting structure when multiple components are combined. Together, these abilities reflect core dimensions of spatial intelligence.

Building on these three abilities, we define three question types, each targeting a distinct aspect of spatial reasoning (Fig. 3(a)):

**Q1: 3D-to-2D projection** This task requires predicting the 2D appearance of a 3D structure from a specific viewing direction. Solving it demands the ability to internally reconstruct the 3D configuration and mentally transform it to the target viewpoint. The model must additionally perform depth-aware occlusion reasoning to determine which blocks remain visible in the final 2D projection.




**(a) Block-Stacking Problems** **(b) Visual Cue-based Extension**

Figure 3: **SpatialBlock-15k**. (a) Three block-stacking tasks: 3D-to-2D projection (Q1), viewpoint transformation (Q2), and structural combination (Q3). (b) Visual cue-based extension with controlled color variations, where colors encode depth ordering (Q1), preserved structural roles across transformations (Q2), and correspondence cues for integration (Q3).

**Q2: Viewpoint transformation** The problem inquires how a given 3D structure appears under self-rotation or viewpoint changes. The model should preserve structural consistency across transformations, maintaining correct relative positions of all components without distortion, disappearance, or unintended displacement.

**Q3: Structural combination** This question asks how the resultant integration is formed when two 3D structures are combined. It requires simulating how individual components interact, identifying contact interfaces, and subsequently inferring the coherent global structure that emerges from their integration.

Collectively, these question types form the foundation of our SpatialBlock-15k dataset, providing a structured framework for evaluating and training spatial reasoning in LVLMs.

### 3.2 Visual Cue-based Extension

In the previous section, we constructed tasks that require understanding 3D structures composed of single-color blocks. Since color does not provide a discriminative feature, these tasks demand that the model focus on purely geometric relationships, such as the relative positions of blocks and appearance changes under viewpoint transformations.

However, real-world scenes typically consist of complex environments with diverse objects, rather than visually uniform ones. Given this intricacy, humans tend to analyze the global scene configurations by selecting a specific object as an anchor and interpreting relationships relative to it (Land & Hayhoe, 2001). That is, visually salient cues serve as the reference points to infer changes and relations.

Motivated by this cognitive characteristic, we extend the monochromatic block-based tasks introduced in Sec. 3.1 by incorporating color as an additional cue. Here, color is not used merely for visual diversity but as a functional guidance that reveals depth information, structural correspondences, and contact regions. This design encourages the model to reconstruct the global structure and track transformations by using specific blocks as reference points. Specifically, we extend the three question types as follows (Fig. 3(b)):

**Q1: 3D-to-2D projection with depth cues** To facilitate depth-aware reasoning, different colors are assigned based on depth from a given viewpoint. Thus, the task requires not only matching the silhouette but also tracking front–back relationships

using color information. Color functions as an explicit indicator of depth ordering and is designed to clearly reflect the hierarchical organization of the 3D structure.

**Q2: Viewpoint transformation with anchor blocks** Among the blocks composing the 3D structure, one anchor block is assigned a distinct color. This block must maintain the same structural role before and after transformation. Accordingly, the task requires not only recognizing geometric structure but also understanding anchor-relative relationships that are preserved through transformation.

**Q3: Structural combination with overlapping blocks** The attachment location between two structures is represented not by arrows but by the overlap of colored blocks indicated by transparent regions. Each structure is presented as a separate image, so aligning the positions of same-colored blocks across images is needed to predict the combined structure. Color thus acts as a shared reference point linking the two inputs, thereby encouraging cross-structure integration.

This visual cue–based extension is designed to maintain the requirement for understanding the 3D structure itself while additionally enabling the learning of scene organization and relational reasoning centered around anchor objects, as required in realistic visual contexts. Consequently, it leads to meaningful improvement in real-scene spatial reasoning performance, as empirically demonstrated in Sec. 5.3.2. Finally, the SpatialBlock-15k dataset comprises three question types with visual cue–based extensions, each containing 5k samples. Further details on data construction can be found in Sec. A.

