BEHAVIOR SPATIAL BENCHMARK: EVALUATING SPATIAL UNDERSTANDING FOR ROBOTICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Robot tasks require grounding spatial language in physical scenes, but task success alone cannot establish whether spatial instructions are followed. We present BEHAVIOR-Spatial, a benchmark built on BEHAVIOR-1K that measures spatial question answering, point-based target grounding, and robot action to examine how VLM spatial understanding relates to VLA behavior. It comprises 3,256 spatial questions across 49 activities, 1,240 grounding queries across 14 manipulation tasks, and action evaluations covering those 14 tasks and 10 navigation tasks. Our manipulation protocol evaluates held-out configurations and instruction following when the requested target changes within a fixed training layout. The baseline achieves 41.4% on held-out configurations but only 16.1% under changed instructions. VLM adaptation and robot-data pretraining provide limited improvements, whereas oracle visual target cues raise these rates to 60.3% and 38.0%, respectively. Perception evaluations reveal policy-dependent correlations between VLM question-answering and grounding accuracy and VLA manipulation performance. Together, these findings motivate evaluating spatial instruction following alongside task completion and examining when spatial perception performance translates into robot behavior.

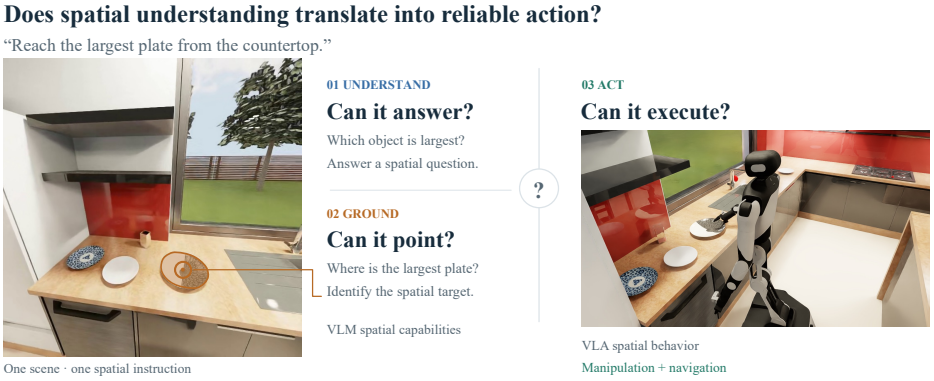

Figure 1: **BEHAVIOR-Spatial** links VQA (Understand), target grounding (Ground), and execution (Act), testing whether VLM spatial capabilities predict VLA task performance.

1 INTRODUCTION

Instructions such as "reach the second tool from the left" require a robot to resolve a spatial relation, identify the intended target, and execute an action. Navigation similarly requires interpreting destinations relative to current and previously observed surroundings. Vision-language-action models (VLAs) are initialized from vision-language models (VLMs) and assume that the backbone's spatial understanding supports robot behavior, yet this connection is rarely studied directly. Task success alone provides ambiguous evidence: a policy may exploit familiar layouts without following spatial language, or select the correct target but fail during execution. Relating VLM spatial capabilities to VLA behavior requires jointly evaluating spatial perception, target grounding, and action.

We introduce BEHAVIOR-Spatial, a benchmark built on BEHAVIOR-1K (Li et al., 2024a) that connects these evaluations across household manipulation and navigation. It comprises three com-

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.

ponents: (1) BEHAVIOR-Spatial-VQA, a spatial VQA suite containing 3,256 questions across 49 task activities, including 1,572 general spatial questions and 1,684 task-specific questions, with both single- and multi-frame evaluations; (2) BEHAVIOR-Spatial-IR, a point-based intermediate representation (IR) generation suite containing 1,240 target-grounding queries across 14 manipulation tasks; and (3) BEHAVIOR-Spatial-Action, a robot action suite covering the same 14 manipulation tasks and 10 navigation tasks. Shared scenes and relations across suites enable studying how VLM question-answering and grounding accuracy relate to VLA performance.

Our evaluation distinguishes generalization to new configurations from responsiveness to spatial instructions. *Held-out* manipulation and navigation configurations test generalization, while a manipulation *prompt probe* changes the requested target within a training layout to test whether behavior follows the instruction. We measure VQA answer accuracy and whether generated points fall within the instructed target region. For robot action, we report *soft success* as a behavioral proxy for target-selection intention, alongside *wrong-target commitments* and *strict success*.

Our experiments reveal persistent spatial instruction-following failures, especially when the target changes within a familiar layout. The tested LoRA adaptation and DROID (Khazatsky et al., 2024) initialization strategies provide limited improvements on this probe, whereas supplying IR visual target cues substantially improves target-consistent behavior. These cues isolate the benefit of providing the grounding decision, rather than demonstrating learned grounding from language. For two of our best manipulation variants, the dot-IR and attention-map policies, grounding accuracy correlates more strongly with soft success than VQA accuracy does. However, the perception–action relationship varies across policy variants, and direction-question accuracy shows little correlation with navigation success across the ten evaluated tasks.

Our contributions are threefold: (1) **a spatial benchmark** connecting question answering, point-based target grounding, and household manipulation and navigation; (2) **a diagnostic evaluation protocol** that tests changed instructions in training layouts and separates target-consistent behavior from task completion; and (3) **an empirical study** of policy adaptation and visual grounding interfaces relating VLM spatial understanding to VLA behavior.

