ARE FRONTIER VLM AGENTS READY TO BE ROBOT GENERALISTS? AN EMPIRICAL STUDY WITH THE EMBODIED AGENT ARENA

Anonymous authors
Paper under double-blind review

ABSTRACT

Frontier vision-language models (VLMs) combine scene estimation, interaction grounding, and executable actions. Understanding how these abilities support complete robotic tasks is central to evaluating their readiness as robot generalists. We introduce Embodied Agent Arena to examine where local competence supports, or falls short of, complete task success across Geometry, Spatial Reasoning, Affordance, Task Planning, and Manipulation. The arena contains 1,000 cases drawn from 32 established sources and GeoProbe, our new benchmark for geometric estimation on Blender renders and real-scene images. A minimal harness preserves source observations and operations while separating metric precision, functional grounding, and native goal completion. We evaluate seven VLMs, analyze Astra's task-specific advantages, and compare richer-observation execution protocols and multi-round review. Across the arena, Astra's advantage is strongest in precise estimation and usable-contact localization; completing coordinated, goal-directed actions remains the key gap to robot generalism.



Figure 1: Overview of Embodied Agent Arena. Each row illustrates six tasks in one capability domain, spanning camera and scene estimation, spatial relations, functional grounding, goal-directed planning, and robot manipulation. Fixed-input probes test estimates and judgments; interactive episodes test goal completion. Spatial-tracing overlays show reference paths and a 3D target region.

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.

1 INTRODUCTION

A robot generalist must turn scene understanding into actions that satisfy a complete goal. Estimating where an object lies, selecting a usable contact, and executing a grasp are useful abilities, yet none establishes that the object reaches its required destination. Vision-language models (VLMs) increasingly address spatial relations (Ma et al., 2024; Yang et al., 2024) and action-relevant grounding (Wang et al., 2025a; Song et al., 2024). Learned visuomotor policies connect observations to robot actions (Kim et al., 2024; Black et al., 2024), while language-model controllers select skills or generate executable programs (Ahn et al., 2022; Liang et al., 2022). Frontier VLM agents now combine observation, programming, and feedback-driven decisions (Fu et al., 2026; Guo et al., 2026), making the relationship between local competence and complete execution a practical evaluation question.

Capability breadth and reliable completion require different evidence. An average score can hide whether errors concern metric scale, an object's reference frame, or the final condition of a placement. Conversely, a terminal failure can conceal useful progress. Evaluating these outcomes separately reveals what a model can do and which additional requirements remain unsatisfied. Generalist readiness therefore calls for both cross-capability comparisons and task-level diagnoses.

We ask: *Are frontier VLM agents ready to be robot generalists?* Embodied Agent Arena evaluates the same models across Geometry, Spatial Reasoning, Affordance, Task Planning, and Manipulation (Figure 1). Its 1,000 cases combine 32 established sources with GeoProbe, our geometric-estimation benchmark. GeoProbe varies camera motion, object motion, and scale references to test which quantities remain accurately estimated as visual conditions change. A minimal harness preserves source observations and operations. Continuous errors, task-specific evidence, and terminal success reveal complementary aspects of agent capability.

Our comparisons locate the remaining gap in task requirements: Astra produces the closest paths yet often misses their endpoints, and its local manipulation strengths coexist with low household-goal completion. Under matched budgets and execution limits, richer visual observations reverse Astra's CALVIN lead while producing gains on other tasks. A separate multi-round review study measures which answers improve after reinspection or specialist measurements. Our contributions are:

- **An evaluation perspective centered on task requirements.** We distinguish local estimates and interaction evidence from complete goal satisfaction across five robotic capability domains.
- **A unified arena with a new geometric benchmark, GeoProbe.** We combine 32 established sources with 168 GeoProbe cases, using controlled Blender scenes and real images to diagnose camera, motion, depth, and scale estimation.
- **A capability baseline for frontier VLM agents.** We characterize seven agents across five robotic application domains within specified inference-time and execution budgets, with task-level comparisons and a deep dive into Astra's strengths and failure modes.

2 RELATED WORK

Embodied evaluation has developed along two complementary directions. Control suites measure execution and task generalization through robot interfaces (James et al., 2019; Mees et al., 2021), whereas visual diagnostics isolate spatial inference (Ma et al., 2024; Yang et al., 2024) and interaction grounding (Song et al., 2024; Qian & Fouhey, 2023). The former establishes whether a goal is achieved; the latter resolves particular perceptual or reasoning requirements. Evaluating the same agents on both reveals how their strengths and rankings vary with these task requirements.

Broader evaluations connect perception with decision-making (Li et al., 2024b; Qi et al., 2025), including navigation and object interaction through executable interfaces (Yang et al., 2025a; Cheng et al., 2025). Embodied Arena combines fixed probes and interaction (Ni et al., 2025). Interfaces also change the model's work: code supports reusable controllers (Fu et al., 2026), while persistent state introduces memory demands (Zhang et al., 2026). Our study explains task-level differences through separately scored geometry, grounding, and terminal outcomes (Table 1). Protocol comparisons make the interface and resource conditions part of the interpretation of model rankings.

Table 1: **Evaluation coverage and diagnostic analysis.** Reported evaluations in the cited papers. Region probes score image grounding; robot manipulation evaluates arm/gripper control; memory spans sessions. Calib./pose tests camera calibration or relative pose; task forms denote type/condition breakdowns.

	Evaluation coverage					Diagnostic analysis	
Benchmark	Fixed probes	Region probes	Online nav.	Robot manip.	State memory	Calib./ pose	Task forms
BEAR (2025)	✓	✓	✗	✓*	✗	✗	✓
EmbodiedBench (2025a)	✗	✗	✓	✓	✗	✗	✓
EmbodiedEval (2025)	✗	✗	✓	✗	✗	✗	✓
Embodied Arena (2025)	✓	✓	✓	✗	✗	✗	✓
Code-based control (2026)	✗	✗	✓	✓	✗	✗	✓
VLABench (2024)	✗	✗	✗	✓	✗	✗	✓
WorldLines (2026)	✓	✗	✗	✗	✓	✗	✓
Embodied Agent Arena	✓	✓	✓	✓	✗	✓	✓

✓ included; ✗ not evaluated in the cited study. *BEAR-Agent's online validation.

3 The Embodied Agent Arena

3.1 Benchmark Construction and Capability Domains

Data construction. Embodied Agent Arena contains 1,000 cases from 32 established sources and one new benchmark, GeoProbe, organized into five capability domains (Figures 1–2). Each case specifies observations, a query or goal, operations, and scoring. We introduce GeoProbe to diagnose the metric ambiguities behind geometric estimates. Its 168 cases vary camera motion, object motion, and scale references in controlled Blender scenes, complemented by real images. For interactive cases, the recorded terminal state connects the agent's action sequence to the task's goal conditions. Appendix A gives source allocation and adaptations.

Geometry. Given images, the agent estimates intrinsics, pose, depth, dimensions, or displacement. InFlux and Map-free test calibration and relative pose (Liang et al., 2025; Arnold et al., 2022); GeoProbe varies camera motion and scale references. TraceSpatial-Bench adds paths on 56 scenes (Zhou et al., 2026), testing reference-trajectory error alongside start, arrival, and collision constraints. These targets separate numerical precision from geometric feasibility.

Spatial Reasoning. The agent infers relations, viewpoints, configurations, and temporal order from images or video (Ma et al., 2024; Wang et al., 2025b; Yang et al., 2024). Examples include comparing object distances, imagining another observer's view, and inferring camera motion. Each question specifies the reference frame and answer format: a category, quantity, or point set.

Affordance. Given a scene and intended action, the agent predicts a contact point or bounding box. UMD tests tool functions such as cutting and scooping (Myers et al., 2015); instruction- and action-conditioned 3DOI queries test interaction contacts (Qian & Fouhey, 2023). Point validity checks whether the contact lies in the functional region; overlap measures coverage of its extent.

Task Planning. The agent organizes source-provided operations into a sequence that achieves a household, navigation, or scientific goal (Shridhar et al., 2019; Wang et al., 2022), using visual, textual, or symbolic feedback. Compound goals require several terminal conditions, such as heating an object and placing it at the destination. For terminal state s_T and required predicates $g_1, \ldots, g_K$, complete success is

$$\text{Success}(s_T, g) = \mathbf{1}\left[\bigwedge_{k=1}^{K} g_k(s_T)\right]. \tag{1}$$

Manipulation. The agent executes robot motions or skills for placement, insertion, articulation, and coordinated handling (James et al., 2019; Chen et al., 2025a). Language-conditioned variants change the target relation or object (Zhang et al., 2024); sequential tasks require preserving progress across ordered goals. Native success checks task completion, while condition-level scores record achieved relations and control events.

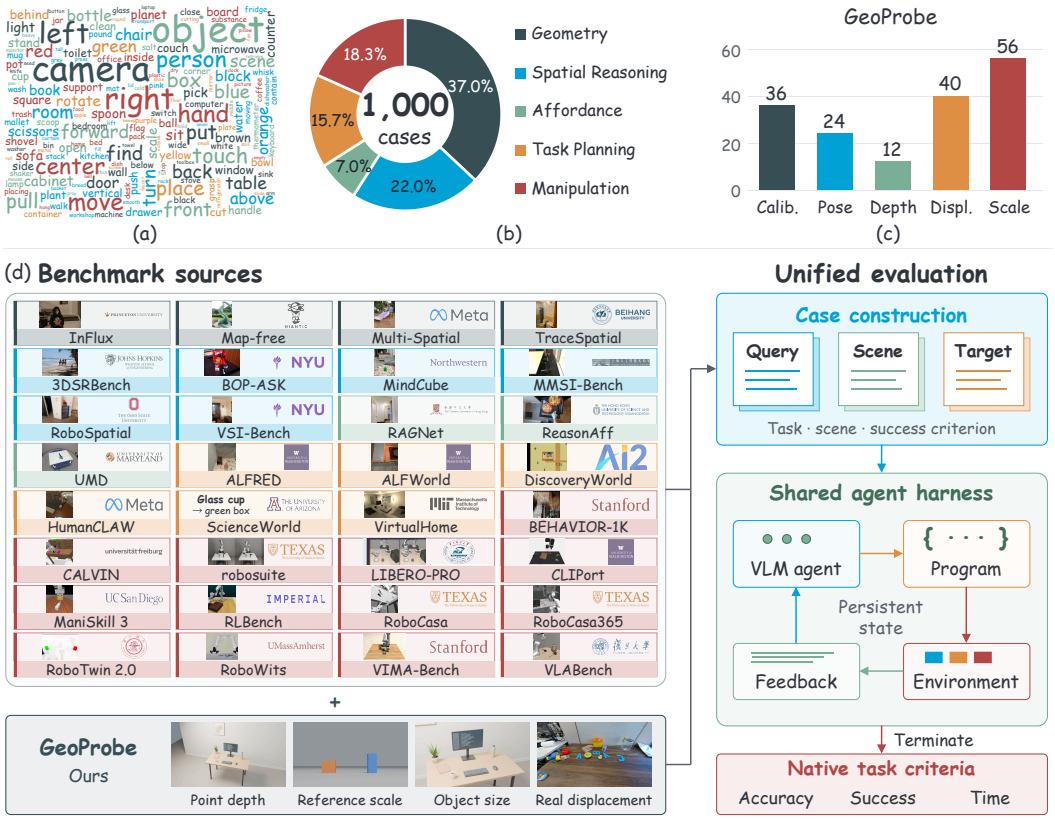

Figure 2: Construction of the 1,000-case arena. (a) Representative task vocabulary across sources. (b) Five-domain composition, including 112 spatial-tracing cases in Geometry. (c) GeoProbe estimation targets. (d) Source queries, scenes, and targets connect model interaction to termination and scoring. Colors identify domains; logos denote source institutions.

3.2 Unified Agent Harness

Model-directed interaction. The harness supports persistent model-directed interaction (Figure 2d). Geometry, Spatial Reasoning, Affordance, and Planning use a Python runtime; Manipulation follows its native agent loop. Models select source-provided operations and use the resulting observations and feedback. Variables persist across Python rounds, and images produced during a round enter the next model request. Appendix B.1 specifies these interfaces; Section 4.4 examines multi-round review with specialist visual tools (Zhang et al., 2025; Huang et al., 2026).

Submission and termination. Fixed-input submissions commit predictions; interactive runs end on agent termination, an environment terminal condition, or budget exhaustion. Scoring independently checks outputs or states. Appendix B.2 details budgets.

3.3 Evaluation Protocol and Metrics

① *Task success.* Spatial Reasoning and Planning report Pass@1. Manipulation reports native success, consecutive goals completed, and task-specific condition or control-evidence scores. Categorical spatial accuracy is reported separately. ② *Geometric precision.* Geometry and numerical spatial queries use MAE, RMSE, AbsRel/MAPE, displacement error, and pose errors according to the target. Tracing pairs trajectory RMSE with offline 3D constraint satisfaction. ③ *Native affordance grounding.* Predicted boxes use GT-mask IoU, with contact validity scored separately. Box–box IoU and strict box success at IoU$\geq$ 0.5 measure localization using the same rectangular representation. ④ *Variation and resource use.* Repeated runs report the mean and standard deviation (SD) across runs; SD measures run-to-run variation. Tokens measure inference use; action calls measure executed control. Section 4.1 specifies repetitions; Appendix B.3 defines aggregation.

Table 2: Performance across five capability domains. Repeated rates show mean ± SD; continuous errors use valid observations. Appendices B–C give full metrics.

	Geometry		Spatial Reasoning		Affordance		Task Planning	Manipulation
Agent	Rot.↓ (°)	Trans.↓ (cm)	Pass@1↑ (%)	AbsRel↓	Mask IoU↑	Point↑ (%)	Pass@1↑ (%)	Task SR↑ (%)
Astra	**17.1**	**106.6**	**69.4**±1.0	0.213	**0.286**	**60.3**±2.6	**80.3**	**41.8**
Sol	27.0	151.9	60.7±2.5	0.273	0.264	52.0±3.6	44.6	19.8
Fable	25.2	121.7	62.0±4.6	0.186	0.284	57.1±2.5	63.7	24.7
Qwen-Max	20.5	138.1	68.9±1.1	0.205	0.272	52.9±4.2	35.7	9.3
Gemini	24.2	140.9	65.4±1.3	**0.161**	0.250	44.6±13.0	48.4	18.7
Qwen-397B	38.4	249.8	50.6±1.8	0.251	0.116	32.6±4.7	17.8	9.9
Qwen-27B	38.4	213.7	48.0±0.9	0.288	0.013	6.3±1.6	19.1	11.0

4 Experiments

4.1 Experimental Setup

① *Models and configuration.* We compare five closed-source VLMs (Astra, Sol, Fable, Qwen-Max, Gemini) and two open-source VLMs (Qwen-397B, Qwen-27B) under the interfaces in Section 3. Budgets are source-specific and matched between Baseline and RGB-expanded; Appendix B.2 records versions and limits. ② *Evaluation and aggregation.* Fixed-input tasks are assigned five runs per case, or three for Fable; tracing and interactive tasks use one adopted run. Manipulation covers 182 binary tasks under Baseline, with an RGB-expanded comparison in Section 4.4. Binary rates retain assigned cases; continuous errors use valid estimates. Appendix B.4 specifies coverage and repetition accounting.