# 4 Method

To foster the spatial intelligence of LVLMs using SpatialBlock-15k, we propose two training strategies, targeting different aspects of the model's capabilities: (1) direct answer prediction, which focuses on mapping visual inputs directly to their corresponding answers (Sec. 4.1), and (2) reasoning-based prediction, which encourages the model to explain its intermediate logical paths that lead to a final answer (Sec. 4.2).

## 4.1 Direct Answer Prediction Model

To maximize the model's capacity for instantaneous spatial problem-solving, we train it with supervision only on the ground-truth answer sequence $y = \{y_1, y_2, ..., y_T\}$. Specifically, given a textual query $q$ and image $v$, we optimize model parameters $\theta$ to predict the next token $y_i$, given the preceding context $y_{(1:i-1)}$, by minimizing standard cross-entropy loss:

$$\mathcal{L}_{\text{ce}}(\theta) = -\sum_i \log P\left(y_i \mid y_{(1:i-1)}, q, v\right). \tag{1}$$

This strategy directly optimizes final answer prediction, thereby establishing a foundation for high-fidelity spatial problem-solving proficiency.

## 4.2 Reasoning-based Prediction Model

To go beyond simple answer prediction, we adopt a second strategy that enables the model to articulate its internal logic. To build a solid foundation for spatial intelligence, we first initialize the model through supervised fine-tuning to establish basic task-solving ability, then use reinforcement learning to enable higher-order structural reasoning.

For model initialization, we train the model with Low-Rank Adaptation (LoRA) instead of full-parameter fine-tuning, as full fine-tuning often degrades the inherent reasoning ability. This lightweight adaptation maintains the model's Chain-of-Thought (CoT) reasoning capabilities while aligning it with our dataset. Notably, we deviate from the conventional "cold-start" phase that relies on synthesized trajectories from larger teacher models (Guo et al., 2025; Li et al., 2026). The rationale behind this decision is that even state-of-the-art proprietary models exhibit relatively low accuracy on our dataset (Tab. 1), making it infeasible to construct reliable high-quality trajectories. Our empirical analysis (Sec. 5.3.4)

demonstrates a clear performance advantage in preserving the model's reasoning ability through LoRA-based tuning. This approach provides a more effective initialization for RL training than attempting to recover the thinking ability via a low-quality cold-start phase.

After LoRA-based initialization, we directly optimize the model using reinforcement learning via Group Relative Policy Optimization (GRPO). We design a multi-objective reward function that evaluates response correctness and reasoning trace quality with three components:

$$\mathcal{R}(y, o) = \mathcal{R}_{\text{acc}}(y, o) + \mathcal{R}_{\text{format}}(o) + \mathcal{R}_{\text{len}}(o), \tag{2}$$

where $o$ is a model-generated prediction given a question $c = (q, v)$ with ground-truth answer $y$. The accuracy reward $\mathcal{R}_{\text{acc}}$ verifies whether the final answer extracted from $o$ matches $y$.

Next, the format reward $\mathcal{R}_{\text{format}}$ enforces the required CoT structure, with a comprehensive reasoning trace followed by the final answer enclosed within `<answer>` and `</answer>` tags. Lastly, to prevent overly brief or excessively verbose output, a length reward $\mathcal{R}_{\text{len}}$ requires the response length $L$ to satisfy $50 < L < 1024$. For simplicity and stability, all components are implemented in binary rewards (0 or 1) based on whether each condition is satisfied.

Given this reward, for each question $c$, the old policy model $\pi_{\theta_{\text{old}}}$ samples a group of candidate responses $\{o_1, o_2, ..., o_G\}$. Each response $o_i$ receives a reward $r_i = \mathcal{R}(y, o_i)$, from which we compute the advantage $A_i = \frac{r_i - \text{mean}(r_1, r_2, ..., r_G)}{\text{std}(r_1, r_2, ..., r_G)}$. The model is then updated by maximizing the following objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{c,\{o_i\}} \left[ \frac{1}{G} \sum_{i=1}^{G} \min \left( \frac{\pi_\theta(o_i|c)}{\pi_{\theta_{\text{old}}}(o_i|c)} A_i, \text{clip} \left( \frac{\pi_\theta(o_i|c)}{\pi_{\theta_{\text{old}}}(o_i|c)}, 1 \pm \varepsilon \right) A_i \right) - \beta \text{KL}[\pi_\theta || \pi_{\text{ref}}] \right], \tag{3}$$