2 Related Work

Spatial understanding in VLMs and its benchmarks. A growing body of work studies spatial reasoning in vision-language models, including 3D-grounded spatial QA (Chen et al., 2024; Cheng et al., 2024), relative position and multi-view correspondence (What'sUp (Kamath et al., 2023), BLINK (Fu et al., 2024)), egocentric video (VSI-Bench (Yang et al., 2025)), spatial mental models (MindCube (Wang et al., 2026a)), and atomic embodied capabilities such as pointing (BEAR (Qi et al., 2026)). Robo2VLM (Chen et al., 2025) is closest to our VQA setting, generating questions from real robot trajectories. These benchmarks largely assess spatial question answering or pointing in isolation. BEHAVIOR-Spatial instead aligns VQA, pointing, and execution on the same scenes, objects, and relations, testing whether spatial understanding transfers to grounded action.

VLMs as the foundation of VLA models. Vision-language-action models augment pretrained VLM backbones with action prediction or control modules (Zitkovich et al., 2023; Kim et al., 2024; Team et al., 2024; Black et al., 2026; Intelligence et al., 2025; NVIDIA et al., 2025; Team et al., 2025; Team, 2025), implicitly relying on the backbone's semantic and spatial knowledge to transfer into the policy. Recent work has begun to examine this assumption: VLM4VLA (Zhang et al., 2026a) studies backbone choice; Hancock et al. (2025) identify catastrophic forgetting of vision-language abilities during action finetuning; and Qwen-VLA (Wang et al., 2026b) uses joint vision-language and action training to mitigate this issue. OpenHelix (Cui et al., 2025) and Mind-to-Hand (Tang et al., 2025) further incorporate explicit embodied reasoning. Our benchmark complements these efforts by separately measuring VQA, referent grounding, and execution.

Benchmarks and evaluation for robot policies. CALVIN (Mees et al., 2022), LIBERO (Liu et al., 2023), VLABench (Zhang et al., 2024), and SimplerEnv (Li et al., 2024b) evaluate language-conditioned policies primarily through end-task success, while BEHAVIOR-1K (Li et al., 2024a) provides the long-horizon household activities and simulator we build on. Factored evaluations such as The Colosseum (Pumacay et al., 2024) and DOMINO (Fang et al., 2026) expose specific

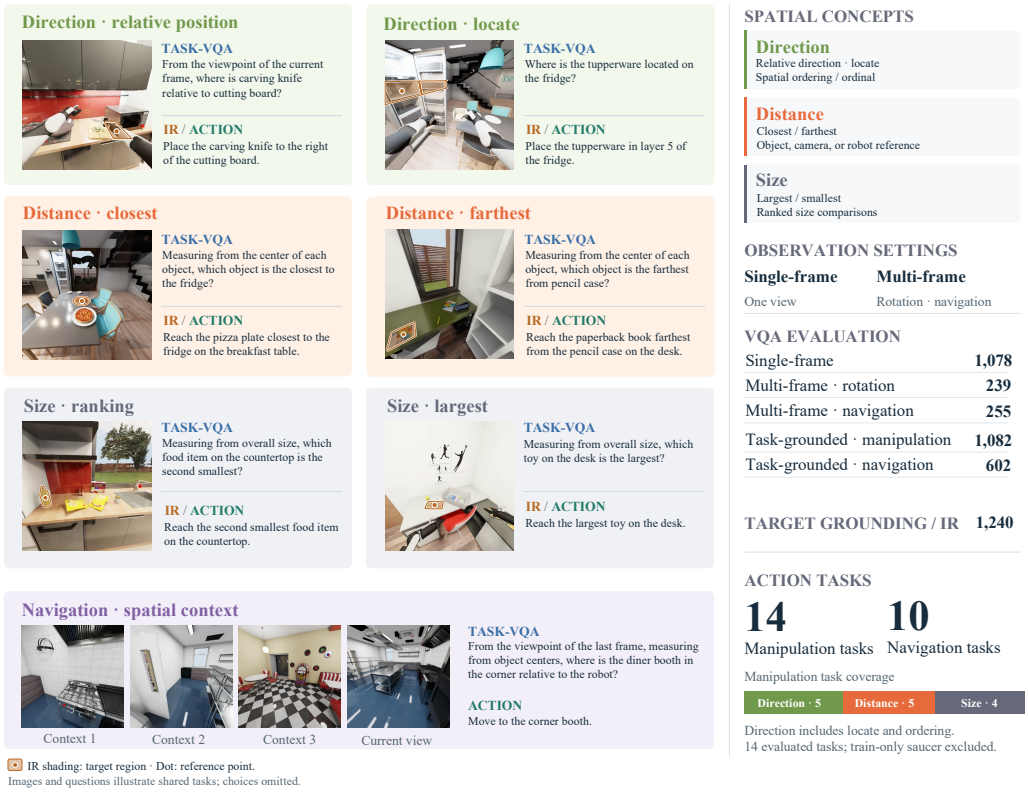

Figure 2: **BEHAVIOR-Spatial Dataset Showcase.** Left: Paired task VQA, grounding, and action examples cover direction, distance, and size; navigation shows three context views and the current view. Right: Spatial taxonomy, VQA counts, grounding/IR representations, and task coverage.

sources of policy brittleness. BEHAVIOR-Spatial similarly adopts a diagnostic evaluation, but targets spatial language grounding: prompt-probe pairs hold the training scene fixed while varying the instruction, and outcome buckets and strict/soft metrics separate grounding from execution.

Intermediate representations and explicit grounding for policies. Points, marks, traces, and sketches have been used as interfaces between perception and control (Lee et al., 2025; Li et al., 2026; Zhang et al., 2026b). We use such an intermediate representation for a different purpose: as a controlled intervention on an otherwise fixed policy. Providing an object-center mark tests whether supplying the correct referent resolves the grounding failure and isolates the remaining execution gap. We then ask whether the grounding signal can instead be supplied by a VLM at rollout time.