4.2 Main Results

Astra's advantage spans precision, grounding, and execution. In Table 2, Astra has the lowest pose errors (17.1°, 106.6 cm), the highest contact validity (60.3%), and the highest Planning success (80.3%). Its baseline Manipulation success is 41.8%, versus Fable's 24.7%. Figure 3 compares the resulting capability profiles using the fixed transformations in Appendix C.3.

The leading model depends on what success requires. Astra's Spatial Pass@1 narrowly exceeds Qwen-Max's (69.4% versus 68.9%), but Gemini has lower numerical AbsRel (0.161 versus 0.213). Astra's mask IoU is also close to Fable's (0.286 versus 0.284). The task mixture matters: excluding ScienceWorld and VirtualHome lowers Astra's Planning score to 66.7%, still above Fable's 33.3% (Appendix C.2). The Planning lead survives this change in task composition; in Spatial Reasoning, task success and numerical precision favor different models.

4.3 A Deep Dive of Astra

We examine source-wide results, then task contrasts in Figures 4–8; Appendix D retains every task group. Equal source weighting preserves Astra's lead in Spatial, contact localization, and Planning (Appendix C.2).

Geometry. ① *Astra leads controlled metric estimation across targets.* It has the lowest errors on all five Blender estimation targets (Figure 4a), including reference height at 0.31% versus Sol's 0.90%. Depth ordering reaches 96.3%, while oracle scale alignment lowers depth AbsRel from 0.104 to 0.038: much of its remaining depth error concerns scale rather than order. InFlux focal error reverses the ranking (53.0% versus Sol's 47.8%), locating the strongest advantage in controlled geometric cues. ② *Astra separates object displacement from moderate viewpoint change.* Across fixed, small-shift, and large-shift Blender conditions, Astra's MAE is 0.24, 0.57, and 1.13 cm; Sol is the closest peer throughout (Figure 4b). The fixed-camera lead survives excluding stationary objects. On four real tabletop photo pairs, Astra achieves 0.05 versus Sol's 0.42 cm MAE. The lower real-scene error occurs on different scenes and object motions. ③ *Coupled motion challenges Astra's otherwise leading pose estimates.* Its aggregate pose errors are lowest (Table 2), yet no model meets both rotation and translation tolerances under mixed motion. ④ *Astra traces the closest paths, but endpoints decide success.* Astra has the lowest 3D trajectory RMSE at 0.382 m versus Gemini's

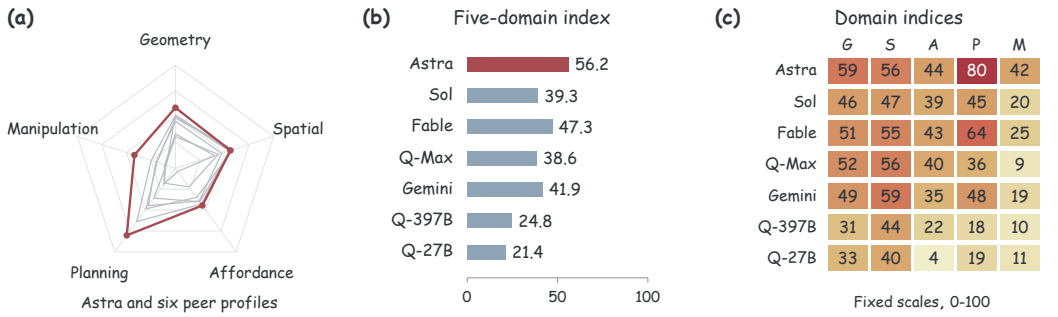

Figure 3: Capability breadth across seven agents. (a) Astra (red) and peers (gray). (b) Mean of five domain indices. (c) Indices by domain, derived from Table 2 using the fixed scales in Appendix C.3. Manipulation uses baseline results; individual profiles appear in Figure A1.

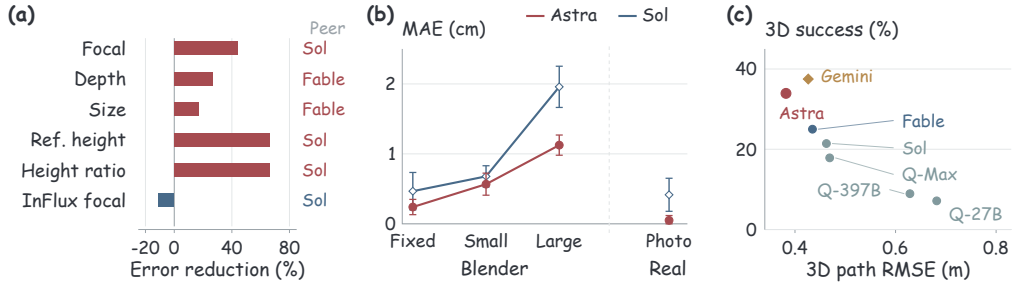

Figure 4: Geometric precision and task completion. (a) Error reduction relative to each target's lowest-error peer; the first five targets are GeoProbe. (b) Displacement under fixed, small-shift, and large-shift Blender cameras, plus separate real photos; points show mean ± run SD. (c) Single-run 3D tracing error versus joint success. Appendix D.1 gives absolute metrics and scene definitions.

0.426, whereas joint success is 33.9% versus 37.5% (Figure 4c). Initiation or arrival fails in 36 of Astra's 37 unsuccessful traces, identifying endpoint placement as the main obstacle to successful tracing.

Spatial Reasoning. ① *Astra's clearest advantage is relating multiple views.* Across all MMSI tasks it reaches 77.0% versus Fable's 66.7%; across MindCube it ties Qwen-Max at 90.8%. Within these sources, object relations reach 96.7% and camera–region relations 100% (Figure 5a,c). Equal weighting of all six sources also ranks Astra first, at 71.9% versus Qwen-Max's 69.1% (Appendix C.2). ② *Astra's strength in relations outpaces its inference of new frames.* MindCube camera-shift accuracy falls to 78.5% versus Qwen-Max's 89.2%; 3DSRBench object-facing judgments reach 56.7% versus Fable's 72.2% (Figure 5a). The shared-run gaps in Figure 5c preserve both deficits. Relating visible landmarks and recovering motion or an object's own axes make different demands. ③ *Depth cues are more reliable than choosing among objects.* Astra achieves 96–100% on BOP-ASK camera-depth ordering, but 50% on VSI nearest-object selection versus Qwen-Max's 76.7% (Figure 5b). RoboSpatial configuration is also 22 points behind its best peer (Figure 5c). Thus its narrow overall Spatial lead combines strong multi-view relations with unresolved object-centered selection and configuration.

Affordance. ① *Astra's contact-localization lead extends across sources.* It leads UMD and ReasonAff at 96.7% and 31.0%, versus Fable's 94.4% and 23.3%, and ties Fable at 35.0% on RAGNet. Figure 6a shows the UMD ranking across all six actions. Across these sources, absolute contact accuracy varies more than Astra's relative ranking. ② *Native boxes locate usable parts without reproducing their shape.* Cutting reaches 100% contact validity and 96% strict box success, despite box–mask IoU of 0.361; box–box IoU is 0.694 (Figure 6b). The rectangular output explains part of this metric gap. Containment still has only 44% strict box success versus Qwen-Max's 96%, showing that contact localization and region extent are distinct even under the same output format. ③ *Action-conditioned contacts expose where Astra's lead changes.* ReasonAff pulling and two-handed movement reach 60% and 48%, above the best peers' 50% and 26.7%; the RAGNet

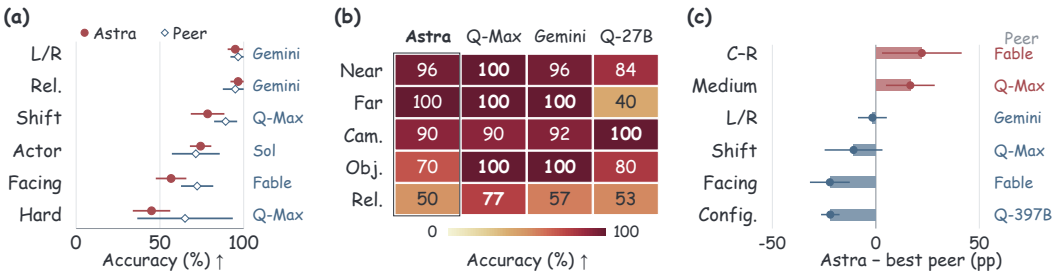

Figure 5: Astra's spatial strengths and reference-frame limits. (a) Relation and viewpoint accuracy compared with each task's best peer. (b) Camera-relative depth (Near/Far/Cam.) and inter-object distance (Obj./Rel.). (c) Astra's accuracy advantage over the best peer on one task per source; C–R denotes camera–region relations. Panels a–b average available runs; c compares shared run indices. Whiskers show run-to-run SD. Appendix D.2 maps labels to all six sources and reports every task.

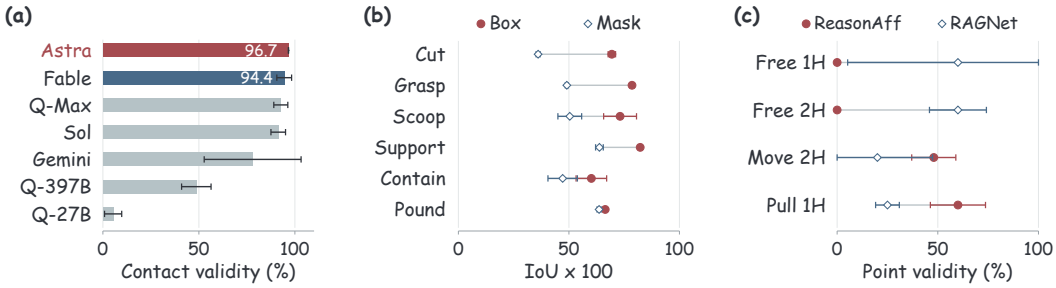

Figure 6: Astra's native affordance localization. (a) Contact validity over all six UMD actions. (b) Astra's boxes evaluated against reference boxes and masks. (c) Contact validity for four corresponding interactions in ReasonAff/RAGNet; 1H/2H means one/two hands. Scores show mean ± run SD; IoUs are multiplied by 100. Appendix D.3 gives all actions and models.

counterparts reach 25% and 20%, below 31.7% and 50% (Figure 6c). The same panel shows zero success on both ReasonAff free-movement categories despite nonzero RAGNet scores. Contact choice therefore depends on the action–scene pairing, beyond recognizing a tool's functional part.

Task Planning. ① *Astra's planning lead spans all six sources.* It leads four sources and ties the best model on two (Figure 7a; Appendix D.4). ALFRED and ALFWorld success is 75% and 70% versus Fable's 35% and 30%. Astra alone solves ALFRED heating and dual placement (Figure 7b), extending its advantage to state changes and maintaining progress across multiple objects. ② *Astra gains most when scientific goals require linked operations.* DiscoveryWorld reaches 100% versus Fable's 70%; ScienceWorld reaches 87.7% versus 80.8%. Figure 7c locates the largest gains in unknown-plant genetics and melting-point measurement. Thermometer use reverses this lead at 66.7% versus 83.3%, identifying instrument operation as a distinct bottleneck within scientific planning. ③ *Astra finds seating targets but fails to establish the sitting state.* It ties Gemini for the best HumanCLAW contact rate (60%), yet every agent fails sitting. Five Astra episodes reach the seat and issue the action, whereas Fable records no target-finding success in these episodes. This execution evidence distinguishes locating a target from establishing the required body–object state.

Manipulation. ① *Astra's baseline lead spans several operation families.* It leads eight of thirteen suites among available baseline records (Figure 8a). Examples include CLIPort at 61.1% versus Fable's 33.3%, RLBench at 60% versus Sol's 20%, and RoboTwin at 38% versus Fable's 14%, spanning placement, articulation, and coordinated handling. Its five-of-nine ManiSkill success is supported by a 60% condition score on the diagnostic subset (Figure 8c). ② *Astra converts baseline control into longer goal sequences.* Astra completes 3.10 consecutive CALVIN goals on average versus Gemini's 1.80, Sol's 1.40, and Fable's 0.90 (Figure 8b). Only Astra finishes full sequences, on three of ten tasks. The expanded-condition markers in panels b–c test the setting dependence of this advantage in Section 4.4. ③ *Household execution exceeds Astra's local control strengths.* Astra completes one of ten RoboCasa tasks and none in RoboCasa365; all agents fail BEHAVIOR-1K



Figure 7: Astra's planning advantage across sources and goals. (a) Overall success for all six sources. (b) ALFRED/ALFWorld success on seven household goal families. (c) All five nonzero Astra–Fable gaps in ScienceWorld; thirteen families tie. K/U means known/unknown conditions. Single-run task outcomes; Appendix D.4 includes all seven models.

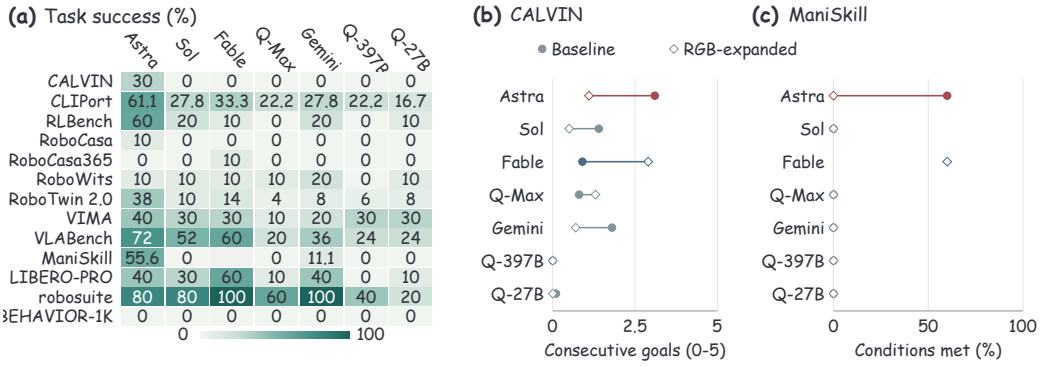

Figure 8: Manipulation across suites and execution settings. (a) Task success on thirteen suites; the last three rows share current results. (b) Average CALVIN goals completed consecutively before the first failure, out of five. (c) ManiSkill goal-condition satisfaction on five diagnostic tasks. Filled/open markers show Baseline/RGB-expanded settings (Section 4.4). Each task uses one run; Appendix D.5 gives metrics and coverage.