where $\varepsilon$ and $\beta$ are hyper-parameters, and $\text{KL}[\pi_\theta || \pi_{\text{ref}}]$ is the KL divergence between the policy model $\pi_\theta$ and reference model $\pi_{\text{ref}}$.

# 5 Experiments

We conduct multiple experiments to validate the effectiveness of SpatialBlock-15k in fostering spatial intelligence. We first describe the experimental setup, including training details and evaluation protocols (Sec. 5.1). We then compare our models with prior methods across several spatial reasoning benchmarks (Sec. 5.2). Next, we perform ablation studies to assess the effects of each component in dataset construction and training (Sec. 5.3). Lastly, we provide an analysis of the generalizability of our proposed dataset (Sec. 5.4).

## 5.1 Experimental Settings

We adopt Qwen2.5-VL-3B (Bai et al., 2025b), Qwen2.5-VL-7B (Bai et al., 2025b), Qwen3-VL-4B (Bai et al., 2025a), and InternVL3-2B (Zhu et al., 2025) as baseline models for our framework. To distinguish the training strategy, we append a suffix to the model name: *SpatialBlock-direct* denotes the model that outputs the answer directly (Sec. 4.1), while *SpatialBlock-reason* refers to the model that performs reasoning before producing the final answer (Sec. 4.2). Both variants are trained using the entire SpatialBlock-15k dataset.

To evaluate the spatial reasoning capability, we conduct experiments on five benchmarks spanning both in-domain and out-of-domain settings. For in-domain evaluation, we introduce SB-Bench, the held-out test split of SpatialBlock-15k, comprising 600 questions. For out-of-domain evaluation, we assess generalization to real-world scenarios on four benchmarks: MindCube (Yin et al., 2025), MMSI-Bench (Yang et al., 2026), and SPBench (Li et al., 2026) for spatial reasoning, and MMMU (Yue et al., 2024) for general visual perception. Across all benchmarks, we restrict evaluation to multiple-choice questions to align with our training answer format. Additional numerical results are provided in Sec. C.1.

## 5.2 Main Results

Tab. 1 shows our method's performance on spatial reasoning benchmarks. Notably, despite being trained with only 15k synthetic SpatialBlock samples, our model consistently outper-

Table 1: **Main Results**. For each dataset, **bold** and underlined indicate the best and second-best performance, respectively. * represents in-domain results where the model was trained on the benchmark. We re-evaluated all previous works for fair comparison.