3 The Behavior Spatial Benchmark

3.1 Design Overview

BEHAVIOR-Spatial comprises three components, overviewed in Figure 2.

BEHAVIOR-Spatial-VQA evaluates spatial understanding through general and task-grounded questions. The general subset covers spatial relations within individual views and reasoning across multiple views, while the task-grounded subset queries objects and relations relevant to the robot tasks. BEHAVIOR-Spatial-IR evaluates whether a model can ground a spatial instruction by predicting a point within the specified target object or placement region. Finally, BEHAVIOR-Spatial-Action evaluates spatially grounded action through manipulation and navigation tasks in which spatial instructions determine the intended target or destination.

Task-grounded VQA and BEHAVIOR-Spatial-IR (together, the task-grounded perception suites) share task contexts, objects, and spatial relations with the action suite, enabling task-level comparison between VLM perception and VLA behavior.

3.2 Spatial Taxonomy

BEHAVIOR-Spatial covers direction, distance, and size. Direction includes relative-position relations in object-, camera-, and robot-relative frames, as well as region identification (*locate*) and spatial ordering (*ordinal*). Distance and size require comparisons of physical separation and object size, respectively. Spatial VQA further distinguishes single-frame understanding from multi-frame reasoning across robot rotation or navigation. Appendix A.1 details the concepts and settings.

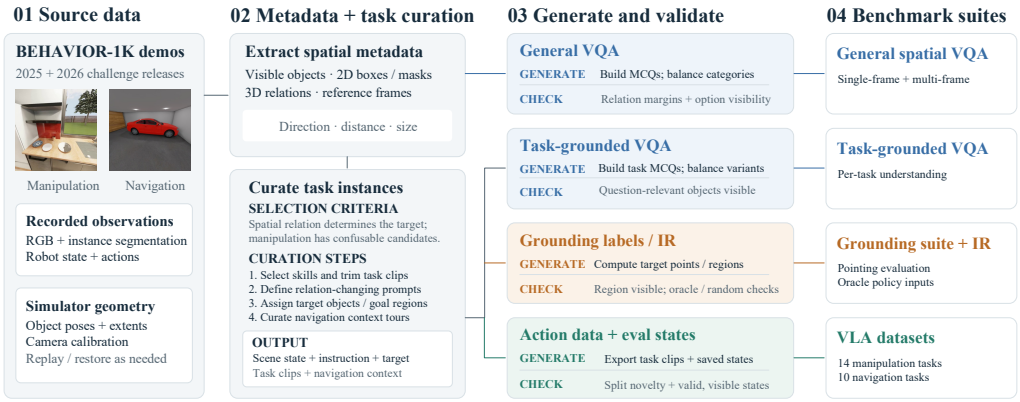

Figure 3: **Data Construction Pipeline.** Demonstration observations and simulator geometry support general VQA and curated spatial tasks. Task instances produce task-grounded VQA, grounding labels, and VLA datasets. Selection criteria guide curation; each suite validates its samples.

3.3 Dataset Construction

We construct BEHAVIOR-Spatial from BEHAVIOR-1K demonstrations, using recorded observations and actions together with object geometry and camera metadata recovered through OmniGibson replay as needed. Figure 3 summarizes the pipeline. We curate **14 manipulation tasks** and **10 navigation tasks** across five navigation scenes. Manipulation tasks require identifying a target among spatially distinct candidates; navigation tasks require grounding an instructed destination. Instruction variants specify different target objects or goal regions, and curated scene-context traversals support navigation. Appendix B describes task curation and lists the tasks and variants.

A metadata-to-template pipeline generates spatial questions and their answers, with visibility checks, relation-margin filters, and sampling controls to reduce ambiguity and imbalance. General VQA spans 49 activities and contains 1,078 single-frame and 494 multi-frame questions. Task-grounded VQA derives from the curated robot task instances and contains 1,082 manipulation and 602 navigation questions. For point grounding, we compute target-object references and placement positions or regions from segmentation and simulator geometry. VQA and grounding use distinct visibility criteria. See Appendix A for curation details.

4 VLA Model

4.1 Model Architecture

BEHAVIOR-Spatial-VLA builds on the $\pi_{0.5}$ architecture (Intelligence et al., 2025), replacing its VLM backbone with Qwen3.5-4B (Team, 2026). The VLM encodes visual observations and the language instruction, with discretized robot proprioception included in the text prefix. A flow-matching action expert attends to the resulting prefix features and predicts 32-step chunks of 23-dimensional actions controlling the torso, dual arms, grippers, and mobile base. Unless otherwise stated, the VLM is frozen and the action expert is trained from scratch.

For manipulation, policies are trained on BEHAVIOR-Spatial-Action's manipulation demonstrations and receive head- and two wrist-camera views. IR-conditioned variants additionally receive oracle grounding images through the same visual encoding interface. For navigation, policies re-

ceive the current head-camera view and a variable-length sequence of curated scene-context frames, together with the instruction and robot state.

