SpatialBench-R: Evaluating Vision-Language Models on Robot-Relevant Spatial Reasoning for Tabletop Manipulation

Mengti Sun[1] Bowen Jiang[1] Minxi Duan[2] Camillo J. Taylor[1]
[1]University of Pennsylvania [2] Northwestern University
{mengti, bwjiang, cjtaylor}@seas.upenn.edu minxiduan2024@u.northwestern.edu

Abstract

VLMs are increasingly deployed as planners in manipulation pipelines, yet their configuration-dependent spatial reasoning (reachability, occlusion, and action effects) remains poorly understood. We present ***SpatialBench-R(obotics)****, a benchmark of 1,632 QA pairs across six robot-relevant task types from 300 synthetic 3D tabletop scenes. We evaluate GPT-5.5, Qwen3-VL-8B, and Gemini-2.5-Flash under image-only, coordinate-augmented, and scene-graph prompting, and introduce an* intervention protocol *that perturbs robot configurations to measure whether models update their spatial judgments. Under image-only prompting, reach-zone estimation and action-effect prediction are near-chance. Structured prompts yield large but task-dependent gains, exposing a gap between static scene description and configuration-dependent inference. Data and code are available at* https://github.com/ChrisSun99/spatialbench-r-release.

1. Introduction

Autonomous robot manipulation requires configuration-dependent spatial judgments: whether the arm can reach an object, whether the target is occluded [2], and how the scene changes after an action. VLMs are an attractive candidate [1, 3], yet existing evaluations focus on general commonsense spatial reasoning [4, 7] rather than robot-centric properties. Crucially, reachability is a function of arm radius and base position, not visual proximity alone. **Critically, robots must not only understand the world but actively interact with it:** a world model is only useful insofar as it supports action selection, precondition checking, and effect prediction under changing physical configurations. To expose whether models track configuration changes rather than relying on appearance, we introduce an *intervention protocol* that perturbs robot configurations and measures *sensitivity*, defined as accuracy on questions whose correct answers change under perturbation.

Table 1. The six SpatialBench-R task types and their grounding.

Task	Question type
T1 Reachability	Can the arm reach target X?
T2 Occlusion	Is target X occluded from the robot?
T3 Spatial relation	What is the spatial relation between objects?
T4 Reach-zone	Which objects lie within the workspace zone?
T5 Counting	How many objects are in a given region?
T6 Action effect	What is the scene state after picking X?

2. SpatialBench-R

Task types. SpatialBench-R defines six QA categories covering the spatial preconditions a manipulation planner must evaluate (Table 1). Each of the 300 procedurally-generated scenes contributes 5–6 QA pairs; scenes vary in object count (3–8), obstacle placement, and robot configuration.

Intervention protocol. For each base scene we generate perturbed variants via four types: **(P1)** *Base relocation*: robot base repositioned; **(P2)** *Radius change*: arm radius scaled up or down; **(P3)** *Obstacle insertion*: obstacle placed between arm and target; **(P4)** *Target relocation*: target moved inside or outside reach. We report **sensitivity**: accuracy on flip-questions whose correct answer changes under perturbation.

Prompting strategies. **(a) Image-only**: rendered 2D top-down scene image with question text. **(b) Coordinates**: image + question + explicit object coordinates, arm base, and radius appended to the prompt. **(c) Scene graph** [6, 10]: image + question + a ground-truth scene graph encoding all pairwise spatial relations and robot configuration, as shown in Figure 1, an oracle upper bound on what explicit relational structure can unlock.

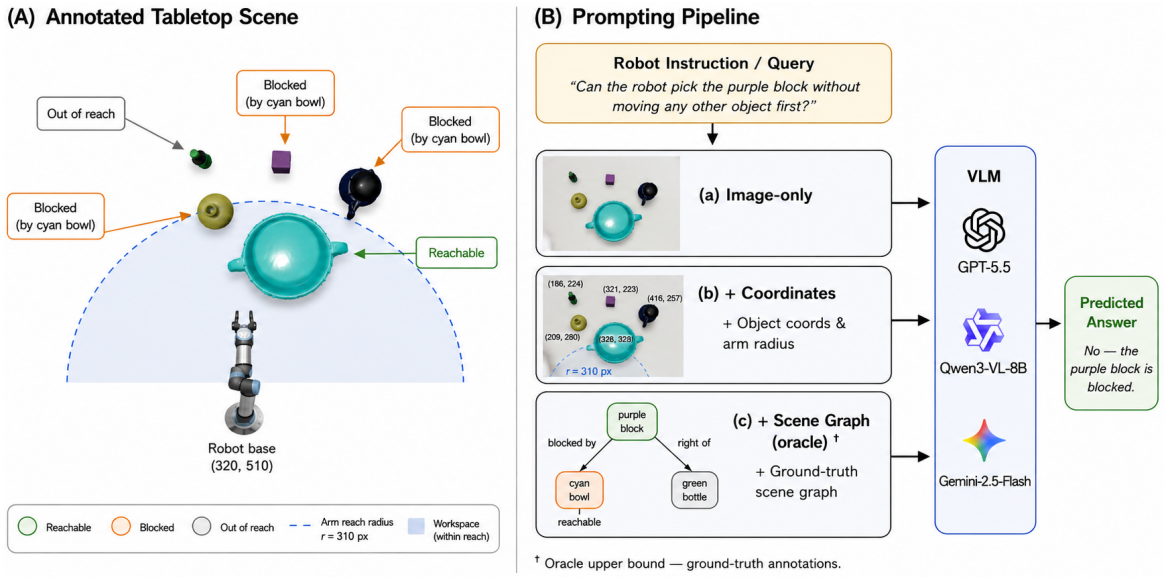

Figure 1. **SpatialBench-R overview. (A)** A procedurally generated top-down tabletop scene with per-object reachability labels (green: reachable, orange: blocked, grey: out of reach). **(B)** The three prompting strategies: image-only, +coordinates, and +scene graph (oracle upper bound).