| Model | # Data | In-domain | Out-of-domain | | | | |
|---|---|---|---|---|---|---|---|
| | | SB-Bench | MindCube | MMSI-Bench | SPBench | MMMU | Overall |
| ***Proprietary Models*** | | | | | | | |
| GPT-5 (Singh et al., 2025) | - | 43.8 | 56.7 | 42.8 | 43.0 | 77.6 | 55.0 |
| Gemini-2.5-flash (Comanici et al., 2025) | - | 36.4 | 51.0 | 32.2 | 44.3 | 58.8 | 46.6 |
| Claude-Sonnet-4.5 (Anthropic, 2025) | - | 46.3 | 34.0 | 32.7 | 43.9 | 69.2 | 43.9 |
| ***Open-Source Models*** | | | | | | | |
| LLaVA-OneVision-7B (Li et al., 2024) | - | 23.8 | 39.0 | 28.4 | 43.4 | 42.6 | 39.2 |
| InternVL3-8B (Zhu et al., 2025) | - | 22.7 | 42.7 | 25.1 | 46.1 | 48.0 | 40.5 |
| Llama-3.2-11B-Vision-Instruct (Dubey et al., 2024) | - | 23.0 | 29.5 | 24.5 | 28.2 | 40.7 | 30.7 |
| ***Spatial Specialist Models*** | | | | | | | |
| *Qwen2.5-VL-3B Based Models* | | | | | | | |
| Qwen2.5-VL-3B-Instruct (Bai et al., 2025b) | - | 29.9 | 39.1 | 26.1 | 40.0 | 45.6 | 37.7 |
| SpaceQwen (Chen et al., 2024) | 2B | 30.6 | 38.8 | 24.5 | 40.8 | 45.6 | 37.4 |
| SpatialMLLM (Wu et al., 2025) | 120K | 30.0 | 36.3 | 25.4 | 40.8 | 40.4 | 35.7 |
| SpatialLadder-3B (Li et al., 2026) | 25K | 25.5 | 46.4 | 26.0 | 76.6* | **47.3** | 39.9 |
| Spatial-SSRL-3B (Liu et al., 2026) | 81K | 29.5 | 41.0 | 24.4 | 19.5 | 45.2 | 32.5 |
| **SpatialBlock-3B-direct (Ours)** | 15K | **94.8** | **49.1** | 28.1 | 40.4 | 46.4 | 41.0 |
| **SpatialBlock-3B-reason (Ours)** | 15K | 90.2 | 45.3 | **29.2** | **43.7** | 46.7 | **41.2** |
| *Qwen2.5-VL-7B Based Models* | | | | | | | |
| Qwen2.5-VL-7B-Instruct (Bai et al., 2025b) | - | 20.8 | 31.3 | 27.4 | 46.7 | 46.6 | 38.0 |
| SpaceR (Ouyang et al., 2025) | 150K | 26.8 | 31.6 | 27.4 | **53.1** | 55.4 | 41.9 |
| Spatial-SSRL-7B (Liu et al., 2026) | 81K | 27.7 | 35.1 | 27.5 | 23.4 | 48.9 | 33.7 |
| **SpatialBlock-7B-direct (Ours)** | 15K | **95.0** | **48.8** | 28.4 | 48.7 | 45.2 | 42.8 |
| **SpatialBlock-7B-reason (Ours)** | 15K | 77.5 | 37.9 | **29.7** | 50.9 | **55.5** | **43.5** |
| *Qwen3-VL-4B Based Models* | | | | | | | |
| Qwen3-VL-4B-Instruct (Bai et al., 2025a) | - | 17.0 | 26.2 | 27.1 | 44.5 | 47.5 | 36.3 |
| Spatial-SSRL-4B (Liu et al., 2026) | 81K | 34.8 | 34.6 | **28.3** | 23.1 | **50.5** | 34.1 |
| **SpatialBlock-4B-direct (Ours)** | 15K | **95.7** | **51.3** | 27.8 | **45.9** | 47.2 | **43.1** |
| **SpatialBlock-4B-reason (Ours)** | 15K | 94.5 | 41.1 | **28.3** | 43.7 | 49.6 | 40.7 |
| *InternVL3-2B Based Models* | | | | | | | |
| InternVL3-2B-Instruct (Zhu et al., 2025) | - | 19.0 | 32.1 | 27.0 | 33.0 | 41.3 | 33.4 |
| **SpatialBlock-2B-direct (Ours)** | 15K | **97.2** | **51.4** | 27.3 | 41.3 | 42.4 | **40.6** |
| **SpatialBlock-2B-reason (Ours)** | 15K | 78.3 | 38.9 | **27.8** | **45.6** | **47.3** | 39.9 |

forms existing spatial specialists on real-scene benchmarks. This result highlights that our block-stacking formulation provides an effective training signal for spatial reasoning.

For **SpatialBlock-direct** models, we observe substantial gains, particularly on MindCube, which requires mental simulation. Among Qwen-based models, the 3B model outperforms the state-of-the-art SpatialLadder by 2.7%, and the 7B model improves upon its backbone by 17.5%. The 4B model achieves an even larger gain of 25.1% over its backbone, reaching 51.3%. Beyond Qwen, our method also improves InternVL3, demonstrating its generalizability across model scales and families.