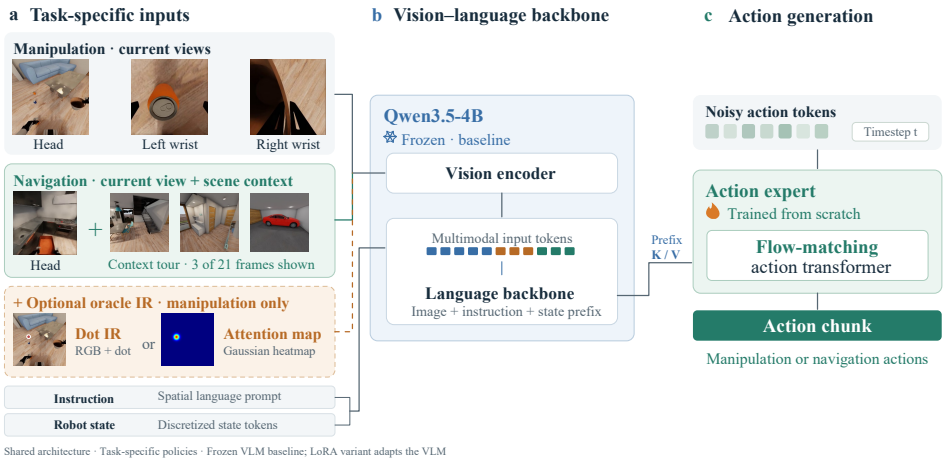

Figure 4: **BEHAVIOR-Spatial-VLA Architecture.** Task specific views share a Qwen3.5-4B backbone and flow matching action expert. Only manipulation uses optional oracle grounding.

4.2 Policy Variants

We compare six manipulation policies, all conditioned on camera observations, the instruction, and robot state: (1) **Baseline** (frozen VLM); (2) **+LoRA-VLM**, which adapts the vision and language components through LoRA; (3) **+DROID pretrain init**, which initializes from DROID pretraining before task finetuning; (4) **+dot IR**, which adds a target-marked head-camera RGB image; (5) **+dot IR + first-frame anchor**, which also supplies a fixed, target-marked initial head view; and (6) **+attn-map IR**, which uses a standalone Gaussian heatmap in place of the dot-marked image.

These variants test backbone adaptation, robot-data transfer, and alternative grounding interfaces. The dot and attention-map comparison changes both spatial spread and image background. For navigation, we train a single frozen-VLM policy jointly on the 10 tasks. Appendices C.1 and C.2 detail adaptation, initialization, and training. Appendix C.3 specifies IR construction.

4.3 Policy Evaluation Protocol

Held-out & Prompt-probe evaluation. The held-out set (n=470) uses scene configurations from unseen demonstrations, paired with unseen instruction variants that reverse the spatial relation used during training. This setting tests generalization under changes to both scene configuration and instruction. The prompt-probe set (n=700) retains scene configurations used during training but changes the instruction to specify a different valid target. This setting tests whether policies follow the instruction when it conflicts with the demonstrated target.

Manipulation evaluation. All manipulation policies are evaluated on the same instances. Episodes use per-task step budgets calibrated to the corresponding demonstration horizons and are scored using the metrics defined in §5.1. For IR-conditioned policies, we provide ground-truth grounding references computed from simulator state. This oracle setting evaluates action performance given correct grounding, independently of errors in VLM grounding predictions.

Navigation evaluation. We evaluate the jointly trained 10-task navigation policy on spawn-validated held-out instances. Appendix C.4 describes instance construction and visualizes initial poses and scene layouts. Training settings and evaluated checkpoints are provided in Appendix C.2.

5 Experiments and Results

5.1 Metrics

VQA and IR Accuracy. VQA accuracy measures the fraction of questions answered correctly under the four-option evaluation interface; unparseable responses count as incorrect. For IR evaluation, we ask model to predict a target point, measuring strict IR accuracy uses task-specific target-region criteria and tolerant IR accuracy that additionally allow small localization errors for reach tasks. We report task-grounded manipulation VQA and IR accuracy as averages of per-task accuracies. Detailed scoring rules, tolerances, and aggregation conventions are provided in Appendix D.

Robot Tasks Success. For manipulation, we primarily use *soft success* as a proxy for target-selection intention. It credits completed correct interactions as well as non-committed episodes that exhibit correct intention for target approaching under task-specific rules and timeout budgets. We also report *strict success* for task completion rate, and the *wrong-target rate* measuring wrong commitments to non-instructed candidates or locations. All three rates report percentage over all evaluated episodes. Task-family rules and exceptions are detailed in Table S3. Navigation uses the recorded binary goal result: the robot must satisfy the instructed region-entry or object-proximity criterion within the episode timeout budget. We report per-task and pooled success, without the strict/soft distinction.

Table 1: **VLM accuracy on BEHAVIOR-Spatial-VQA.** Single-frame Direction includes its locate subtype; All is pooled over all questions.

Model	Single-frame				Multi-frame	
	All	Dir.	Dist.	Size	All	Dir.
Proprietary						
GPT-6 Astra	**82.1**	**85.8**	**77.3**	**77.7**	**74.4**	**74.6**
Gemini 3	77.8	80.5	75.2	72.7	71.1	69.6
GPT-5	64.7	61.7	68.1	69.1	63.2	59.6
Gemini 2.5	59.9	55.0	65.2	68.3	60.1	57.4
Open-weight						
Molmo2-8B	51.8	45.0	57.5	66.9	36.8	30.7
Qwen3.5-9B	50.6	42.0	58.1	69.1	54.5	51.4
Qwen3.5-4B	46.8	44.8	48.1	51.8	46.6	44.1
Molmo2-4B	45.5	34.2	58.7	62.6	32.4	23.7
PaliGemma	36.2	32.5	40.7	41.0	37.9	34.4
PaliGemma 2	34.8	29.8	39.2	45.3	27.3	23.4
Gemma 3-4B	34.4	35.0	33.6	33.8	29.2	21.3
Qwen2.5-VL	33.4	31.2	37.8	32.4	28.9	22.9
DeepSeek-VL2-Tiny	31.4	30.5	34.8	26.6	25.0	17.6
VLM backbones extracted from released VLA models						
OpenHelix	9.6	5.8	14.2	15.1	—	—
$\pi_{0.5}$	0.9	0.8	0.3	2.9	—	—
Chance	25.0	25.0	25.0	25.0	25.0	25.0