(Figure 8a). These tasks combine mobile-base positioning and articulated fixtures with object handling (Nasiriany et al., 2024). Household success requires local control to support coordinated navigation, placement, and environment-state changes.

4.4 Further Investigations

Richer observations under matched budgets. ① *Astra's control lead can reverse.* With inference budgets and execution-step limits fixed, per-round RGB changes CALVIN success from 30% to zero for Astra and zero to 40% for Fable (Table 3b); average consecutive-goal scores also reverse (Figure 8b). ② *Astra processes more input while executing fewer actions.* Astra's median input tokens rise from 0.32M to 1.21M, while action calls fall from 1,795 to 481. In sequence 0557, its reusable controller appears at round 12 rather than 5; expanded Fable builds one at round 2. This trajectory contrast links completion to when agents establish reusable feedback control (Appendix E.1). ③ *Stable totals hide opposing task effects.* VIMA improves for all three models, and RoboTwin improves for Astra. Across 153 common tasks, Astra stays at 63 successes while Fable rises from 34 to 65. ManiSkill condition scores likewise fall from 60% to zero for Astra, while expanded Fable reaches 60% (Figure 8c). Task-specific gains offset losses in Astra's aggregate.

Multi-round review on fixed inputs. Agents revisit initial answers on 46 shared cases. Self-review and Crop review allow two additional responses; Tool-assisted review allows six and adds specialist measurements (Appendix E.2). Tool-assisted review raises Astra's Affordance success from 22.2% to 33.3%, but improves five and worsens four Geometry entries (Table 3a). Astra's Spatial accuracy remains at 72.2%, while Qwen-27B rises from 55.6% to 61.1%. Qwen-27B's self-review yields

Table 3: Multi-round review (a) and RGB-expanded Manipulation (b). Baseline and RGB-expanded share budgets and execution-step limits; RGB-expanded supplies per-round images. I/T/W counts improved/tied/worsened entries; SR_{50} requires box IoU$\geq$ 0.5. Goals gives average consecutive completions (0–5); Common pools nine suites (153 tasks). Details: Appendix E.

(a)		Geometry	Spatial Reasoning	Affordance	
Model	Condition	I/T/W	Acc. (%)↑	Box IoU↑	SR_{50} (%)↑
Astra	Initial	–	**72.2**	0.2903	22.2
	Self-review	3/20/1	66.7	0.2914	22.2
	Crop review	4/19/1	66.7	0.2912	22.2
	Tool-assisted	5/15/4	**72.2**	**0.3174**	**33.3**
Qwen-27B	Initial	–	55.6	0.0336	0.0
	Self-review	6/15/3	50.0	**0.1524**	0.0
	Crop review	3/19/1	44.4	0.0937	0.0
	Tool-assisted	10/6/8	**61.1**	0.1452	**11.1**

(b)		CALVIN		ManiSkill	VIMA	Common
Model	Condition	SR (%)↑	Goals↑	SR (%)↑	SR (%)↑	SR (%)↑
Astra	Baseline	**30.0**	**3.10**	**55.6**	40.0	**41.2**
	RGB-expanded	0.0	1.10	11.1	**100.0**	**41.2**
Sol	Baseline	0.0	**1.40**	**0.0**	30.0	19.0
	RGB-expanded	**10.0**	0.50	**0.0**	**90.0**	**23.5**
Fable	Baseline	0.0	0.90	–	30.0	22.2
	RGB-expanded	**40.0**	**2.90**	55.6	**100.0**	**42.5**

higher box IoU than tool assistance (0.1524 versus 0.1452), yet tool assistance alone produces a box passing the 0.5 threshold. Mean overlap and threshold success capture different effects of review.

Case studies. Appendix F links spatial judgments and contact localization to concrete inputs, then compares controller construction and goal completion in CALVIN and PickCube. CloseLaptopLid and PullCube frames contrast visible motion with complete execution.

5 Conclusion

Findings. The gap to robot generalism lies in converting local competence into complete goal satisfaction. Astra's strongest advantages span geometric estimation, usable-contact localization, and compound household planning. Richer visual observations also change control rankings under matched execution limits. Across the arena, success depends on satisfying the task's coupled requirements: geometric boundaries, object relations, and final physical states. These requirements connect local progress to complete outcomes.

Limitations. The arena samples five mainstream robotic capability domains through bounded tasks and source-specific budgets. Tasks exceeding two hours, substantially larger budgets, and broader interactive repetition remain to be studied. Public task assets leave unseen-environment generalization open, while source-specific helpers vary the balance between skill selection and low-level control. Physical-robot transfer and sustained recovery from accumulated errors require further evaluation.

Future work. Future studies should vary observations and budgets independently, extend evaluation to physical robots and multi-hour tasks, and compare specialized controllers. Tracking error accumulation, recovery, and preservation of achieved conditions will reveal whether local progress persists across actions and changing scene configurations. The central test is whether agents can coordinate these capabilities and reliably complete robotic goals over extended operation.

AI USE STATEMENT

Generative AI assisted with literature retrieval, planning experimental analyses, data aggregation and analysis code, result interpretation, manuscript drafting and editing, and the overview logo and wordmark. Reported performance values derive from evaluation records and spatial-tracing summaries; task scenes are benchmark inputs or recorded observations. The authors are responsible for the final content.

REFERENCES

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, et al. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. *arXiv preprint arXiv:2204.01691*, 2022. URL https://arxiv.org/abs/2204.01691.

Eduardo Arnold, Jamie Wynn, Sara Vicente, Guillermo Garcia-Hernando, Áron Monszpart, Victor Adrian Prisacariu, Daniyar Turmukhambetov, and Eric Brachmann. Map-free Visual Relocalization: Metric Pose Relative to a Single Image. *arXiv preprint arXiv:2210.05494*, 2022. URL https://arxiv.org/abs/2210.05494.

Vineet Bhat, Sungsu Kim, Valts Blukis, Greg Heinrich, Prashanth Krishnamurthy, Ramesh Karri, Stan Birchfield, Farshad Khorrami, et al. BOP-ASK: Object-Interaction Reasoning for Vision-Language Models. *arXiv preprint arXiv:2511.16857*, 2025. URL https://arxiv.org/abs/2511.16857.

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. *arXiv preprint arXiv:2410.24164*, 2024. URL https://arxiv.org/abs/2410.24164.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025. URL https://arxiv.org/abs/2511.16719.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, et al. RoboTwin 2.0: A Scalable Data Generator and Benchmark with Strong Domain Randomization for Robust Bimanual Robotic Manipulation. *arXiv preprint arXiv:2506.18088*, 2025a. URL https://arxiv.org/abs/2506.18088.

Zeren Chen, Xiaoya Lu, Zhijie Zheng, Pengrui Li, Lehan He, Yijin Zhou, Jing Shao, Bohan Zhuang, and Lu Sheng. Geometrically-constrained agent for spatial reasoning. *arXiv preprint arXiv:2511.22659*, 2025b. doi: 10.48550/arXiv.2511.22659. URL https://arxiv.org/abs/2511.22659.

Zhili Cheng, Yuge Tu, Ran Li, Shiqi Dai, Jinyi Hu, Shengding Hu, Jiahao Li, Yang Shi, et al. EmbodiedEval: Evaluate Multimodal LLMs as Embodied Agents. *arXiv preprint arXiv:2501.11858*, 2025. URL https://arxiv.org/abs/2501.11858.

Seokju Cho, Ryo Hachiuma, Abhishek Badki, Hang Su, Byung-Kwan Lee, Chan Hee Song, Sifei Liu, Subhashree Radhakrishnan, Seungryong Kim, Yu-Chiang Frank Wang, and Min-Hung Chen. SpatialClaw: Rethinking action interface for agentic spatial reasoning. Preprint, 2026. URL https://spatialclaw.github.io/static/pdfs/spatialclaw.pdf.

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018. URL https://arxiv.org/abs/1712.07629.

Letian Fu, Justin Yu, Karim El-Refai, Ethan Kou, Haoru Xue, Huang Huang, Wenli Xiao, Guanzhi Wang, et al. CaP-X: A Framework for Benchmarking and Improving Coding Agents for Robot Manipulation. *arXiv preprint arXiv:2603.22435*, 2026. URL https://arxiv.org/abs/2603.22435.

Pengyuan Guo, Zhonghao Mai, Zhengtong Xu, Kaidi Zhang, Quan Khanh Luu, Heng Zhang, Zichen Miao, Arash Ajoudani, Zachary Kingston, Qiang Qiu, and Yu She. PLanAR: Planning-Language-Grounded Agentic Reasoning for Robot Manipulation. *arXiv preprint arXiv:2602.01662*, 2026. URL https://arxiv.org/abs/2602.01662.

Haojian Huang, Jiahao Shi, Yinchuan Li, and Yingcong Chen. Affordance agent harness: Verification-gated skill orchestration. *arXiv preprint arXiv:2605.00663*, 2026. URL https://arxiv.org/abs/2605.00663.

Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J. Davison. RLBench: The Robot Learning Benchmark & Learning Environment. *arXiv preprint arXiv:1909.12271*, 2019. URL https://arxiv.org/abs/1909.12271.

Peter Jansen, Marc-Alexandre Côté, Tushar Khot, Erin Bransom, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Oyvind Tafjord, and Peter Clark. DISCOVERYWORLD: A Virtual Environment for Developing and Evaluating Automated Scientific Discovery Agents. *arXiv preprint arXiv:2406.06769*, 2024. URL https://arxiv.org/abs/2406.06769.

Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, et al. VIMA: General Robot Manipulation with Multimodal Prompts. *arXiv preprint arXiv:2210.03094*, 2022. URL https://arxiv.org/abs/2210.03094.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, et al. OpenVLA: An Open-Source Vision-Language-Action Model. *arXiv preprint arXiv:2406.09246*, 2024. URL https://arxiv.org/abs/2406.09246.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, et al. BEHAVIOR-1K: A Human-Centered, Embodied AI Benchmark with 1,000 Everyday Activities and Realistic Simulation. *arXiv preprint arXiv:2403.09227*, 2024a. URL https://arxiv.org/abs/2403.09227.

Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, et al. Embodied Agent Interface: Benchmarking LLMs for Embodied Decision Making. *arXiv preprint arXiv:2410.07166*, 2024b. URL https://arxiv.org/abs/2410.07166.

Erich Liang, Roma Bhattacharjee, Sreemanti Dey, Rafael Moschopoulos, Caitlin Wang, Michel Liao, Grace Tan, Andrew Wang, et al. InFlux: A Benchmark for Self-Calibration of Dynamic Intrinsics of Video Cameras. *arXiv preprint arXiv:2510.23589*, 2025. URL https://arxiv.org/abs/2510.23589.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as Policies: Language Model Programs for Embodied Control. *arXiv preprint arXiv:2209.07753*, 2022. URL https://arxiv.org/abs/2209.07753.

Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://github.com/cvg/LightGlue.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso M de Melo, and Alan Yuille. 3DSRBench: A Comprehensive 3D Spatial Reasoning Benchmark. *arXiv preprint arXiv:2412.07825*, 2024. URL https://arxiv.org/abs/2412.07825.

Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. CALVIN: A Benchmark for Language-Conditioned Policy Learning for Long-Horizon Robot Manipulation Tasks. *arXiv preprint arXiv:2112.03227*, 2021. URL https://arxiv.org/abs/2112.03227.

Austin Myers, Ching L. Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance Detection of Tool Parts from Geometric Features. In *IEEE International Conference on Robotics and Automation*, 2015. URL https://users.umiacs.umd.edu/~fermulcm/affordance/part-affordance-dataset/index.html.

Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. RoboCasa: Large-Scale Simulation of Everyday Tasks for Generalist Robots. *arXiv preprint arXiv:2406.02523*, 2024. URL https://arxiv.org/abs/2406.02523.

Fei Ni, Min Zhang, Pengyi Li, Yifu Yuan, Lingfeng Zhang, Yuecheng Liu, Peilong Han, Longxin Kou, et al. Embodied Arena: A Comprehensive, Unified, and Evolving Evaluation Platform for Embodied AI. *arXiv preprint arXiv:2509.15273*, 2025. URL https://arxiv.org/abs/2509.15273.

Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. VirtualHome: Simulating Household Activities via Programs. *arXiv preprint arXiv:1806.07011*, 2018. URL https://arxiv.org/abs/1806.07011.

Yu Qi, Haibo Zhao, Ziyu Guo, Siyuan Ma, Ziyan Chen, Yaokun Han, Renrui Zhang, Zitiantao Lin, et al. Dissecting Embodied Abilities in Multimodal Language Models through Skill-level Evaluation and Diagnosis. *arXiv preprint arXiv:2510.08759*, 2025. URL https://arxiv.org/abs/2510.08759.

Shengyi Qian and David F. Fouhey. Understanding 3D Object Interaction from a Single Image. *arXiv preprint arXiv:2305.09664*, 2023. URL https://arxiv.org/abs/2305.09664.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. *arXiv preprint arXiv:1912.01734*, 2019. URL https://arxiv.org/abs/1912.01734.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning Text and Embodied Environments for Interactive Learning. *arXiv preprint arXiv:2010.03768*, 2020. URL https://arxiv.org/abs/2010.03768.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. CLIPort: What and Where Pathways for Robotic Manipulation. *arXiv preprint arXiv:2109.12098*, 2021. URL https://arxiv.org/abs/2109.12098.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. RoboSpatial: Teaching Spatial Understanding to 2D and 3D Vision-Language Models for Robotics. *arXiv preprint arXiv:2411.16537*, 2024. URL https://arxiv.org/abs/2411.16537.

Stone Tao, Fanbo Xiang, Arth Shukla, Yuzhe Qin, Xander Hinrichsen, Xiaodi Yuan, Chen Bao, Xinsong Lin, et al. ManiSkill3: GPU Parallelized Robotics Simulation and Rendering for Generalizable Embodied AI. *arXiv preprint arXiv:2410.00425*, 2024. URL https://arxiv.org/abs/2410.00425.