The reasoning-enhanced variant, **SpatialBlock-reason**, further demonstrates strong results on MMSI-Bench, which requires complex logical inference. The 3B model outperforms SpatialLadder-3B by 3.2% using a simpler training pipeline based solely on synthetic block-stacking data, whereas the 7B model outperforms SpaceR and Spatial-SSRL with only 15K training samples. Beyond final-answer accuracy, our model also improves reasoning quality, achieving a higher alignment score of 21.1 with ground-truth reasoning steps, compared to 17.8 for the baseline. (See Sec. B.4 for details.) On SPBench, most of our models improve relative spatial reasoning despite such questions not being explicitly included in the training data, demonstrating strong generalization beyond the supervised task distribution.

In summary, direct models excel on benchmarks that require canonical 90-degree viewpoint transformations, while reasoning models perform better on tasks requiring diverse viewpoint changes and multi-image reasoning. Also, despite being trained without any real-scene images, our models maintain stable performance on the general visual understanding benchmark MMMU, indicating that SpatialBlock training enhances 3D spatial reasoning without degrading broader visual perception.

### 5.3 Ablation Studies

#### 5.3.1 Ablation on Question Types

We evaluate models trained on a single question type (Tab. 2(a)). Each question type targets a distinct aspect of spatial reasoning: Q1 focuses on viewpoint changes, Q2 em-

Table 2: **Ablation Studies**. * denotes in-domain results from benchmark-trained models.

| Model | # Data | In-domain | Out-of-domain | | | | |
|---|---|---|---|---|---|---|---|
| | | SB-Bench | MindCube | MMSI-Bench | SPBench | MMMU | Overall |
| Qwen2.5-VL-3B-Instruct (Bai et al., 2025b) | - | 29.9 | 39.1 | 26.1 | 40.0 | 45.6 | 37.7 |
| SpatialBlock-3B-direct (Ours) | 15K | 94.8 | 49.1 | 28.1 | 40.4 | 46.4 | 41.0 |
| SpatialBlock-3B-reason (Ours) | 15K | 90.2 | 45.3 | 29.2 | 43.7 | 46.7 | 41.2 |
| *(a) Ablation on question types* | | | | | | | |
| SpatialBlock-3B-reason (Q1 only) | 15K | 48.0 | 41.8 | 27.6 | 42.8 | 43.2 | 38.8 |
| SpatialBlock-3B-reason (Q2 only) | 15K | 51.5 | 38.8 | 24.2 | 42.1 | 47.1 | 38.0 |
| SpatialBlock-3B-reason (Q3 only) | 15K | 40.8 | 41.3 | 25.3 | 43.2 | 46.5 | 39.1 |
| *(b) Ablation on visual cues* | | | | | | | |
| SpatialBlock-3B-direct (w/o visual cues) | 15K | 61.7 | 41.2 | 26.2 | 38.4 | 46.6 | 38.1 |
| SpatialBlock-3B-reason (w/o visual cues) | 15K | 55.8 | 38.0 | 24.4 | 44.8 | 48.8 | 39.0 |
| *(c) Ablation on task design* | | | | | | | |
| SpatialBlock-3B-direct (SpatialLadder-26k) | 15K | 24.7 | 44.5 | 27.2 | 72.1* | 45.3 | 47.3 |
| SpatialBlock-3B-reason (SpatialLadder-26k) | 15K | 28.0 | 44.9 | 27.5 | 68.8* | 44.7 | 46.5 |
| *(d) Ablation on model initialization* | | | | | | | |
| SpatialBlock-3B-reason (No Init.) | 15K | 67.8 | 41.3 | 25.4 | 43.5 | 46.7 | 39.2 |
| SpatialBlock-3B-reason (Full-SFT Init., Qwen2.5-VL-7B) | 15K | 96.5 | 43.0 | 27.6 | 47.8 | 44.7 | 40.8 |
| SpatialBlock-3B-reason (Full-SFT Init., GPT-5) | 15K | 75.3 | 41.7 | 28.0 | 44.3 | 48.1 | 40.5 |

phasizes spatial transformations such as 90-degree rotations, and Q3 requires integrating spatial information across multiple perspectives. The single-type variants exhibit different performance patterns across benchmarks, suggesting that each type provides distinct spatial supervision. However, while single-type variants show limited gains, training on all three types achieves the best performance across most benchmarks. These results indicate that the three types provide complementary supervision, and combining them leads to more robust spatial reasoning and generalization.