5.2 VLM Evaluation on BEHAVIOR-Spatial-VQA and BEHAVIOR-Spatial-IR

How well do VLMs answer spatial questions? Table 1 evaluates 13 VLMs, comprising four proprietary – GPT-6 Astra (OpenAI, 2026), Gemini 3 (Google DeepMind, 2025), GPT-5 (Singh et al., 2026), and Gemini 2.5 (Comanici et al., 2025)– and nine open-weight models spanning Qwen (Qwen Team, 2026), Molmo (Deitke et al., 2024), Gemma (Gemma Team et al., 2025), and DeepSeek (Wu et al., 2024) families. VQA evaluation prompts, including spatial conventions and multi-frame context instructions, are provided in Appendix E.1. The highest single-frame accuracy is 82.1%, compared with 51.8% for the strongest open-weight model. Multi-frame performance varies substantially across models, with several open-weight models at or below the random-guess baseline on direction questions. Spatial VQA thus remains challenging, especially across multiple views. Extracted VLA backbones also score poorly (Appendix F.1).

Does general VQA performance transfer to task-grounded questions? We evaluate the best open-source models from the general VQA comparison on task-grounded questions from the manipulation suite (Table 2); only open-source models are evaluated, as only these can serve as our VLA backbones. Qwen3.5-9B achieves the highest accuracy at 57.2%, followed by Molmo2-4B at 53.0% and Molmo2-8B at 50.0%. Even Qwen3.5-9B attains 56.3% on direction and 61.4% on distance, but only 47.3% on locate sub-types of direction questions. Strong general VQA thus does not guarantee task-grounded perception, motivating task-grounded evaluation of candidate backbones.

How accurately can VLMs ground the instructed target? The IR evaluation asks models to predict a target point directly from the instruction and head-camera view (Table 2), using the prompts

provided in Appendix E.2. Qwen3.5-9B achieves the highest strict accuracy at 51.4%, followed by Molmo2-8B at 46.7% and Qwen3.5-4B at 43.3%. Under tolerant scoring, these accuracies increase to 59.7%, 58.4%, and 56.2%, respectively. Qwen3.5-4B and Molmo2-8B reverse order between VQA and pointing, so answer selection and localization probe different aspects of perception. We relate these task scores to policy behavior in §5.3.2; qualitative examples appear in Appendix F.2.

Table 2: **Task-grounded perception performance for manipulation.** VQA and point-grounding accuracy (%) for the same VLMs. VQA reports overall and category-level accuracy; IR reports strict and tolerant accuracy. Higher is better in all columns; bold indicates the highest accuracy.

Model	Task-grounded VQA				IR	
	All	Dir.	Dist.	Size	Strict	Tolerant
Qwen3.5-9B	**57.2**	**56.3**	**61.4**	**52.5**	**51.4**	**59.7**
Molmo2-4B	53.0	54.1	56.4	46.8	27.1	40.5
Molmo2-8B	50.0	49.9	53.4	45.3	46.7	58.4
Qwen3.5-4B	49.4	50.9	48.6	48.6	43.3	56.2
Gemma 3-4B	33.5	23.0	42.4	36.0	36.8	47.9

Table 3: **Task-level perception–action correlation.** Spearman correlations across 14 manipulation tasks between VQA or tolerant IR accuracy and policy success. Ties receive average ranks.

Policy	VQA (ρ)	IR (ρ)
Baseline	0.713	0.658
LoRA VLM	0.616	0.581
IR	0.560	0.701
IR + first-frame IR	0.612	0.618
Attention map	0.608	0.753

5.3 VLA EVLAUATION ON ROBOT TASKS

We evaluate the policy variants of §4.2 on the manipulation and navigation tasks of BEHAVIOR-Spatial-Action (Tables 4 and S9). Checkpoints are selected around similar training steps for fair comparisons.

The policy predicts 32-step action chunks using 10 flow-matching denoising steps per inference. Execution uses asynchronous receding-horizon control with a 10-step replanning threshold measured from the committed plan's input-observation timestamp. Actions are temporally ensembled over up to three committed plans, with gripper commands taken from the newest plan. Details are in Appendix F.3.

5.3.1 VLA PERFORMANCE ON MANIPULATION TASKS

Results are task-dependent, with particularly limited instruction following when the requested target changes in training layouts. VLM adaptation and robot-data pretraining offer limited gains, while oracle target grounding yields larger improvements.

How well do policies act toward spatially specified targets? Averaged across manipulation tasks, on held-out instances the baseline achieves 41.4% soft and 27.8% strict success, with 38.2% wrong targets.

The baseline's per-task results (Appendix F.4) reveal distinct bottlenecks. For `place_carving_knife`, 70.8% soft success versus 29.2% strict success suggests execution difficulty despite frequent target-directed progress. Bowl placement achieves only 15.0% on both measures, with a 77.5% wrong-target rate, indicating that target selection itself is difficult. Both target selection and execution thus limit completion.

Table 4: **Success rates of all policy variants on the two evalutaion sets.** S = strict success; +Soft = soft success; W = fraction of episodes committing to a non-instructed object.