Alexander Veicht, Paul-Edouard Sarlin, Philipp Lindenberger, and Marc Pollefeys. GeoCalib: Single-image calibration with geometric optimization. In *European Conference on Computer Vision*, 2024. URL https://github.com/cvg/GeoCalib.

Hanqing Wang, Shaoyang Wang, Yiming Zhong, Zemin Yang, Jiamin Wang, Zhiqing Cui, Jiahao Yuan, Yifan Han, Mingyu Liu, and Yuexin Ma. Affordance-R1: Reinforcement Learning for Generalizable Affordance Reasoning in Multimodal Large Language Model. *arXiv preprint arXiv:2508.06206*, 2025a. URL https://arxiv.org/abs/2508.06206.

Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, et al. MindCube: Spatial Mental Modeling from Limited Views. *arXiv preprint arXiv:2506.21458*, 2025b. URL https://arxiv.org/abs/2506.21458.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. ScienceWorld: Is your Agent Smarter than a 5th Grader? *arXiv preprint arXiv:2203.07540*, 2022. URL https://arxiv.org/abs/2203.07540.

Dongming Wu, Yanping Fu, Saike Huang, Yingfei Liu, Fan Jia, Nian Liu, Feng Dai, Tiancai Wang, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Jianbing Shen. Ragnet: Large-scale reasoning-based affordance segmentation benchmark towards general grasping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11980–11990, October 2025.

Runsen Xu, Weiyao Wang, Hao Tang, Xingyu Chen, Xiaodong Wang, Fu-Jen Chu, Matt Feiszli, and Kevin J. Liang. Multi-SpatialMLLM: Multi-Frame Spatial Understanding with Multi-Modal Large Language Models. *arXiv preprint arXiv:2505.17015*, 2025. URL https://arxiv.org/abs/2505.17015.

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. *arXiv preprint arXiv:2412.14171*, 2024. URL https://arxiv.org/abs/2412.14171.

Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, et al. EmbodiedBench: Comprehensive Benchmarking Multi-modal Large Language Models for Vision-Driven Embodied Agents. *arXiv preprint arXiv:2502.09560*, 2025a. URL https://arxiv.org/abs/2502.09560.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, et al. MMSI-Bench: A Benchmark for Multi-Image Spatial Intelligence. *arXiv preprint arXiv:2505.23764*, 2025b. URL https://arxiv.org/abs/2505.23764.

Shiduo Zhang, Zhe Xu, Peiju Liu, Xiaopeng Yu, Yuan Li, Qinghui Gao, Zhaoye Fei, Zhangyue Yin, et al. VLABench: A Large-Scale Benchmark for Language-Conditioned Robotics Manipulation with Long-Horizon Reasoning Tasks. *arXiv preprint arXiv:2412.18194*, 2024. URL https://arxiv.org/abs/2412.18194.

Yehang Zhang, Jianchong Su, Haojian Huang, Yifan Chang, Tianhao Zhou, Xinli Xu, Yingjie Xu, Yinchuan Li, Zexi Li, and Ying-Cong Chen. WorldLines: Benchmarking and Modeling Long-Horizon Stateful Embodied Agents. *arXiv preprint arXiv:2606.18847*, 2026. URL https://arxiv.org/abs/2606.18847.

Zixin Zhang, Kanghao Chen, Hanqing Wang, Hongfei Zhang, Harold Haodong Chen, Chenfei Liao, Litao Guo, and Ying-Cong Chen. A4-Agent: An agentic framework for zero-shot affordance reasoning. *arXiv preprint arXiv:2512.14442*, 2025. URL https://arxiv.org/abs/2512.14442.

Enshen Zhou, Yibo Li, Jingkun An, Jiayuan Zhang, Shanyu Rong, Mengzhen Liu, Yi Han, Yuheng Ji, Huajie Tan, Jiawei He, Pengwei Wang, Zhongyuan Wang, Cheng Chi, Lu Sheng, and Shanghang Zhang. Towards spatial trace with reasoning in vision-language models for robotics. *arXiv preprint arXiv:2512.13660*, 2026. URL https://arxiv.org/abs/2512.13660.

APPENDIX: SUPPORTING EVIDENCE AND PROTOCOLS

Reading guide. Appendices A–B define task construction, interfaces, and scoring. Appendix C expands Table 2, including source-composition sensitivity and profile normalization. Appendix D follows the five domains of Section 4.3, from full task comparisons to Manipulation's auxiliary metric definitions. Appendix E specifies the execution-protocol and multi-round review comparisons, and Appendix F illustrates recorded actions and terminal outcomes.

A Benchmark Construction and Source Inventory

Embodied Agent Arena contains 1,000 cases from 32 established sources and our GeoProbe benchmark (Table A1). A case is one query or episode; multiple queries can share a scene. Benchmark-specific allocation uses seed 42 and targets diagnostic coverage of the five capabilities in Section 3.1.

Table A1: The unified 1,000-case arena, drawn from 32 established sources and our GeoProbe benchmark. Detailed source allocations follow below.

Domain	Cases	Share (%)	Task coverage
Geometry	370	37.0	Calibration, pose, scale, tracing
Spatial Reasoning	220	22.0	Relations, views, video
Affordance	70	7.0	Functional parts, contact
Task Planning	157	15.7	Household, navigation, science
Manipulation	183	18.3	Robot operations, sequences
Total	1,000	100.0	660 fixed-input; 340 interactive

Query vocabulary. Figure 2a counts each term once per query, balances weights across source groups and domains, and includes English translations of GeoProbe queries. GeoProbe's ten protocols form one source group.

Geometry (370 cases). InFlux-Real (Liang et al., 2025) contributes 30 single-frame intrinsic-estimation cases, of which four have invalid reference annotations and are excluded from continuous-error calculations. Map-free (Arnold et al., 2022) contributes 30 relative-pose image pairs with known camera intrinsics. MultiSPA-derived uses Map-free images and poses with the camera-motion question formulation associated with Multi-SpatialMLLM (Xu et al., 2025): 12 angle, ten distance, and eight displacement-vector questions. This local adaptation retains the Map-free image and pose references. The pose transformation and camera-center displacement have different coordinate conventions and are evaluated against their respective targets.

GeoProbe comprises ten diagnostic protocols: 168 queries over 130 unique RGB images, predominantly rendered in Blender, with four real-photo displacement cases. Shared scenes allow comparisons of calibration, pose, depth, displacement, and scale requirements. Table A2 gives the allocation.

TraceSpatial-Bench (Zhou et al., 2026), introduced with RoboTracer, adds paired 2D and 3D queries on 56 scenes. These 112 Geometry cases predict offline object trajectories, scored for reference agreement and geometric constraints.

Table A2: GeoProbe allocation across ten diagnostic protocols. Several protocols share images and scenes.

Diagnostic	Cases	Diagnostic	Cases
Object displacement, fixed camera	12	Full camera intrinsics	24
Object displacement, small view change	12	Focal length, centered principal point	12
Object displacement, large view change	12	Relative camera pose	24
Real tabletop displacement	4	Sparse depth	12
Object dimensions	12	Reference-based scale	44

Spatial Reasoning (220 cases). 3DSRBench (Ma et al., 2024) contributes 50 relation questions; 28 flip-tagged entries were displayed unflipped. Scores use corrected references for five affected answers under the unflipped presentation. MindCube (Wang et al., 2025b) contributes 50 multi-view cases: 39 among-view, six around-view, and five rotation. VSI-Bench (Yang et al., 2024) contributes 50 video questions: nine absolute-distance, nine object-size, fourteen relative-direction, six relative-distance, five counting, four room-size, two appearance-order, and one route-planning. Eight frames form its initial observation. RoboSpatial-Home (Song et al., 2024) contributes ten cases each for compatibility, configuration, and context. BOP-ASK (Bhat et al., 2025) contributes twelve relative-position and eight relative-depth questions. MMSI-Bench (Yang et al., 2025b) contributes

20 cross-image questions. Together these comprise 183 categorical, 27 numerical, and ten point-set questions, scored by the arena criteria in Appendix B.3.

Affordance (70 cases). UMD (Myers et al., 2015) contributes 30 views spanning five tool classes and six functions: grasp, cut, scoop, contain, pound, and support. Multiple views can depict the same object. The 20 ReasonAff-style cases adapt instruction-conditioned queries motivated by ReasonAff (Wang et al., 2025a) to 3DOI (Qian & Fouhey, 2023) interaction annotations. Another 20 RAGNet-3DOI cases use explicit action-conditioned queries (Wu et al., 2025). The two subsets use disjoint images. The arena uses these images and annotations to evaluate native point and bounding-region outputs.

Task Planning (157 cases). ALFRED (Shridhar et al., 2019) and ALFWorld (Shridhar et al., 2020) contribute 20 visual cases each, covering placement, heating, cooling, cleaning, containers, multiple objects, and illumination. ScienceWorld (Wang et al., 2022) contributes 73 text-based cases across eighteen topics. VirtualHome (Puig et al., 2018) contributes 21 symbolic programs in seven scenes. The composition analysis removes these 94 cases together. DiscoveryWorld (Jansen et al., 2024) contributes ten tasks with structured observations. HumanCLAW contributes thirteen Habitat navigation and sitting episodes: five sofa, three toilet, two chair, two plant, and one television target. The source interfaces thus span visual, textual, and symbolic observations.

Table A3: Manipulation task allocation across 13 source benchmarks. All agents use the same assigned task set.

Benchmark	Cases	Task coverage
CALVIN (Mees et al., 2021)	10	Ten sequences, each with five ordered subgoals
CLIPort (Shridhar et al., 2021)	18	Eighteen task and color-generalization configurations
RLBench (James et al., 2019)	10	Ten task configurations
RoboCasa (Nasiriany et al., 2024)	10	Seven atomic and three composite tasks
RoboCasa365	10	Three atomic and seven composite tasks
RoboWits	10	Ten official task configurations
RoboTwin 2.0 (Chen et al., 2025a)	50	Fifty distinct manipulation tasks
VIMA-Bench (Jiang et al., 2022)	10	Six placement-generalization and four new-task cases
VLABench (Zhang et al., 2024)	25	8 ID, 3 category, 4 commonsense, 7 semantic, 3 texture
ManiSkill (Tao et al., 2024)	10	Nine binary tabletop tasks; one quadruped reward task
LIBERO-PRO (Fu et al., 2026)	10	Ten object, spatial, goal, and multi-object configurations
robosuite	5	Two lift seeds; stacking, restacking, and wiping
BEHAVIOR-1K (Li et al., 2024a)	5	Five multi-object household activities

Manipulation (183 cases). Table A3 lists 182 binary tasks and one continuous-reward task. LIBERO-PRO, robosuite, and BEHAVIOR-1K use benchmark-specific code interfaces (Fu et al., 2026). RoboCasa365 uses variations designated for pretraining in the source inventory. Appendix B.1 specifies the native assistance available to agents.

B Evaluation Protocol and Reproducibility

B.1 Observations and Model-Directed Interaction

Geometry, Spatial Reasoning, and Affordance inspect fixed observations; Planning and Manipulation interact with an environment. All domains except Manipulation use a persistent Python runtime; Manipulation uses its native agent loop. Both enforce episode-level budgets and termination. Variables and environment state persist between code rounds, and images produced by a cell reach the next model request after that cell ends.

Table A4: Model-visible observations, actions, and feedback. Private scoring annotations are unavailable to the agent.

Domain	Observation support	Actions and feedback
Geometry	Image or image pair; intrinsics for Map-free	One final numerical submission
Spatial Reasoning	Images, crops, and video-frame queries	Evidence inspection and answer submission
Affordance	Image and interaction instruction	One point or bounding-region submission
Task Planning	Images, text, or symbolic state	Navigation, interaction, programs, public progress
Manipulation	Protocol-specific scene/state observations and native grounding helpers	Motion or skill calls; execution and task checks

ALFRED and visual ALFWorld expose observation, search, navigation, interaction, and memory primitives. CLIPort and VIMA expose instance inspection and pick-and-place actions. RoboTwin provides actor localization, pose inspection, planning probes, and execution. LIBERO-PRO, robosuite, and BEHAVIOR-1K expose benchmark-specific observation and action interfaces; RoboCasa includes fixture and contact localization. The model makes decisions using these native helpers; the baseline adds no specialist perception or verification pipeline. Geometry instead requests direct visual estimates, with pose-solver implementation excluded from its interface.

A runtime exception can be followed by another code round if budget remains; prior state changes are not automatically undone. Model requests, code rounds, primitive calls, and environment steps count different units. Stored visualization frames are also distinct from images actually sent to the model. Qualitative figures sample the recorded camera streams.

B.2 Models and Execution Budgets

Table A5: Evaluated models and configured identifiers. All seven agents are represented across the five domains.

Model label	Identifier
GPT-6 Astra	`gpt-6-astra`
GPT-5.6 Sol	`gpt-5.6-sol`
Claude Fable 5.1	`claude-fable-5-1`
Gemini 3.8 Flash	`gemini-3.8-flash`
Qwen 3.8 Max	`qwen3.8-max`
Qwen 3.5 397B-A17B	`qwen3.5-397b-a17b`
Qwen 3.5 27B	`qwen3.5-27b`

Fixed-input and Planning configurations use temperature zero. Baseline request limits are 1,800 output tokens and 360 seconds; fixed-input profiles also allow up to 16,384 tokens and 900 seconds. On the configured CLI interface, reasoning effort is maximum for Sol and high for Astra and Fable; API requests omit this parameter. Astra's first repetition on the original 90-case Geometry core uses the high-effort CLI configuration. Manipulation's request settings and shared execution limits are specified in Appendix E.1.

Table A6: Baseline episode budgets. Tokens accumulate within each case; rounds are code interactions, not repetitions. Dashes denote unspecified fields. Appendix B.4 lists additional fixed-input configurations.

Entry	Code rounds	Tokens	Env. steps	Primitives	Episode seconds	Task seconds
Geometry	12	120k	–	–	–	960
Spatial: non-VSI	6	96k	–	–	900	960
Spatial: VSI	8	240k	–	–	1,200	1,500
ALFRED, ALFWorld	8	320k	120	540	1,200	1,500
DiscoveryWorld	8	320k	100	400	1,200	1,500
HumanCLAW	101	2,000k	100	400	10,800	11,000
ScienceWorld	8	320k	100	500	1,200	1,500
VirtualHome	8	320k	16	120	1,200	1,500
Affordance	12	120k	–	–	–	960

Execution and native-call time limits are clipped by remaining budgets. Appendix E.1 gives Manipulation's shared execution limits, camera presets, and native step-limit semantics.