#### 5.3.2 Ablation on Visual Cues

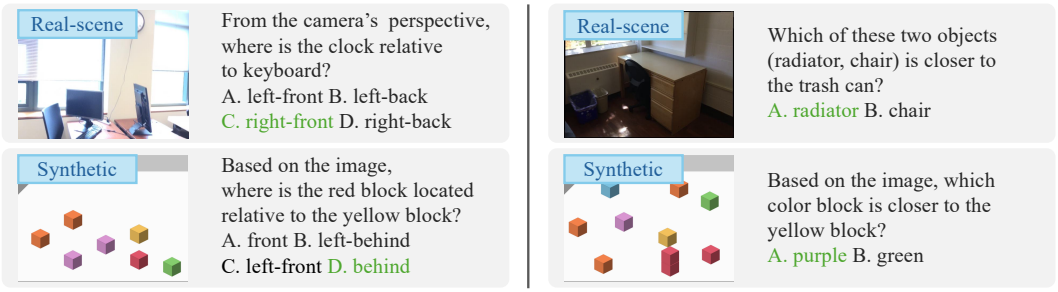
We introduce color manipulation as a visual cue to approximate how humans interpret real-world scenes using visually salient reference objects (Sec. 3.2). To evaluate its effectiveness, we compare models trained with and without visual cues. As shown in Tab. 2(b), removing visual cues leads to performance drops in *direct* and *reason* models of 7.9% and 7.3% in MindCube, and 1.9% and 4.8% in MMSI-Bench, respectively. These results suggest that color cues help models better understand complex scenes and mimic human scene perception, improving spatial reasoning in real-world environments.

#### 5.3.3 Ablation on Task Design

In order to analyze the effect of the task design itself more thoroughly, we train a model on SpatialLadder-26k using the identical training recipe. For a fair comparison, we use 11.5k multiple-choice questions matching our response format and train for 1.3 epochs to match the total number of training samples. As shown in Tab. 2(c), this model underperforms ours across all benchmarks except the in-domain benchmark SPBench, suggesting that our performance gains mainly stem from the proposed block-stacking task design.

#### 5.3.4 Ablation on model initialization for reasoning model

To train a reasoning-capable model via reinforcement learning, we adopt a LoRA-based initialization strategy to enhance baseline spatial intelligence while preserving inherent reasoning ability. Tab. 2(d) compares different initialization strategies and their effects on performance. Without prior fine-tuning, performance drops noticeably across benchmarks, indicating the importance of establishing basic spatial understanding. We further compare our approach with commonly used cold-start phases based on full-parameter tuning (Guo et al., 2025; Li et al., 2026). While cold-start training can improve instruction following and achieve higher accuracy convergence, generating reliable trajectories for our tasks is challenging: both a larger model (Qwen2.5-VL-7B) and a proprietary model (GPT-5) frequently produce inaccurate block descriptions and reasoning, as shown in Fig. 8. Consequently, training with these noisy reasoning traces yields only marginal gains. In contrast, LoRA-based initialization provides substantial improvements, offering a strong starting point for reinforcement learning with both basic spatial intelligence and reliable instruction-following ability.




**(a) Relative Direction** **(b) Relative Distance**

Figure 4: **Synthetic-Real Dataset**. Widely used spatial question types, (a) relative direction and (b) relative distance, are adapted into a block-based environment while preserving their semantic structure.