Policy	Held-out		Prompt-probe		
	+Soft↑	S↑	+Soft↑	S↑	W↓
$\pi_{0.5}$-Qwen (frozen VLM)	41.4	27.8	16.1	10.0	70.0
+ LoRA-VLM	43.5	31.0	17.9	9.6	68.1
+ DROID pretrain init	40.1	24.7	16.6	8.9	66.0
IR-conditioned					
+ dot IR	**60.3**	**44.8**	**38.0**	**24.9**	43.9
+ dot IR + first-frame anchor	59.1	41.8	36.3	23.7	42.4
+ attn-map IR	59.3	39.0	35.0	17.0	46.0
Data scaling (§F.8)					
+ dot IR @ 50% data	53.9	34.8	30.6	19.4	45.7
+ dot IR @ 30% data	56.1	34.5	32.6	19.9	46.6

Do policies follow changed instructions in training layouts? The baseline often fails to redirect its behavior toward a newly requested target, even in training layouts. On prompt-probe episodes, it achieves only 16.1% soft and 10.0% strict success, while 70.0% of episodes commit to a non-instructed target. The policy relies on scene–action associations rather than the instruction. In the failed probe example in Figure S12, the instruction requests the largest food item, but the policy

reaches the chip bag associated with the original training instruction rather than the indicated oatmeal box. Held-out target-directed behavior thus does not imply reliable instruction following.

Does VLM adaptation or robot-data pretraining improve instruction following? LoRA adaptation increases held-out soft success from 41.4% to 43.5%, but yields only a small change on prompt-probe episodes (16.1% to 17.9%), with 68.1% wrong-target commitments remaining. DROID initialization achieves 40.1% held-out and 16.6% prompt-probe soft success. Neither recipe improves strict probe success over the baseline (9.6% and 8.9%, respectively, versus 10.0%).

Does explicit target grounding improve policy performance? Oracle IR produces substantially larger gains: dot-IR conditioning raises soft success from 41.4% to 60.3% on held-out set and from 16.1% to 38.0% on the prompt probe. Across the three IR variants, probe wrong-target commitments decrease from 70.0% to 42.4–46.0%. Strict success also improves, reaching 44.8% held-out and 24.9% on the probe for the strongest variant with dot IR. Explicit grounding thus redirects policy behavior and improves completion.

5.3.2 Perception–Action correlation for manipulation

Do perception scores track spatially directed behavior? We compare Qwen3.5-4B's task-level VQA and tolerant IR accuracies with each policy's soft held-out success across 14 manipulation tasks, using Spearman's rank correlation with average ranks for ties (Table 3). For two of the strongest-performing variants, dot IR and attention-map IR, grounding accuracy correlates more strongly with success than VQA accuracy ($\rho = 0.701$ versus 0.560, and 0.753 versus 0.608, respectively). This suggests that the metric that tracks the policy is the one the VLM has to generate, not the one it has to answer. Correlation plots are provided in Appendix F.5.

5.3.3 Ablation Studies for manipulation tasks

We examine how target cues and demonstration design affect manipulation performance.

Does the IR rendering affect target-consistent behavior? We compare two ways of presenting the same oracle target center: a dot on the head-camera RGB image and a standalone Gaussian attention map. Both yield similar soft success on held-out and probe sets, but dot IR achieves higher strict success, possibly because it preserves the surrounding RGB context.

With VLM-predicted instead of oracle grounding, the dot-IR policy retains part of this benefit: Qwen3.5-4B predictions yield 48.8% held-out and 27.1% prompt-probe soft success, compared with 41.6% and 16.1% for the baseline on the matched episodes. Performance remains below oracle grounding (60.3% and 38.0%). Appendix F.6 provides the rollout protocol and full results, including held-out comparisons across three grounding VLMs.

Does a first-frame reference improve performance? The target can leave the head-camera view as the robot moves toward it, motivating a fixed reference that retains the initial scene and target mark. We add the target-marked first frame as a fifth image input to the dot IR policy, held constant throughout the rollout. Both soft and strict success decrease slightly, so a persistent initial view adds no benefit over the current-view cue.

Does changing the demonstration distribution affect target selection? To examine the influence of demonstration geometry, we study the `reach_paperback_book` task in which the referred book is associated with a fixed side in the training data. We prefix each demonstration with motion from the two candidate books' perpendicular bisector to the original start, changing the approach geometry while retaining the original interaction. The main 14-task evaluation uses the augmented paperback demonstrations. On the shared prompt-probe set, this augmentation reduces wrong-target commitments from 79.7% to 43.5% and raises soft probe success from 11.6% to 36.2%, while strict success remains 10.1%. This intervention changes the failure distribution without improving completion. Details are in Appendix F.7.

How much does performance depend on demonstration volume? We retrain our best-performing variant, the dot-IR policy, on nested 30% and 50% subsets of the demonstrations, pre-

serving the task mixture and training for the same number of updates. Increasing the data from 30% to 100% raises held-out soft/strict success from 56.1%/34.5% to 60.3%/44.8%, and prompt-probe soft/strict success from 32.6%/19.9% to 38.0%/24.9%. The full-data policy performs best overall, although gains are not monotonic: the 50% subset yields similar or lower success than the 30% subset. Even with 30% of the demonstrations, dot IR substantially outperforms the full-data frozen-VLM baseline on both evaluation sets. See Appendix F.8 for details.

5.4 VLA PERFORMANCE ON NAVIGATION TASKS

How does navigation success vary across tasks? Success ranges from 22.5% for the garage to 92.5% for the bookcase (Table S9). The bookcase and bathroom tasks achieve high success on short-range navigations, while the garage task remains challenging. The diner scene and two-floor house also exhibit substantial variation across destinations, with success from 25–35% to 75–85%. Scene identity alone therefore does not account for the task ordering. Reaching a nearby object or region and navigating across a larger environment impose different demands on context use and motion. Route length, obstacles, and doorway crossings are plausible contributors to these differences. See Appendix F.11.

Does navigation VQA accuracy track policy success? We evaluate Qwen3.5-4B on multi-frame direction-question for navigation, resulting accuracy ranges from 16.7% to 38.6%. Across the ten tasks, the weak correlation with navigation success indicates that VQA accuracy alone does not explain variation in task success. Navigation-VQA examples are provided in Appendix S14. The complete results for all ten navigation tasks are reported in Table S9 (Appendix F.10). The table pairs VQA accuracy and policy success by task and scene to show variation across environments.