B.3 Scoring and Repeated Evaluation

Submission and termination. A final submission commits a fixed-input prediction. Interactive runs end when the agent stops, the environment reaches a terminal condition, or a source-specific round, token, action, or time limit is reached. The evaluator scores the submitted output or recorded state independently of the agent's declaration. A complete episode includes all its internal reasoning, actions, and recovery; these steps are not separate attempts.

Task scores and aggregation. Categorical questions use answer accuracy and interactive tasks use native goal success. Invalid or absent predictions remain in success denominators. Invalid reference annotations are excluded before scoring. Continuous errors pool valid predictions with compatible metrics and units, alongside their coverage. Numerical Spatial Reasoning uses exact-match success, $|\hat{y} - y| < 10^{-6}$, alongside continuous errors; this differs from VSI-Bench's official mean relative accuracy. Point-set questions retain their local set criterion.

Pass@1 and repeated scores. For N assigned cases and R repetitions, let $s_{ir} \in \{0, 1\}$ denote success on case i in run r. The average single-episode success rate is

$$\text{Pass@1} = \frac{1}{R} \sum_{r=1}^{R} \left(\frac{1}{N} \sum_{i=1}^{N} s_{ir} \right). \tag{2}$$

This estimates success in one complete episode under the specified condition. Fresh agent state is used for formal repetitions. For any per-run score M_r, we compute

$$\bar{M} = \frac{1}{R} \sum_{r=1}^{R} M_r, \qquad v_M = \frac{1}{R-1} \sum_{r=1}^{R} (M_r - \bar{M})^2, \qquad \text{SD}_M = \sqrt{v_M}. \tag{3}$$

Tables using $\bar{M} \pm \text{SD}_M$ report variation between run-level scores. Variance is supplied separately in squared metric units. Figure 4 averages valid-case errors within each run and then weights the recorded runs equally. Its coverage sensitivity instead fixes the cases valid in every shared repetition for each model pair; those common-cohort estimates are reported separately. Figure 5 uses per-run categorical accuracies; its pairwise gaps average accuracy differences over shared run indices, with the comparator fixed per subtask. Run indices align reporting repetitions; seeds and tool configurations can differ. Run-to-run SD is defined for $R > 1$.

Scalar estimates and depth. For J valid targets, $e_j = \hat{y}_j - y_j$ gives $\text{MAE} = J^{-1} \sum_j |e_j|$ and $\text{RMSE} = \sqrt{J^{-1} \sum_j e_j^2}$. Absolute relative error is $\text{AbsRel} = J^{-1} \sum_j |e_j|/|y_j|$ on nonzero targets, and $\text{MAPE} = 100\,\text{AbsRel}$ is expressed in percent. Size and focal-length estimates use absolute and relative errors. Depth additionally reports RMSE in meters. Depth errors are computed over positive finite predictions within each image, then averaged across images, with valid-point coverage

reported. The fraction within 10% relative error and $\delta_1 = J_{\text{all}}^{-1} \sum_j \mathbf{1}[\max(\hat{y}_j/y_j, y_j/\hat{y}_j) < 1.25]$ use all queried points as their denominator; invalid predictions contribute zero. Ordinal accuracy is the fraction of eligible point pairs with the correct depth ordering, with invalid predictions counted as incorrect. Scale-aligned errors quantify depth shape after removing a global scale factor.

Displacement and camera pose. For a nonzero scalar or vector target y, motion relative error is $\|\hat{y} - y\|_2 / \|y\|_2$. The MultiSPA-derived tolerance is 20%; a zero target requires zero absolute error. GeoProbe object-displacement diagnostics separately test movement status, direction, absolute distance error, and direction agreement with distance error at most 1 cm. Stationary cases are evaluated without division by zero. Camera translation error is $e_t = \|\hat{\mathbf{t}} - \mathbf{t}\|_2$, reported in centimeters. Rotation error is $e_R = \arccos[(\text{tr}(R^\mathsf{T}\hat{R}) - 1)/2]$, converted to degrees with the argument clipped to $[-1, 1]$. GeoProbe reports component errors and joint success within 2° and 2 cm.

Camera calibration. Focal MAPE averages the absolute relative errors of f_x and f_y per case. Principal-point error is $e_c = \|\hat{\mathbf{c}} - \mathbf{c}\|_2$ in pixels, also normalized by the image diagonal D. Joint calibration requires both focal components to satisfy 10% relative error and $e_c \leq 0.02D$. Main-table pose errors and complementary continuous metrics use pooled valid-observation means; Figure 4 instead uses equal-run means, as specified above.

Spatial tracing. For each case, predicted and reference paths are independently resampled to $L = 100$ equally spaced arc-length positions. Trajectory error is $\text{RMSE}_{\text{trace}} = \sqrt{L^{-1} \sum_{\ell=1}^{L} \|\hat{\mathbf{p}}_\ell - \mathbf{p}_\ell\|_2^2}$. We average this per-case value over valid scenes. The 2D metric uses normalized image coordinates; the 3D metric uses meters. The two metrics retain their respective units. The supplied variance is the population variance of the 56 per-scene RMSE values for a single run.

Offline 3D success requires valid initiation, target arrival, and collision checks to pass together. The predicted start must lie within 20 cm of the object point cloud or the reference start. At least one of the final three predicted points must lie within 20 cm of the target 3D box; for two-point paths, only the endpoint is checked. Translating the object point cloud to each checked predicted position must yield at most 20% overlap with the environment. Collision is checked at the sampled positions. Missing or invalid trajectories count as failures. 2D tracing reports continuous error. Qwen-27B trajectory outputs undergo format normalization that preserves all predicted coordinate values.

Native affordance grounding. For a directly predicted bounding box B and ground-truth mask G, the box interior is rasterized and $\text{IoU}_{\text{mask}} = |B \cap G|/|B \cup G|$. Contact validity is $\mathbf{1}[p \in G]$ for the submitted point p, with a reference-box fallback only where a mask is unavailable. A separate diagnostic uses $\text{IoU}_{\text{box}} = |B \cap B_G|/|B \cup B_G|$ against the reference box B_G. Strict box success requires $\text{IoU}_{\text{box}} \geq 0.5$; contact validity is scored independently. The source field named `giou` contains ordinary box–mask IoU, not generalized IoU.

Resource accounting. Inference tokens, action calls, model rounds, and elapsed time measure different aspects of execution. The Manipulation comparison reports token usage and action calls from the supplied execution analysis; protocol-matched per-case durations are unavailable. Request, trial, and native-call time limits are specified separately in Appendix E.1.

B.4 Repetition Protocol and Result Accounting

Repetitions. Non-tracing Geometry, Spatial Reasoning, and Affordance use five repetitions per case for six models and three for Fable. Spatial tracing, Planning, and Manipulation have one reported episode per assignment and therefore no run-to-run SD. Manipulation's two configurations appear in Appendix E.1.

Additional execution configurations. Fixed-input evaluation also uses per-case limits of up to 48 code rounds and 480k tokens for Geometry and Affordance, 24 rounds and 384k tokens for non-VSI Spatial Reasoning, and 32 rounds and 960k tokens for VSI. Task time limits reach 7,200 seconds, or 10,800 seconds for VSI. Budgets vary by source and execution configuration.

Result selection. Fixed-input and Planning trial slots use the first valid outcome, independent of correctness. Manipulation uses the fixed adopted baseline and RGB-expanded manifests. Table 2

follows the baseline CSV over 182 binary assignments. Appendix E.1 defines missing-record accounting, shared-current sources, and the common subset used in Table 3.

Coverage and scoring denominators. Table A7 reports valid fixed-input submissions and recorded Planning outcomes. InFlux excludes four invalid references. Binary criteria retain all assigned tasks and require verified success. Manipulation covers 182 binary tasks, including all five BEHAVIOR-1K assignments per model; AnymalC-Spin is excluded from this binary comparison. Continuous metrics retain finite valid predictions.

Table A7: Valid submission coverage for fixed-input domains and recorded terminal outcomes for Planning. Denominators include repetitions; Manipulation uses the fixed allocation in Table A3.

Domain	Astra	Sol	Fable	Q-Max	Gemini	Q-397B	Q-27B
Geometry	1382/1402	1382/1402	819/886	1382/1402	1314/1402	1382/1402	1376/1402
Spatial Reasoning	1060/1100	1100/1100	586/660	1097/1100	1077/1100	1033/1100	982/1100
Affordance	350/350	350/350	208/210	350/350	335/350	349/350	349/350
Task Planning	157/157	157/157	157/157	157/157	157/157	157/157	157/157

Manipulation diagnostics. CALVIN reports average consecutive goals completed. Other auxiliary metrics average task-normalized native criteria or control evidence over fixed applicable subsets. They differ in their scoring items and evidence time windows; Appendix D.5 defines the aggregation, coverage, and missing-record treatment.

Main-table metric scope. Geometry rotation and translation pool the 30 Map-free and 24 GeoProbe pose cases in degrees and centimeters. Spatial Pass@1 covers all 220 cases; numerical AbsRel covers 27, and categorical accuracy separately covers 183. Affordance mask IoU and contact validity cover all three sources, with 70 cases. Planning uses terminal success across 157 cases. Manipulation uses task-weighted native success across 182 binary assignments; Table A9 reports each source separately.

Dispersion and valid subsets. Pooled continuous statistics use $\bar{x} = K^{-1}\sum_{k=1}^{K} x_k$ and $v_{\text{pool}} = K^{-1}\sum_k (x_k - \bar{x})^2$ over valid case–run observations. This dispersion differs from the variance of run means in Equation 3. Main-table continuous metrics use pooled means; the task-level Geometry and Spatial analyses use equal-run means and SD, with common-valid-cohort checks in Appendices D.1 and D.2.

C Complete Model Comparisons

This section supports Section 4.2 at three resolutions: complementary native metrics, differences among source benchmarks, and the fixed transformations used in Figure 3. Task-level explanations follow in Appendix D.

C.1 Metrics and Source-Level Results

Table A8 gives complementary native metrics. Table A9 separates source results, with denominators and repetition rules from Appendix B.4.

Table A8: Complementary native metrics across four domains. Repeated binary rates show mean ± run SD; single-run rates have no SD. Continuous values are pooled valid-observation means. N counts eligible cases. GP denotes GeoProbe; Manipulation source results appear in Table A9.

Metric	N	Astra	Sol	Fable	Q-Max	Gemini	Q-397B	Q-27B
Geometry (370 cases)								
InFlux focal MAPE (%)↓	26	53.02	47.79	52.65	55.50	52.36	67.82	65.86
InFlux principal error (px)↓	26	3.82	3.82	3.82	3.82	3.88	3.82	3.82
Angle relative L2↓	12	0.760	0.915	1.32	1.41	0.962	1.30	2.04
Distance relative L2↓	10	0.294	0.418	0.785	0.560	0.419	0.411	0.331
Vector relative L2↓	8	0.425	0.591	0.625	0.454	0.647	1.02	1.03

Complementary metrics, continued.

Metric	N	Astra	Sol	Fable	Q-Max	Gemini	Q-397B	Q-27B
Map-free rotation (°)↓	30	29.94	46.60	52.75	35.11	46.32	63.03	64.94
Map-free translation (cm)↓	30	188.21	265.85	259.75	243.06	280.99	430.82	374.96
Map-free direction (°)↓	30	26.28	48.91	53.02	38.92	52.42	95.15	78.78
GP focal MAPE (%)↓	36	6.92	9.57	31.71	14.67	26.59	33.39	38.56
GP principal error (px)↓	24	51.95	110.62	63.40	78.41	66.83	58.23	59.05
Depth MAE (m)↓	12	0.307	0.633	0.412	0.490	0.589	0.934	0.944
Depth RMSE (m)↓	12	0.338	0.753	0.517	0.598	0.699	1.05	1.06
Depth AbsRel↓	12	0.104	0.229	0.140	0.163	0.193	0.312	0.292
Displacement MAE (cm)↓	40	0.585	0.973	1.49	1.52	2.28	2.48	2.94
GP rotation (°)↓	24	1.17	2.54	2.83	2.21	3.10	7.55	5.26
GP translation (cm)↓	24	5.36	9.50	9.30	6.99	7.85	23.59	12.05
Scale MAPE (%)↓	44	1.90	2.84	3.76	3.77	3.93	16.07	16.45
Scale judgment Pass@1 (%)↑	12	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0	93.3±7.0
2D trace RMSE↓	56	0.120	0.132	0.118	0.118	0.127	0.155	0.284
3D trace RMSE (m)↓	56	0.382	0.463	0.435	0.469	0.426	0.629	0.682
3D constraint success (%)↑	56	33.9	21.4	25.0	17.9	37.5	8.9	7.1
Spatial Reasoning (220 cases)								
Native Pass@1 (%)↑	220	69.4±1.0	60.7±2.5	62.0±4.6	68.9±1.1	65.4±1.3	50.6±1.8	48.0±0.9
Categorical Pass@1 (%)↑	183	82.2±0.6	71.3±3.0	72.7±5.2	81.2±1.3	76.7±1.9	60.0±2.2	56.8±0.7
Numerical AbsRel↓	27	0.213	0.273	0.186	0.205	0.161	0.251	0.288
Affordance (70 cases)								
BBox–GT box IoU↑	70	0.390	0.366	0.386	0.382	0.322	0.146	0.021
BBox–GT mask IoU↑	70	0.286	0.264	0.284	0.272	0.250	0.116	0.013
BBox success@0.5 (%)↑	70	46.6±1.3	43.7±1.6	46.2±0.8	45.7±1.7	33.1±11.8	5.4±3.7	0.3±0.6
Contact-point validity (%)↑	70	60.3±2.6	52.0±3.6	57.1±2.5	52.9±4.2	44.6±13.0	32.6±4.7	6.3±1.6
Task Planning (157 cases)								
Native Pass@1 (%)↑	157	80.3	44.6	63.7	35.7	48.4	17.8	19.1

Table A9: Source results. Spatial/Affordance: mean ± run SD (%); Planning: solved/assigned cases; Manipulation: success (%). N counts eligible cases.