Table 3: **Comparison of Training on SpatialBlock and Synthetic-Real**. While Synthetic-Real improves in-domain performance, training on SpatialBlock leads to better generalization on out-of-domain benchmark datasets. All experiments are conducted on Qwen2.5-VL-3B (Bai et al., 2025b), and in-domain performance is evaluated on SPBench.

| | **In-domain** | | **Out-of-domain** | | |
|---|---|---|---|---|---|
| Train Dataset | Relative Direction | Relative Distance | MindCube | MMSI-Bench | MMMU |
| Baseline | 31.4 | 74.1 | 39.1 | 26.1 | 45.6 |
| Synthetic-Real | **43.4** | **76.9** | 32.8 | 25.5 | 43.6 |
| SpatialBlock | 31.4 | 67.4 | **49.1** | **28.1** | **46.4** |

### 5.4 Generalizability of Block-Stacking Questions

As shown in Tab. 1, our models trained on SpatialBlock-15k generalize well to real-scene benchmark datasets despite involving question types that differ from those commonly used. This observation raises the question of whether our block-stacking tasks provide a more effective learning signal than conventional spatial tasks. To investigate this, we construct *Synthetic-Real*, a dataset of the same size as SpatialBlock-15k, which adapts two widely used spatial question types – relative direction and distance – into our block-based environment (Fig. 4). Although both datasets can be generated with comparable effort, the results in Tab. 3 reveal that the model trained on Synthetic-Real shows gains only on in-domain tasks but does not transfer these gains to the out-of-domain benchmark datasets. In contrast, our model consistently improves across benchmarks. These results suggest that SpatialBlock tasks promote a more holistic understanding of 3D scenes, leading to more generalizable and transferable spatial reasoning than standard spatial supervision.

## 6 Conclusion

In this paper, we revisited the problem of improving spatial intelligence in large vision-language models from a new perspective: learning foundational spatial skills through structured tasks rather than relying on annotation-heavy real-scene supervision. To this end, we introduced **SpatialBlock-15k**, a novel synthetic dataset of block-stacking problems, including 3D-to-2D projection, viewpoint transformation, and structural combination. The dataset further incorporated controlled color variation as visual cues to encourage anchor-based reasoning, mimicking how humans interpret scenes by using specific objects as reference points. Using this dataset, we developed two specialized models: **SpatialBlock-direct** for immediate answer generation and **SpatialBlock-reason** for detailed logical inference. Experimental results demonstrated that our models consistently outperform existing spatial specialists on various real-world spatial benchmarks. Furthermore, extensive analyses show that the learned capabilities generalize across tasks and domains, indicating that the models capture the underlying spatial structure of scenes. These findings highlighted that robust spatial intelligence can be achieved through training on carefully designed synthetic spatial tasks, offering a scalable and effective path for improving spatial reasoning in LVLMs.

## AI Use Statement

We used generative AI tools solely as a reference for improving the clarity and phrasing of the manuscript. The tools were not used to generate or draft substantive content, nor were they used for research ideation, methodology, experimental design, or interpretation of results. All final wording and content were reviewed and determined by the authors, who take full responsibility for the manuscript.

## References

Anthropic. Claude sonnet 4.5 system card. Technical report, Anthropic, 2025. URL https://assets.anthropic.com/m/12f214efcc2f457a/original/Claude-Sonnet-4-5-System-Card.pdf.

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

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

Yvonne M Caldera, Anne McDonald Culp, Marion O'Brien, Rosemarie T Truglio, Mildred Alvarez, and Aletha C Huston. Children's play preferences, construction play with blocks, and visual-spatial skills: Are they related? *International Journal of Behavioral Development*, 23(4):855–872, 1999.