6 LIMITATIONS AND CONCLUSION

6.1 LIMITATIONS

Our evaluation is conducted entirely in simulation; transfer to physical robots remains to be established. In the tupperware task, correlations between the acting hand and the target layer can involve spatial instruction following with hand-specific manipulation behavior. Success rates also depend on the episode timeout budget, so cross-run comparisons should account for differences in the available execution time. Finally, the IR-conditioned policies require simulator-derived IR target centers during training; applying this training approach to datasets without such annotations would require an alternative source of grounding supervision.

6.2 CONCLUSION

We introduced BEHAVIOR-Spatial to jointly evaluate spatial question answering, target grounding, and robot action in household tasks. The evaluated policies often fail to act on the instructed target, particularly when prompts change in familiar layouts. Explicit IR target cues improve target-directed behavior, with gains persisting under VLM-predicted grounding, although reliable completion remains challenging. Perception scores offer policy-dependent diagnostics but do not consistently explain action success. By distinguishing target selection, partial progress, and task completion, BEHAVIOR-Spatial provides a framework to measure when spatial understanding supports successful action and when it fails.

AI USE STATEMENT

Generative AI tools were used to assist with drafting and editing manuscript text, writing figure-generation scripts, and aggregating evaluation results from logged metrics. The authors reviewed the AI-assisted text and code and verified the reported results against the experimental logs. The authors take responsibility for the final manuscript, including all claims and artifacts.

Ethics Statement

Our benchmark uses simulated environments and programmatically generated labels. The grounding failures identified in this work highlight potential risks of deploying robots that do not reliably follow instructions. Simulation performance should not be interpreted as evidence of safe real-world deployment.

Reproducibility Statement

To support reproducibility, we describe benchmark construction and label generation in Section 3.3, evaluation splits and protocols in Section 4.3, and experimental settings in Section 5. We will release the benchmark and code upon publication.

Author Contributions

References

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control, 2026. URL https://arxiv.org/abs/2410.24164.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities, 2024. URL https://arxiv.org/abs/2401.12168.

Kaiyuan Chen, Shuangyu Xie, Zehan Ma, Pannag R Sanketi, and Ken Goldberg. Robo2vlm: Visual question answering from large-scale in-the-wild robot manipulation datasets, 2025. URL https://arxiv.org/abs/2505.15517.

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, 2024. URL https://arxiv.org/abs/2406.01584.

Gheorghe Comanici et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Can Cui, Pengxiang Ding, Wenxuan Song, Shuanghao Bai, Xinyang Tong, Zirui Ge, Runze Suo, Wanqi Zhou, Yang Liu, Bofang Jia, Han Zhao, Siteng Huang, and Donglin Wang. Openhelix: A short survey, empirical analysis, and open-source dual-system vla model for robotic manipulation, 2025. URL https://arxiv.org/abs/2505.03912.

Matt Deitke, Christopher Clark, Jordi Salvador, Dustin Schwenk, Aniruddha Kembhavi, Ali Farhadi, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models, 2024. URL https://arxiv.org/abs/2409.17146.

Heng Fang, Shangru Li, Shuhan Wang, Xuanyang Xi, Dingkang Liang, and Xiang Bai. Towards generalizable robotic manipulation in dynamic environments, 2026. URL https://arxiv.org/abs/2603.15620.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive, 2024. URL https://arxiv.org/abs/2404.12390.

Gemma Team et al. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Google DeepMind. Gemini 3 pro model card. https://deepmind.google/models/gemini/pro/, 2025. Model gemini-3-pro-preview, accessed April 2026.

Asher J. Hancock, Xindi Wu, Lihan Zha, Olga Russakovsky, and Anirudha Majumdar. Actions as language: Fine-tuning vlms into vlas without catastrophic forgetting, 2025. URL https://arxiv.org/abs/2509.22195.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization, 2025. URL https://arxiv.org/abs/2504.16054.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning, 2023. URL https://arxiv.org/abs/2310.19785.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Sakthivel, Archit Wang, Megha Nair, Daya Jain, et al. Droid: A large-scale in-the-wild robot manipulation dataset. In *arXiv preprint arXiv:2403.12945*, 2024.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model, 2024. URL https://arxiv.org/abs/2406.09246.

Jason Lee, Jiafei Duan, Haoquan Fang, Yuquan Deng, Shuo Liu, Boyang Li, Bohan Fang, Jieyu Zhang, Yi Ru Wang, Sangho Lee, Winson Han, Wilbert Pumacay, Angelica Wu, Rose Hendrix, Karen Farley, Eli VanderBilt, Ali Farhadi, Dieter Fox, and Ranjay Krishna. Molmoact: Action reasoning models that can reason in space, 2025. URL https://arxiv.org/abs/2508.07917.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Wensi Ai, Benjamin Martinez, Hang Yin, Michael Lingelbach, Minjune Hwang, Ayano Hiranaka, Sujay Garlanka, Arman Aydin, Sharon Lee, Jiankai Sun, Mona Anvari, Manasi Sharma, Dhruva Bansal, Samuel Hunter, Kyu-Young Kim, Alan Lou, Caleb R Matthews, Ivan Villa-Renteria, Jerry Huayang Tang, Claire Tang, Fei Xia, Yunzhu Li, Silvio Savarese, Hyowon Gweon, C. Karen Liu, Jiajun Wu, and Li Fei-Fei. Behavior-1k: A human-centered, embodied ai benchmark with 1,000 everyday activities and realistic simulation. *arXiv preprint arXiv:2403.09227*, 2024a.