Source	N	Astra	Sol	Fable	Q-Max	Gemini	Q-397B	Q-27B
Spatial Reasoning: categorical accuracy (%)								
3DSRBench	50	73.6 ± 2.6	71.2 ± 1.8	72.7 ± 3.1	74.0 ± 3.7	70.0 ± 2.4	69.2 ± 3.6	57.6 ± 2.6
BOP-ASK	20	96.0 ± 4.2	87.0 ± 4.5	88.3 ± 2.9	92.0 ± 2.7	97.0 ± 2.7	85.0 ± 3.5	84.0 ± 4.2
MindCube	50	90.8 ± 1.1	68.0 ± 7.3	79.3 ± 8.1	90.8 ± 2.7	89.2 ± 3.0	46.0 ± 2.8	50.4 ± 4.1
MMSI-Bench	20	77.0 ± 5.7	55.0 ± 11.7	66.7 ± 7.6	55.0 ± 8.7	56.0 ± 9.6	35.0 ± 9.4	29.0 ± 7.4
RoboSpatial	20	85.0 ± 3.5	81.0 ± 8.9	88.3 ± 2.9	88.0 ± 5.7	76.0 ± 4.2	90.0 ± 3.5	92.0 ± 5.7
VSI-Bench	23	72.2 ± 7.3	70.4 ± 5.7	36.2 ± 28.9	83.5 ± 8.4	65.2 ± 5.3	44.3 ± 6.4	39.1 ± 3.1
Affordance: strict box success at IoU ≥ 0.5 (%)								
ReasonAff-style	20	19.0 ± 5.5	6.0 ± 4.2	16.7 ± 7.6	13.0 ± 5.7	5.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
RAGNet-3DOI	20	14.0 ± 4.2	15.0 ± 5.0	13.3 ± 2.9	7.0 ± 2.7	6.0 ± 2.2	4.0 ± 2.2	1.0 ± 2.2
UMD	30	86.7 ± 3.3	88.0 ± 4.5	87.8 ± 5.1	93.3 ± 2.4	70.0 ± 27.0	10.0 ± 7.8	0.0 ± 0.0
Task Planning: terminal success (single run)								
ALFRED	20	15/20	3/20	7/20	3/20	2/20	1/20	0/20
ALFWorld	20	14/20	4/20	6/20	2/20	4/20	3/20	4/20
DiscoveryWorld	10	10/10	4/10	7/10	3/10	5/10	1/10	1/10
HumanCLAW	13	3/13	2/13	1/13	0/13	3/13	0/13	0/13
ScienceWorld	73	64/73	38/73	59/73	29/73	45/73	11/73	14/73
VirtualHome	21	20/21	19/21	20/21	19/21	17/21	12/21	11/21
Total	157	126/157	70/157	100/157	56/157	76/157	28/157	30/157
Manipulation: baseline verified success (%)								
CALVIN	10	30.0	0.0	0.0	0.0	0.0	0.0	0.0
CLIPort	18	61.1	27.8	33.3	22.2	27.8	22.2	16.7
RLBench	10	60.0	20.0	10.0	0.0	20.0	0.0	10.0
RoboCasa	10	10.0	0.0	0.0	0.0	0.0	0.0	0.0
RoboCasa365	10	0.0	0.0	10.0	0.0	0.0	0.0	0.0

Source-level comparisons, continued.

Source	N	Astra	Sol	Fable	Q-Max	Gemini	Q-397B	Q-27B
RoboTwin 2.0	50	38.0	10.0	14.0	4.0	8.0	6.0	8.0
RoboWits	10	10.0	10.0	10.0	10.0	20.0	0.0	10.0
VIMA-Bench	10	40.0	30.0	30.0	10.0	20.0	30.0	30.0
VLABench	25	72.0	52.0	60.0	20.0	36.0	24.0	24.0
ManiSkill	9	55.6	0.0	–	0.0	11.1	0.0	0.0
LIBERO-PRO	10	40.0	30.0	60.0	10.0	40.0	0.0	10.0
robosuite	5	80.0	80.0	100.0	60.0	100.0	40.0	20.0
BEHAVIOR-1K	5	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Total	182	41.8	19.8	24.7†	9.3	18.7	9.9	11.0

†Coverage, source configurations, and aggregation are specified in Appendix E.1.

Reading the metric families. GeoProbe focal MAPE combines full-intrinsics and focal-only queries; displacement MAE combines its four camera/environment conditions. Scale MAPE combines object dimensions and numerical reference-scale queries, while qualitative scale judgments retain Pass@1. TraceSpatial reports 2D and 3D errors separately because their coordinates and units differ. Each metric retains the target's physical interpretation.

C.2 Sensitivity to Source Composition

Table A10: Task-composition analysis (success %). Spatial uses only the 183 categorical queries; Planning removes ScienceWorld and VirtualHome; Affordance removes UMD and uses strict box success. All columns retain the assigned-case denominator of the indicated subset.

Agent	Spatial without VSI	Planning without text/symbolic	Affordance without UMD
Astra	83.6	66.7	16.5
Sol	71.4	20.6	10.5
Fable	77.9	33.3	15.0
Qwen-Max	80.9	12.7	10.0
Gemini	78.4	22.2	5.5
Qwen-397B	62.2	7.9	2.0
Qwen-27B	59.4	7.9	0.5

Removing ScienceWorld and VirtualHome changes Astra's Planning success from 80.3% to 66.7%, and Qwen-Max's from 35.7% to 12.7%. The same table isolates categorical Spatial Reasoning without VSI-Bench and strict affordance-box success without UMD. These subset comparisons measure sensitivity to the task mixture.

Equal source weighting. To test whether source allocation drives the lead, we average complete source scores with equal weight, $M_{\text{source}} = B^{-1} \sum_{b=1}^{B} M_b$, retaining every assigned case within each source. Across all six Spatial sources, Astra scores 71.9% versus Qwen-Max's 69.1%; across all three Affordance sources, contact validity is 54.2% versus Fable's 50.9%; across all six Planning sources, success is 75.2% versus Fable's 53.1%. These are the highest non-Astra macro scores. The ranking therefore persists under both case weighting and equal source weighting. These analyses use the complete domain allocations, separate from the 46-case multi-round review diagnostic in Section 4.4.

C.3 Interpreting the Capability Profiles

Fixed metric transformations. Figures 3 and A1 use a visualization index derived from Table 2. Geometry covers the 54 pose cases summarized in the main table. With rotation error r in degrees and translation error u in centimeters, its score is $50[\exp(-r/45) + \exp(-u/150)]$. Spatial Reasoning averages native Pass@1 (in percent) and $100\exp(-e/0.25)$, where e is numerical AbsRel. Affordance averages $100\,\text{IoU}$ and contact validity in percent. Planning and Manipulation use domain

success percentages; Manipulation follows the baseline aggregation in Table 2. The five-domain index is their arithmetic mean. The same reference scales apply to all models. Native metrics provide the task-level comparisons in the text.

Interpretation. The individual radars retain the same axes, transformations, and limits for every model. The index supports the overview in Figure 3; comparisons of accuracy, precision, and completion use the native metrics in Tables 2 and A8.

D Task-Level Analysis of Astra

The subsections follow Section 4.3 and Figures 4–8. Each first identifies the task requirement behind the main-text claim, then gives the supporting comparison and broader task coverage. Repetition and valid-coverage conventions are defined in Appendix B.4.

D.1 Geometry: Task Requirements and Source Conditions

Task-level comparisons. Table A11 supplies the absolute errors behind Figure 4a. Its relative reduction is $100(1 - e_{\text{Astra}}/e_{\text{peer}})$, using the lowest-error non-Astra peer for each target. Equal-run means retain task-specific units: MAPE for focal length, size, and height, and $100 \times \text{AbsRel}$ for depth. Astra leads the five GeoProbe targets; Sol leads on InFlux focal estimation.

Table A11: Geometric errors (%) across seven agents, mean ± run SD. GP denotes GeoProbe; lower is better.

Task	Astra	Sol	Fable	Q-Max	Gemini	Q-397B	Q-27B
GP focal-only	6.2±2.1	11.2±5.5	29.5±1.8	13.8±3.1	23.7±12.9	38.6±3.2	38.4±5.0
InFlux focal	53.0±2.1	47.8±2.0	52.6±2.2	55.5±1.0	52.4±3.4	67.8±1.8	65.9±2.1
GP depth	10.4±2.9	22.9±2.7	14.2±2.7	16.3±1.3	19.2±2.1	31.2±2.8	29.2±1.3
GP object size	6.3±1.0	8.5±0.7	7.6±0.2	9.6±0.8	8.3±0.3	11.4±0.6	9.4±0.3
GP ref. height	0.31±0.11	0.90±0.36	2.31±0.08	1.69±0.72	2.73±0.75	17.55±1.22	13.69±0.86
GP height ratio	0.32±0.15	0.95±0.41	2.23±0.17	2.17±0.22	2.67±0.28	20.35±0.87	17.08±2.70

Coverage sensitivity. Table A12 restricts each Astra–peer comparison to cases valid in every shared run. All six comparison directions remain unchanged. The ranking is thus stable on the common valid cohort.

Table A12: Errors (%) on the common-valid cohort. Both agents use the same cases and runs within each row.

Task	Cases	Rounds	Peer	Astra	Peer
GP focal-only	12	5	Sol	6.21±2.10	11.16±5.47
InFlux focal	26	5	Sol	53.02±2.15	47.79±2.00
GP depth	9	3	Fable	10.16±3.49	14.22±2.88
GP object size	12	3	Fable	6.44±1.36	7.63±0.18
GP ref. height	12	5	Sol	0.31±0.11	0.90±0.36
GP height ratio	12	5	Sol	0.32±0.15	0.95±0.41

Camera motion and real photographs. The rendered displacement subsets share 12 object interventions while changing the camera condition. Table A13 compares Astra with Sol, the lowest-MAE non-Astra agent under each displacement condition. The fixed-camera advantage persists when the two stationary Blender cases are excluded. Under the native direction-and-1 cm criterion, both agents achieve 96.7% fixed-camera success: the continuous error distinguishes precision that this tolerance does not resolve.

The four real-photo pairs form a separate tabletop subset. Their displacement labels are supplied annotations, without independent camera or metric calibration. The Blender comparison uses different scenes and object motions.

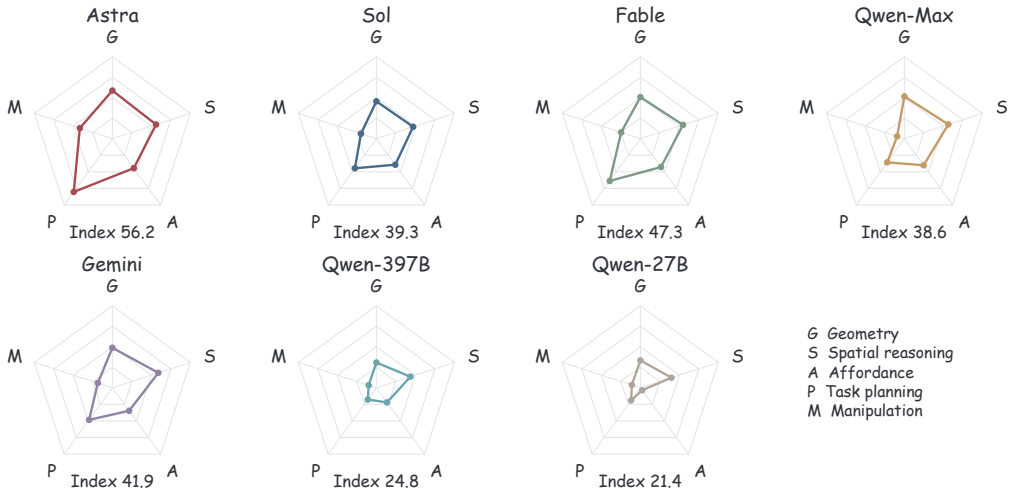

Figure A1: Individual capability profiles with the same axes and fixed scales as Figure 3a. G, S, A, P, and M denote Geometry, Spatial Reasoning, Affordance, Task Planning, and Manipulation. All axes derive from the metrics in Table 2.

Table A13: Displacement MAE (cm), mean ± run SD. Sol is the strongest non-Astra comparator under each listed condition. Moving-only excludes stationary cases.

Condition	Cases	Astra	Sol
Blender: fixed camera	12	0.240±0.109	0.467±0.267
Blender: fixed, moving-only	10	0.288±0.131	0.560±0.321
Blender: small view change	12	0.567±0.158	0.678±0.152
Blender: large view change	12	1.125±0.144	1.958±0.295
Real photographs	4	0.050±0.068	0.415±0.236

Coupled motion and metric scale reveal distinct estimation errors. Table A14 separates the GeoProbe pose modes: mixed rotation and translation produce 14.16 cm translation error and zero joint success, despite much lower errors on simpler motions. The best non-Astra translation estimate on the mixed subset is Qwen-Max at 16.77 cm; no evaluated agent achieves joint success on this subset. Translation-only cases have zero target rotation. Depth also separates ordering from metric scale: Astra's ordinal accuracy is 96.3%, whereas raw AbsRel is 0.104; oracle scale alignment reduces it to 0.038. The aligned diagnostic fits a global scale using reference depth.

Explicit scale references and regular rendered geometry provide useful cues for these estimates.

Table A14: GeoProbe camera-pose modes. Errors are pooled valid-observation means; joint success uses 2°/2 cm tolerances and shows mean ± run SD.

Motion	Cases	Rotation (°)	Translation (cm)	Joint success (%)
Stationary	4	0.00	0.00	100.0±0.0
Rotation only	6	0.35	1.00	96.7±7.5
Translation only	6	0.00	1.55	76.7±25.3
Rotation + translation	8	3.26	14.16	0.0±0.0

TraceSpatial-Bench: path matching and boundary constraints. Astra has lower 3D path RMSE than Gemini (0.382 versus 0.426 m), but fewer jointly valid paths (19/56 versus 21/56; Table A8). Among Astra's trajectories, 27 fail initiation, 24 fail arrival, and ten fail collision checks; these sets overlap. Of its 37 joint failures, 36 violate initiation or arrival, while only one satisfies both boundaries and fails collision avoidance alone. This supports the boundary-anchoring diagnosis in Figure 4c. The offline tolerances and sampled-position collision rule are defined in Appendix B.3.

D.2 Spatial Reasoning: Cross-Source Task Contrasts

Complete task coverage. Table A16 covers all 43 groups across six sources. Source categories are retained; MindCube uses task axes and question templates, prioritizing composed rotation and turn–translation before individual motion and relation queries.

Figure labels and comparisons. Figure 5a shows BOP-ASK left/right (L/R), MindCube relations (Rel.) and camera translation (Shift), 3DSRBench actor-relative side (Actor) and object orientation (Facing), and VSI hard directions (Hard). Panel b compares BOP camera depth (Near/Far), 3DSR camera/object distance (Cam./Obj.), and VSI relative distance (Rel.); each displayed peer leads or ties on a row. Source scenes are retained; the Obj. row has two cases.