3. Experiments

Task accuracy. We evaluate GPT-5.5 [8], Gemini-2.5-Flash [5], and Qwen3-VL-8B [9] across all three strategies; Table 2 reports per-task accuracy. Two tasks prove consistently difficult under image-only prompting across all models: T4 (reach-zone, 14.7–32.0%) and T6 (action effect, 19.7–32.6%), while T2 (occlusion) and T5 (counting) are already strong (88.0–97.3% and 70.3–90.7%), confirming that the difficulty is geometric rather than perceptual. On reachability (T1), GPT-5.5 substantially outperforms the others under image-only (79.7% vs. 49.3% and 55.7%), suggesting that model scale aids configuration-aware visual reasoning. Across all models, coordinate augmentation recovers reach-zone accuracy to 92.7–100%, while scene graphs push action-effect to 96.2–97.0%.

Notably, structured prompts do not help uniformly. Reachability degrades for all models once coordinates are added (dropping to ∼50%), and spatial-relation accuracy shows no consistent improvement, indicating that geometric structure is helpful only when it directly matches the task computation.

Intervention sensitivity. Image-only sensitivity varies substantially across models: GPT-5.5 [8] achieves 87.6%, Gemini-2.5-Flash [5] reaches 78.0%, and Qwen3-VL-8B [9] only 61.8%. The failure modes differ. Qwen3-VL-8B collapses on P2 (radius change, 40.3%) and Gemini-2.5-Flash is weakest on P1 (base relocation, 53.3%), suggesting that each model's visual priors create distinct configuration blind spots. The intervention-conditioned scene graph closes the gap across all three models: GPT-5.5 reaches 100%, Gemini-2.5-Flash 98.7%, and Qwen3-VL-8B 94.8%, confirming that structured relational representations are a universal remedy regardless of model size or architecture.

Table 2. Task accuracy (%) per strategy. Each cell: GPT-5.5 | Qwen3-VL-8B | Gemini-2.5-Flash. **Bold**: the best result per task.

Task	Image-only	+Coords	+Scene Graph
T1 Reachability	**79.7** |49.3 |55.7	46.3 |50.0 |50.0	50.0 |49.7 |50.0
T2 Occlusion	88.0 |95.7 |97.3	85.0 |86.3 |90.0	89.0 |89.3 |**100.0**
T3 Spatial rel.	35.7 |46.7 |36.7	43.3 |50.0 |49.0	22.3 |**52.0** |46.7
T4 Reach-zone	14.7 |32.0 |24.7	**100.0** |93.3 |92.7	**100.0** |77.0 |71.7
T5 Counting	70.3 |89.7 |**90.7**	57.0 |88.7 |78.0	83.3 |89.0 |90.3
T6 Action eff.	19.7 |32.6 |32.6	94.7 |87.9 |**98.5**	**97.0** |96.2 |96.2

4. Conclusion

SpatialBench-R reveals that VLMs struggle specifically on configuration-dependent tasks (reach-zone and action-effect), while handling perceptual tasks such as occlusion well. Structured prompts help unevenly, with large gains where explicit geometric computation is required but limited or negative effects elsewhere. Our intervention protocol provides a sharper diagnostic than aggregate accuracy. Image-only sensitivity varies widely across models (61.8–87.6%), while scene-graph prompting restores near-perfect sensitivity across all models, pointing toward structured representations as a practical path for robust embodied reasoning.

References

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, et al. Do as I can, not as I say: Grounding language in robotic affordances. In *CoRL*, 2022. 1

[2] Bibit Bianchini, Minghan Zhu, Mengti Sun, Bowen Jiang, Camillo J. Taylor, and Michael Posa. Vysics: Object reconstruction under occlusion by fusing vision and contact-rich physics. In *Robotics: Science and Systems (RSS)*, 2025. 1

[3] Anthony Brohan, Noah Brown, Justice Carbajal, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control, 2023. 1

[4] Bohao Chen, Zixian Xu, Issam Laradji, Michael Rabbat, Marcus Rohrbach, et al. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *CVPR*, 2024. 1

[5] Google DeepMind. Gemini 2.5 flash. https://deepmind.google/technologies/gemini/flash/, 2025. 2

[6] Bowen Jiang, Zhijun Zhuang, Shreyas S Shivakumar, and Camillo J Taylor. Enhancing scene graph generation with hierarchical relationships and commonsense knowledge. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8883–8894. IEEE, 2025. 1

[7] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. 1

[8] OpenAI. GPT-5.5 system card. https://openai.com/, 2025. 2

[9] Qwen Team. Qwen3-vl technical report. https://github.com/QwenLM/Qwen3-VL, 2025. 2

[10] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5419, 2017. 1