Hao Li, Ziqin Wang, Zi han Ding, Shuai Yang, Yilun Chen, Yang Tian, Xiaolin Hu, Tai Wang, Dahua Lin, Feng Zhao, Si Liu, and Jiangmiao Pang. Robointer: A holistic intermediate representation suite towards robotic manipulation, 2026. URL https://arxiv.org/abs/2602.09973.

Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*, 2024b.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning, 2023. URL https://arxiv.org/abs/2306.03310.

Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks, 2022. URL https://arxiv.org/abs/2112.03227.

NVIDIA, Johan Bjorck, Nikita Cherniadev Fernando Castañeda, Xingye Da, Runyu Ding, Linxi "Jim" Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu,

Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. GR00T N1: An open foundation model for generalist humanoid robots. In *ArXiv Preprint*, March 2025.

OpenAI. GPT-6 Astra System Card, September 2026. URL https://deploymentsafety.openai.com/gpt-6-astra. Accessed: 2026-09-26.

Wilbert Pumacay, Ishika Singh, Jiafei Duan, Ranjay Krishna, Jesse Thomason, and Dieter Fox. The colosseum: A benchmark for evaluating generalization for robotic manipulation, 2024. URL https://arxiv.org/abs/2402.08191.

Yu Qi, Haibo Zhao, Ziyu Guo, Siyuan Ma, Ziyan Chen, Yaokun Han, Renrui Zhang, Zitiantao Lin, Yizhe Zhu, Shiji Xin, Yijian Huang, Boce Hu, Kai Cheng, Peiheng Wang, Jiazheng Liu, Jiayi Zhang, Yizhe Zhu, Wenqing Wang, Yiran Qin, Haojie Huang, and Lawson L. S. Wong. Dissecting embodied abilities in multimodal language models through skill-level evaluation and diagnosis, 2026. URL https://arxiv.org/abs/2510.08759.

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

Aaditya Singh et al. OpenAI GPT-5 system card, 2026. URL https://arxiv.org/abs/2601.03267.

Peijun Tang, Shangjin Xie, Binyan Sun, Baifu Huang, Kuncheng Luo, Haotian Yang, Weiqi Jin, and Jianan Wang. Mind to hand: Purposeful robotic control via embodied reasoning, 2025. URL https://arxiv.org/abs/2512.08580.

Gemini Robotics Team et al. Gemini robotics: Bringing ai into the physical world, 2025. URL https://arxiv.org/abs/2503.20020.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy, 2024. URL https://arxiv.org/abs/2405.12213.

Qwen Team. Qwen3.5-4b: A unified vision-language foundation model. https://huggingface.co/Qwen/Qwen3.5-4B, 2026.

Xiaomi Embodied Intelligence Team. Mimo-embodied: X-embodied foundation model technical report, 2025. URL https://arxiv.org/abs/2511.16518.

Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Jiajun Wu, Li Fei-Fei, and Manling Li. Mindcube: Spatial mental modeling from limited views, 2026a. URL https://arxiv.org/abs/2506.21458.

Qiuyue Wang, Mingsheng Li, Jian Guan, Jinhui Ye, Sicheng Xie, Yitao Liu, Junhao Chen, Zhixuan Liang, Jie Zhang, Xintong Hu, Xuhong Huang, Pei Lin, Junyang Lin, Dayiheng Liu, Shuai Bai, Jingren Zhou, Jiazhao Zhang, Haoqi Yuan, Gengze Zhou, Hang Yin, Ye Wang, Yiyang Huang, Zixing Lei, Wujian Peng, Delin Chen, Yingming Zheng, Jingyang Fan, Xianwei Zhuang, Xin Zhou, Haoyang Li, Anzhe Chen, Tong Zhang, Xuejing Liu, Yuchong Sun, Ruizhe Chen, Zhaohai Li, Chenxu Lü, Zhibo Yang, Tao Yu, and Xionghui Chen. Qwen-vla: Unifying vision-language-action modeling across tasks, environments, and robot embodiments, 2026b. URL https://arxiv.org/abs/2605.30280.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024. URL https://arxiv.org/abs/2412.10302.

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, 2025. URL https://arxiv.org/abs/2412.14171.

Jianke Zhang, Xiaoyu Chen, Qiuyue Wang, Mingsheng Li, Yanjiang Guo, Yucheng Hu, Jiajun Zhang, Shuai Bai, Junyang Lin, and Jianyu Chen. Vlm4vla: Revisiting vision-language-models in vision-language-action models, 2026a. URL https://arxiv.org/abs/2601.03309.

Shiduo Zhang, Zhe Xu, Peiju Liu, Xiaopeng Yu, Yuan Li, Qinghui Gao, Zhaoye Fei, Zhangyue Yin, Zuxuan Wu, Yu-Gang Jiang, and Xipeng Qiu. Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks, 2024. URL https://arxiv.org/abs/2412.18194.

Zhengshen Zhang, Hao Li, Yalun Dai, Zhengbang Zhu, Lei Zhou, Chenchen Liu, Dong Wang, Francis E. H. Tay, Sijin Chen, Ziwei Liu, Yuxiao Liu, Xinghang Li, and Pan Zhou. From spatial to actions: Grounding vision-language-action model in spatial foundation priors, 2026b. URL https://arxiv.org/abs/2510.17439.

Anthony Zitkovich, Soravit Yu, Stefan Wang, Alex Anh, Pannag Sundararajan, Karol Petrenko, Quan Yu, Mohit Bansal, Karol Hausman, Yevgen Chebotar, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *7th Annual Conference on Robot Learning*, 2023. URL https://arxiv.org/abs/2307.15818.