Beth M Casey, Nicole Andrews, Holly Schindler, Joanne E Kersh, Alexandra Samper, and Juanita Copley. The development of spatial skills through interventions involving block building activities. *Cognition and instruction*, 26(3):269–309, 2008.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aur'elien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany M. Biron, Binh Tang, Bobbie Chern, Char lotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel J. Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab A. AlBadawy, E I Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guanglong Pang, Guillem Cu-curell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert,

Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Ju-Qing Jia, Kalyan Vasuden Alwala, K. Upasani, Kate Plawiak, Keqian Li, Kenneth Heafield, Kevin R. Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuen ley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Ma hesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melissa Hall Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bash-lykov, Nikolay Bogoychev, Niladri S. Chatterji, Olivier Duchenne, Onur cCelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasić, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, R. S. Mughil Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ron nie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sa hana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Chandra Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whit-man, Sten Sootla, Stéphane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whit ney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yiqian Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zhengxu Yan, Zhengxing Chen, Zoe Papakipos, Aaditya K. Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adi Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Fein-stein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin R. Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Po-Yao (Bernie) Huang, Beth Loyd, Beto de Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Shang-Wen Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smoth-ers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco (Paco) Guzmán, Frank J. Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Han Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kaixing(Kai) Wu, U KamHou, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, A Lavender, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthias Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer,

Michal Valko, Michelle Re-strepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Mun ish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navy ata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pe dro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollár, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Ro han Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, S. Yu. Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Zha, Shiva Shankar, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Kumar Gupta, Sung-Bae Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria O Ajayi, Victoria Montanez, Vijai Mohan, Vinay Kumar, Vishal Mangla, Vlad Ionescu, Vlad Andrei Poenaru, Vlad T. Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xia Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024. URL `https://arxiv.org/abs/2407.21783`.

Zhiyuan Feng, Zhaolu Kang, Qijie Wang, Zhiying Du, Jiongrui Yan, Shubin Shi, Chengbo Yuan, Huizhi Liang, Yu Deng, Qixiu Li, et al. Seeing across views: Benchmarking spatial reasoning of vision-language models in robotic scenes. *arXiv preprint arXiv:2510.19400*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Jamie J Jirout and Nora S Newcombe. Building blocks for developing spatial skills: Evidence from a large, representative us sample. *Psychological science*, 26(3):302–310, 2015.

Michael F Land and Mary Hayhoe. In what ways do eye movements contribute to everyday activities? *Vision research*, 41(25-26):3559–3565, 2001.

Susan C Levine, Kristin R Ratliff, Janellen Huttenlocher, and Joanna Cannon. Early puzzle play: a predictor of preschoolers' spatial transformation skill. *Developmental psychology*, 48(2):530, 2012.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. Spatialladder: Progressive training for spatial reasoning in vision-language models. In *International Conference on Learning Representations*, volume 2026, pp. 76566–76592, 2026.

Yuhong Liu, Beichen Zhang, Yuhang Zang, Yuhang Cao, Long Xing, Xiaoyi Dong, Haodong Duan, Dahua Lin, and Jiaqi Wang. Spatial-ssrl: Enhancing spatial understanding via self-supervised reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9570–9581, 2026.

Haoran Lou, Chunxiao Fan, Ziyan Liu, Yuexin Wu, and Xinliang Wang. Llava-sp: Enhancing visual representation with visual spatial tokens for mllms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22014–22024, 2025.

Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso de Melo, Jianwen Xie, and Alan Yuille. Spatialreasoner: Towards explicit and generalizable 3d spatial reasoning. *arXiv preprint arXiv:2504.20024*, 2025a.

Wufei Ma, Luoxin Ye, Celso M de Melo, Alan Yuille, and Jieneng Chen. Spatialllm: A compound 3d-informed design towards spatially-intelligent large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17249–17260, 2025b.

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

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Kexin Tian, Jingrui Mao, Yunlong Zhang, Jiwan Jiang, Yang Zhou, and Zhengzhong Tu. Nuscenes-spatialqa: A spatial understanding and reasoning benchmark for vision-language models in autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4567–4576, 2025.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri Iyer, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *arXiv preprint arXiv:2505.23747*, 2025.

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*, pp. 10632–10643, 2025.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. In *International Conference on Learning Representations*, volume 2026, pp. 157051–157088, 2026.

Junting Lu Yaowei Zheng, Zhangchi Feng Shenzhi Wang, Yuwen Xiong Dongdong Kuang, and Richong Zhang. Easyr1: An efficient, scalable, multi-modality rl training framework. `https://github.com/hiyouga/EasyR1`, 2025.

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.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.