Shared-round gaps. Panel c selects one task per source and its highest-mean non-Astra peer. Gaps and SD then use shared run indices (Table A15; protocol in Appendix B.3). On object-facing queries, Fable's three-run mean is 72.2% versus Astra's 50.0%; Astra's full five-run mean is 56.7%.

Table A15: Source-specific Astra–peer comparisons for Figure 5c. Accuracy and gaps use shared available runs and are in percentages and percentage points. Run-index alignment follows Appendix B.3.

Task	Peer	Rounds	Astra	Peer	Gap ± SD
MMSI camera–region	Fable	3	100.0	77.8	+22.2±19.2
VSI medium direction	Qwen-Max	5	100.0	83.3	+16.7±11.8
BOP-ASK left/right	Gemini	5	95.0	96.7	−1.7±7.0
MindCube camera shift	Qwen-Max	5	78.5	89.2	−10.8±14.0
3DSR object facing	Fable	3	50.0	72.2	−22.2±9.6
RoboSpatial configuration	Q-397B	5	78.0	100.0	−22.0±4.5

Table A16: All Spatial task groups, separated by benchmark. Entries are mean ± run SD (%); bold means mark the highest score in each row, including ties. Astra's column is shaded. N counts distinct cases; each assigned case remains in the success denominator.

3DSRBench								50 cases
Task	N	**Astra**	Sol	Fable	Q-Max	Gemini	Q-397B	Q-27B
World height	4	60.0±13.7	50.0±0.0	**66.7**±14.4	55.0±11.2	25.0±0.0	50.0±0.0	50.0±0.0
Above / below	7	**77.1**±7.8	60.0±12.0	71.4±0.0	**77.1**±12.8	74.3±12.0	74.3±6.4	68.6±12.0
Camera depth	10	90.0±7.1	98.0±4.5	93.3±5.8	90.0±0.0	92.0±4.5	96.0±5.5	**100.0**±0.0
Next to	1	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	**60.0**±54.8	0.0±0.0
Object distance	2	70.0±27.4	80.0±27.4	83.3±28.9	**100.0**±0.0	**100.0**±0.0	50.0±0.0	80.0±27.4
Parallel / perpendicular	1	**100.0**±0.0	**100.0**±0.0	**100.0**±0.0	**100.0**±0.0	20.0±44.7	**100.0**±0.0	80.0±44.7
Same/different facing	4	55.0±11.2	50.0±0.0	50.0±0.0	50.0±0.0	50.0±0.0	**75.0**±17.7	35.0±13.7
Object-facing side	6	56.7±9.1	53.3±13.9	**72.2**±9.6	66.7±0.0	66.7±0.0	36.7±18.3	23.3±9.1
In front of actor	4	**100.0**±0.0	**100.0**±0.0	**100.0**±0.0	**100.0**±0.0	**100.0**±0.0	**100.0**±0.0	95.0±11.2
Actor left/right	7	**74.3**±6.4	71.4±14.3	61.9±8.2	71.4±10.1	65.7±7.8	45.7±12.0	5.7±7.8
Object viewpoint	4	**70.0**±11.2	**70.0**±11.2	50.0±0.0	60.0±13.7	**70.0**±11.2	**70.0**±11.2	65.0±28.5
BOP-ASK								20 cases
Task	N	**Astra**	Sol	Fable	Q-Max	Gemini	Q-397B	Q-27B
Closer to camera	5	96.0±8.9	96.0±8.9	86.7±11.5	**100.0**±0.0	96.0±8.9	80.0±14.1	84.0±8.9
Farther from camera	3	**100.0**±0.0	**100.0**±0.0	**100.0**±0.0	**100.0**±0.0	**100.0**±0.0	66.7±0.0	40.0±14.9
Left / right	12	95.0±4.6	80.0±4.6	86.1±4.8	86.7±4.6	**96.7**±4.6	91.7±0.0	95.0±4.6

Table A16 continued. Values are mean ± run SD (%).

MindCube								50 cases
Task	N	**Astra**	Sol	Fable	Q-Max	Gemini	Q-397B	Q-27B
Behind observer	7	97.1±6.4	**100.0**±0.0	**100.0**±0.0	**100.0**±0.0	88.6±12.0	80.0±16.3	65.7±16.3
Camera translation	13	78.5±10.0	64.6±8.8	56.4±16.0	**89.2**±6.9	84.6±5.4	20.0±6.9	35.4±12.9
Imagined observer	2	**90.0**±22.4	70.0±27.4	50.0±0.0	80.0±27.4	70.0±27.4	30.0±27.4	30.0±27.4
Object relations	12	**96.7**±4.6	63.3±12.6	91.7±8.3	85.0±3.7	95.0±7.5	31.7±7.0	35.0±3.7
Composed rotation	5	**100.0**±0.0	96.0±8.9	**100.0**±0.0	**100.0**±0.0	**100.0**±0.0	**100.0**±0.0	96.0±8.9
Turn then translate	11	**90.9**±6.4	43.6±7.6	75.8±10.5	**90.9**±6.4	87.3±10.4	49.1±13.8	58.2±10.4
MMSI								20 cases
Task	N	**Astra**	Sol	Fable	Q-Max	Gemini	Q-397B	Q-27B
Attribute (Appr.)	2	**90.0**±22.4	70.0±27.4	50.0±0.0	60.0±22.4	50.0±0.0	70.0±27.4	40.0±22.4
Attribute (Meas.)	2	**100.0**±0.0	50.0±0.0	66.7±57.7	40.0±22.4	50.0±0.0	0.0±0.0	40.0±22.4
MSR	2	40.0±41.8	30.0±27.4	**50.0**±0.0	40.0±22.4	30.0±27.4	30.0±44.7	40.0±22.4
Motion (Cam.)	1	80.0±44.7	80.0±44.7	**100.0**±0.0	**100.0**±0.0	60.0±54.8	0.0±0.0	0.0±0.0
Motion (Obj.)	1	**100.0**±0.0	80.0±44.7	**100.0**±0.0	80.0±44.7	80.0±44.7	80.0±44.7	20.0±44.7
Camera–camera relation	3	**100.0**±0.0	73.3±27.9	**100.0**±0.0	66.7±0.0	73.3±14.9	26.7±14.9	33.3±0.0
Camera–object relation	1	0.0±0.0	40.0±54.8	**66.7**±57.7	0.0±0.0	60.0±54.8	40.0±54.8	0.0±0.0
Camera–region relation	3	**100.0**±0.0	60.0±14.9	77.8±19.2	66.7±0.0	73.3±14.9	66.7±23.6	53.3±18.3
Object–object relation	1	**100.0**±0.0	80.0±44.7	**100.0**±0.0	60.0±54.8	20.0±44.7	80.0±44.7	60.0±54.8
Object–region relation	4	**50.0**±0.0	30.0±11.2	25.0±25.0	45.0±20.9	**50.0**±30.6	5.0±11.2	0.0±0.0
RoboSpatial								30 cases
Task	N	**Astra**	Sol	Fable	Q-Max	Gemini	Q-397B	Q-27B
Placement compatibility	10	92.0±4.5	92.0±8.4	**96.7**±5.8	90.0±0.0	64.0±8.9	80.0±7.1	90.0±7.1
Configuration	10	78.0±4.5	70.0±10.0	80.0±0.0	86.0±11.4	88.0±4.5	**100.0**±0.0	94.0±8.9
Free-space point set	10	**0.0**±0.0	**0.0**±0.0	**0.0**±0.0	**0.0**±0.0	**0.0**±0.0	**0.0**±0.0	**0.0**±0.0
VSI-Bench								50 cases
Task	N	**Astra**	Sol	Fable	Q-Max	Gemini	Q-397B	Q-27B
Appearance order	2	40.0±22.4	80.0±27.4	50.0±50.0	**100.0**±0.0	40.0±41.8	40.0±22.4	80.0±27.4
Absolute distance	9	2.2±5.0	2.2±5.0	3.7±6.4	0.0±0.0	**6.7**±6.1	0.0±0.0	2.2±5.0
Object count	5	32.0±11.0	20.0±14.1	**40.0**±20.0	36.0±16.7	36.0±16.7	28.0±11.0	12.0±11.0
Direction: easy	4	**100.0**±0.0	85.0±22.4	33.3±28.9	**100.0**±0.0	85.0±13.7	30.0±20.9	45.0±11.2
Direction: hard	4	45.0±11.2	30.0±27.4	25.0±43.3	**65.0**±28.5	60.0±22.4	30.0±32.6	5.0±11.2
Direction: medium	6	**100.0**±0.0	73.3±9.1	38.9±38.5	83.3±11.8	66.7±11.8	40.0±14.9	20.0±13.9
Relative distance	6	50.0±20.4	**76.7**±22.4	38.9±9.6	**76.7**±9.1	56.7±14.9	60.0±19.0	53.3±13.9
Object size	9	4.4±6.1	**22.2**±7.9	11.1±0.0	13.3±5.0	11.1±7.9	2.2±5.0	8.9±5.0
Room area	4	**0.0**±0.0	**0.0**±0.0	**0.0**±0.0	**0.0**±0.0	**0.0**±0.0	**0.0**±0.0	**0.0**±0.0
Route planning	1	**100.0**±0.0	**100.0**±0.0	33.3±57.7	**100.0**±0.0	**100.0**±0.0	**100.0**±0.0	**100.0**±0.0

Numerical error and acceptance are complementary. The numerical rows in Table A16 use the exact-match criterion from Appendix B.3. Table A17 instead measures continuous error on valid predictions. Fixing Astra's cohort across runs preserves low object-size AbsRel (0.091), while absolute-distance coverage is only 25/45 submissions (Table A18). The table reports error together with valid-prediction coverage.

Table A17: VSI-Bench AbsRel, equal-run mean ± SD on valid predictions; lower is better. Each model uses its valid-prediction cohort; coverage is reported below.

Task	Astra	Sol	Fable	Q-Max	Gemini	Q-397B	Q-27B
Object count	0.35±0.05	0.40±0.11	0.20±0.11	0.23±0.11	0.31±0.08	0.41±0.06	0.42±0.04
Absolute distance	0.28±0.11	0.42±0.05	0.28±0.07	0.26±0.02	0.12±0.03	0.29±0.09	0.31±0.08
Object size	0.09±0.01	0.09±0.02	0.10±0.01	0.12±0.02	0.05±0.01	0.13±0.02	0.19±0.02
Room area	0.24±0.03	0.19±0.03	0.21±0.04	0.26±0.07	0.27±0.04	0.22±0.05	0.26±0.10

Table A18: Astra numerical coverage and a fixed-cohort sensitivity check. Common cases have a valid numerical prediction in every recorded round.

Task	Cases	Valid / assigned	Common cases	Common AbsRel
Object count	5	25/25	5	0.350 ± 0.053
Absolute distance	9	25/45	2	0.279 ± 0.083
Object size	9	42/45	8	0.091 ± 0.017
Room area	4	20/20	4	0.240 ± 0.029

Other task requirements. RoboSpatial separates placement compatibility (92.0%), configuration (78.0%), and free-space point sets (0.0%) for Astra. All agents fail the point-set criterion. The scores separate successful compatibility and configuration judgments from failure on point-set submissions.

D.3 Affordance: Contact and Region Extent

From valid contact to functional-region extent. Table A19 retains all fifteen action/hand groups, including zero-scoring conditions. UMD contact validity averages 96.7% for Astra and 94.4% for Fable. Cutting combines 100% valid contacts and 96% strict box success with mask IoU of 0.361; the box IoU is 0.694 (Figure 6b). This separates successful localization from the limitations of rectangular mask coverage. Containment adds a same-representation comparison: Astra's strict box success is 44%, versus Qwen-Max's 96% (Table A20).

Table A19: Astra Affordance metrics, mean $\pm$ run SD. IoU compares the native box to the GT mask; strict box success compares it to the GT box at IoU ≥ 0.5. Point and box success are percentages.

Action / hands	Cases	Mask IoU	Point (%)	Box (%)
RAGNet-3DOI				
free movement / two hands	2	0.046 ± 0.063	20.0 ± 27.4	20.0 ± 27.4
free / one hand	1	0.384 ± 0.351	60.0 ± 54.8	60.0 ± 54.8
free / two hands	5	0.249 ± 0.070	60.0 ± 14.1	36.0 ± 16.7
pull / one hand	12	0.008 ± 0.002	25.0 ± 5.9	0.0 ± 0.0
ReasonAff-style				
free movement / two hands	5	0.270 ± 0.055	48.0 ± 11.0	48.0 ± 11.0
free / one hand	4	0.000 ± 0.000	0.0 ± 0.0	0.0 ± 0.0
free / two hands	4	0.001 ± 0.002	0.0 ± 0.0	0.0 ± 0.0
pull / one hand	4	0.098 ± 0.033	60.0 ± 13.7	0.0 ± 0.0
push / one hand	3	0.318 ± 0.221	46.7 ± 29.8	46.7 ± 29.8
UMD				
contain	5	0.472 ± 0.067	100.0 ± 0.0	44.0 ± 16.7
cut	5	0.361 ± 0.009	100.0 ± 0.0	96.0 ± 8.9
grasp	5	0.491 ± 0.007	100.0 ± 0.0	100.0 ± 0.0
pound	5	0.637 ± 0.009	100.0 ± 0.0	100.0 ± 0.0
scoop	5	0.504 ± 0.054	80.0 ± 0.0	80.0 ± 0.0
support	5	0.638 ± 0.017	100.0 ± 0.0	100.0 ± 0.0

Contact, extent, and shape give different comparisons. Astra leads mask IoU on UMD cutting, grasping, scooping, and support, while Sol leads containment and pounding. In the interaction-conditioned sources, one-handed pulling has zero strict box success for every model despite nonzero contact validity. Table A20 identifies the best peer separately for each reported criterion. ReasonAff and RAGNet retain their source-specific scenes and interaction annotations.

Table A20: Metric-specific Affordance peers. Values are mean ± run SD, using the repetition counts in Appendix B.4. Each peer has the highest non-Astra mean for the indicated criterion; comparisons use available runs.

Task / criterion	Astra	Best peer	Model
UMD cutting: mask IoU	0.361±0.009	0.341±0.022	Fable
UMD cutting: strict box (%)	96.0±8.9	100.0±0.0	Fable
UMD containment: mask IoU	0.472±0.067	0.673±0.069	Sol
UMD containment: strict box (%)	44.0±16.7	96.0±8.9	Qwen-Max
ReasonAff pulling: point (%)	60.0±13.7	50.0±0.0	Fable
RAGNet pulling: point (%)	25.0±5.9	31.7±7.0	Qwen-Max
ReasonAff two-hand movement: point (%)	48.0±11.0	26.7±11.5	Fable
RAGNet two-hand movement: point (%)	20.0±27.4	50.0±0.0	Sol / Fable

D.4 Task Planning: Goal Families and Terminal States

Compound goals and source-specific requirements. Table A21 covers all 56 native families across the six sources. Each row compares Astra with the highest-success peer for that family, whereas Figure 7 keeps Fable fixed for source-level and household comparisons. Heating and dual placement test state change and multiple-object goal retention; ScienceWorld distinguishes experimental topics and instrument use. This preserves the task-specific advantages and reversals behind the aggregate Planning lead.

Table A21: Task Planning by native goal family, with one episode per assigned case. Entries are solved/assigned cases. One highest-success peer is shown per row; * denotes a tie with another peer. Full model results are in the accompanying CSV.

Goal family	Astra	Best peer	Model
ALFRED			
Illumination	1/2	1/2	Fable*
Placement	4/4	2/4	Sol*
Container transport	2/2	1/2	Fable*
Cleaning	2/2	1/2	Sol*
Cooling	2/4	2/4	Fable
Heating	2/3	0/3	Sol*
Dual placement	2/3	0/3	Sol*
ALFWorld			
Illumination	3/3	2/3	Fable*
Placement	3/4	2/4	Sol*
Container transport	2/2	1/2	Gemini*
Cleaning	3/4	1/4	Sol*
Cooling	0/1	0/1	Sol*
Heating	1/1	0/1	Sol*
Dual placement	2/5	1/5	Sol*
DiscoveryWorld			
Archaeology dating	1/1	1/1	Sol*
Combinatorial chemistry	1/1	1/1	Sol*
It's not rocket science	1/1	1/1	Fable
Lost in translation	1/1	1/1	Fable
Plant nutrients	1/1	0/1	Sol*
Discovery feed	1/1	1/1	Sol*
Doors	1/1	1/1	Qwen-Max
Navigation in a house	1/1	1/1	Fable*
Pick and give	1/1	1/1	Sol*
Space sick	1/1	1/1	Fable
HumanCLAW			
Navigate to contact	3/5	3/5	Gemini
Sit on target	0/8	0/8	Sol*

Continued from previous page.

Goal family	Astra	Best peer	Model
ScienceWorld			
Find animal	3/3	3/3	Fable
Find living thing	2/2	2/2	Sol*
Find non living thing	3/3	3/3	Sol*
Find plant	3/3	3/3	Sol*
Grow fruit	2/3	2/3	Fable
Inclined plane determine angle	2/2	2/2	Fable
Inclined plane friction named surfaces	15/16	16/16	Fable
Inclined plane friction unnamed surfaces	1/1	1/1	Fable*
Lifespan longest lived	2/2	2/2	Sol*
Lifespan longest lived then shortest lived	1/1	1/1	Sol*
Lifespan shortest lived	1/1	1/1	Sol*
Measure melting point known substance	8/8	6/8	Gemini
Measure melting point unknown substance	1/1	1/1	Sol*
Mendelian genetics known plant	1/1	1/1	Sol*
Mendelian genetics unknown plant	3/5	1/5	Qwen-Max*
Test conductivity	6/7	6/7	Gemini
Test conductivity of unknown substances	6/8	6/8	Fable
Use thermometer	4/6	5/6	Fable*
VirtualHome			
Grab	1/1	1/1	Sol*
Greet	1/1	1/1	Sol*
Lie	1/1	1/1	Sol*
Navigation or observation	1/1	1/1	Sol*
Read	2/2	2/2	Fable*
Rinse	1/1	1/1	Sol*
Sit	3/3	3/3	Sol*
Switchoff	1/1	1/1	Sol*
Switchon	7/7	7/7	Sol*
Touch	1/1	1/1	Sol*
Type	1/1	0/1	Sol*
Wash	0/1	1/1	Sol*

Attempted actions and terminal goals. All eleven Astra household failures exhaust their assigned budget. ScienceWorld's nine failures have scored task-failure outcomes with no recorded parser-no-match, invalid-action, or code-exception events; negative score sentinels are not physical-progress measures. HumanCLAW identifies a different limit (Table A22): six Astra sitting episodes find the target, five approach within 20 cm and issue a sit action, yet none succeeds. Fable records no target-finding success on these episodes. The shared zero terminal rate conceals different progress before the required body–object condition.

Table A22: HumanCLAW sitting diagnostics for eight assigned cases per model. Counts pool independently recorded conditions. † denotes lower bounds because process fields are missing for two Gemini cases; dashes denote unavailable process metrics. Terminal success retains all eight assignments.

Model	Process coverage	Target found	Within 20 cm	Near + sit issued	Success
Astra	8/8	6	5	5	0
Sol	8/8	3	1	1	0
Fable	8/8	0	0	0	0
Qwen-Max	0/8	–	–	–	0
Gemini	6/8	3†	3†	3†	0
Q-397B	0/8	–	–	–	0
Q-27B	0/8	–	–	–	0

D.5 Manipulation: Control and Protocol Dependence

Baseline performance and coverage. Table A23 supports Table 2 and Figure 8a with all thirteen sources and seven agents. The baseline aggregate combines ten-source baseline results with the three sources shared across conditions. Astra records 76 successes out of 182 assignments; Fable records 45. Appendix E.1 defines the fixed denominator and coverage accounting.

Consecutive goals and task-specific evidence. Table A24 reports auxiliary metrics under both evaluation conditions. CALVIN reports average consecutive goals completed. The other rows average task-normalized native criteria or control evidence, with definitions below.

Table A23: Complete Manipulation success by protocol (%). Baseline totals use 182 assignments; – marks missing records. The last three sources share current results. The common nine-source subset isolates sources with records under both protocols.

Benchmark	N	Astra	Sol	Fable	Q-Max	Gemini	Q-397B	Q-27B
Baseline								
CALVIN	10	30.0	0.0	0.0	0.0	0.0	0.0	0.0
CLIPort	18	61.1	27.8	33.3	22.2	27.8	22.2	16.7
RLBench	10	60.0	20.0	10.0	0.0	20.0	0.0	10.0
RoboCasa	10	10.0	0.0	0.0	0.0	0.0	0.0	0.0
RoboCasa365	10	0.0	0.0	10.0	0.0	0.0	0.0	0.0
RoboTwin 2.0	50	38.0	10.0	14.0	4.0	8.0	6.0	8.0
RoboWits	10	10.0	10.0	10.0	10.0	20.0	0.0	10.0
VIMA-Bench	10	40.0	30.0	30.0	10.0	20.0	30.0	30.0
VLABench	25	72.0	52.0	60.0	20.0	36.0	24.0	24.0
ManiSkill	9	55.6	0.0	–	0.0	11.1	0.0	0.0
LIBERO-PRO	10	40.0	30.0	60.0	10.0	40.0	0.0	10.0
robosuite	5	80.0	80.0	100.0	60.0	100.0	40.0	20.0
BEHAVIOR-1K	5	0.0	0.0	0.0	0.0	0.0	0.0	0.0
All assigned	182	41.8	19.8	24.7	9.3	18.7	9.9	11.0
Common nine sources	153	41.2	19.0	22.2	8.5	15.7	10.5	11.8
RGB-expanded								
CALVIN	10	0.0	10.0	40.0	10.0	0.0	0.0	0.0
CLIPort	18	55.6	33.3	50.0	27.8	38.9	22.2	16.7
RLBench	10	40.0	10.0	30.0	10.0	10.0	0.0	0.0
RoboCasa	10	10.0	0.0	10.0	0.0	0.0	0.0	0.0
RoboCasa365	10	0.0	0.0	20.0	0.0	0.0	0.0	0.0
RoboTwin 2.0	50	42.0	12.0	34.0	14.0	6.0	4.0	6.0
RoboWits	10	10.0	10.0	20.0	20.0	10.0	0.0	10.0
VIMA-Bench	10	100.0	90.0	100.0	70.0	60.0	20.0	20.0
VLABench	25	64.0	48.0	68.0	32.0	52.0	24.0	28.0
ManiSkill	9	11.1	0.0	55.6	0.0	0.0	0.0	0.0
LIBERO-PRO	10	40.0	30.0	60.0	10.0	40.0	0.0	10.0
robosuite	5	80.0	80.0	100.0	60.0	100.0	40.0	20.0
BEHAVIOR-1K	5	0.0	0.0	0.0	0.0	0.0	0.0	0.0
All assigned	182	39.6	23.6	44.5	19.2	22.0	8.8	9.9
Common nine sources	153	41.2	23.5	42.5	20.3	20.3	9.2	10.5

Table A24: Manipulation auxiliary metrics under both protocols. CALVIN reports average consecutive goals completed (0–5); other rows average task-normalized criteria or control evidence (%). Definitions and time windows appear in Table A25. – marks unavailable records; the final three rows share current results.

Benchmark	N	Astra	Sol	Fable	Q-Max	Gemini	Q-397B	Q-27B
Baseline								
CALVIN goals (0–5)	10	3.10	1.40	0.90	0.80	1.80	0.00	0.10
RoboCasa	4	–	–	–	–	–	–	–
RoboCasa365	7	–	–	–	–	–	–	–
ManiSkill conditions	5	60.00	0.00	–	0.00	0.00	0.00	0.00

Continued from previous page.

Benchmark	N	Astra	Sol	Fable	Q-Max	Gemini	Q-397B	Q-27B
LIBERO-PRO	10	20.00	20.00	30.00	5.00	20.00	0.00	5.00
robosuite	5	60.00	60.00	80.00	40.00	80.00	30.00	16.67
BEHAVIOR-1K control	5	13.33	16.67	30.00	16.67	6.67	6.67	6.67
RGB-expanded								
CALVIN goals (0–5)	10	1.10	0.50	2.90	1.30	0.70	0.00	0.00
RoboCasa	4	25.00	33.33	83.33	50.00	50.00	16.67	25.00
RoboCasa365	7	33.33	23.81	57.14	28.57	47.62	28.57	23.81
ManiSkill conditions	5	0.00	0.00	60.00	0.00	0.00	0.00	0.00
LIBERO-PRO	10	20.00	20.00	30.00	5.00	20.00	0.00	5.00
robosuite	5	60.00	60.00	80.00	40.00	80.00	30.00	16.67
BEHAVIOR-1K control	5	13.33	16.67	30.00	16.67	6.67	6.67	6.67

Comparisons with matching records. The common nine-source subset excludes ManiSkill and the three shared-current sources, leaving 153 assignments per agent and protocol. Astra has 63 successes in both configurations, but task-specific gains and losses offset; Fable increases from 34 to 65. Appendix E.1 examines the corresponding execution evidence.

Metric definitions and coverage. The 183-case Manipulation inventory contains 182 binary tasks and one continuous-return AnymalC-Spin task. AnymalC-Spin is evaluated separately through continuous return. For each model and configuration, task success is $\mathrm{SR} = 100N^{-1}\sum_i s_i$, where $s_i = 1$ requires native acceptance. Runs without native acceptance receive zero success credit, including timeouts, execution or infrastructure failures, and missing records. Each task contributes one run to the assigned denominator.

CALVIN scores a five-goal sequence as successful only when all five goals are completed in order. Its average number of consecutive goals completed is $\bar{L} = 10^{-1}\sum_{i=1}^{10} L_i$, where $L_i \in \{0, \ldots, 5\}$ is the verified consecutive prefix. Unknown prefixes contribute zero: baseline Qwen-Max has one such record, while expanded Gemini's five available records contribute seven goals over the fixed ten-case denominator, giving 0.70.

For the other auxiliary metrics, case i has K_i scoring items and verified indicators z_{ij}. Its score is $u_i = K_i^{-1}\sum_{j=1}^{K_i} z_{ij}$; the reported mean is $100N^{-1}\sum_i u_i$. Failed or unknown items contribute zero. This weights tasks equally even when K_i differs. The diagnostics summarize recorded control events on fixed task subsets using the scoring items and evidence windows in Table A25.

Table A25: The six auxiliary evidence metrics, averaged over task-normalized items within each fixed subset.

Source	N	Scoring items	Evidence window
RoboCasa	4	Approach, grasp and lift, stable release	Any verified event during execution
RoboCasa365	7	Same three transport milestones	Any verified event during execution
ManiSkill	5	Two or three native conditions per task	Final verification state
LIBERO-PRO	10	Two task-specific conditions	Final state and limited process records
robosuite	5	Two lifting, three stacking, or four wiping items	Native success and terminal reward-derived evidence
BEHAVIOR-1K	5	Six control-evidence items	Before the first episode-end signal

Transport control versus destination success. RoboCasa uses ArrangeBreadBasket, LoadDishwasher, PickPlaceToasterToCounter, and RecycleBottlesByType. RoboCasa365 uses BreadAndCheese, MakeBananaMilkshake, MicrowaveThawingFridge, MixCakeFrosting, PlaceMeatInMarinade, PlaceVegetablesEvenly, and RemoveCuttingBoardItems. Door, appliance, and navigation-only tasks are outside these subsets. Approach requires executed motion and confirmed arrival; grasp-and-lift requires both grasp and lift confirmation. Stable release requires prior grasp-and-lift of the same canonically named object, followed by confirmed release and stability. Each item scores at most once per case. These events need not cover every task object, and release does not establish the requested destination relation.

Table A26: ManiSkill's five-task diagnostic. Items are evaluated at final verification; grasp history is separate. The other four binary tasks remain in complete-task success.

Task	K_i	Scored conditions
PickCube	2	Object placed at its goal; placed and robot static. The suspended goal requires no release.
PlaceSphere	3	Sphere on its support; on support and released; both with the sphere static.
StackCube	3	Cube A on B; on B and released; both with cube A static.
PlugCharger	2	Goal distance ≤ 0.005 m; distance satisfied and angular error ≤ 0.2 rad.
PegInsertionSide	2	Peg-head $x \geq -0.015$ m in the hole frame; official insertion success.

LiftPegUpright, PullCube, PushCube, and RollBall contribute to the nine-task success rate but not this diagnostic. Fable's baseline records are unavailable for both sets and displayed as dashes.

Task-specific conditions in LIBERO-PRO and robosuite. LIBERO-PRO uses fixed init-0 instances: spatial 1/7, object 2/5, goal 1/7, libero_10 1/9, spatial_swap 0, and object_swap 4. Ordinary placement scores recorded grasp and the official target relation; stove opening scores operation of the correct control and an open stove. The two longer tasks score a cup inside a microwave followed by the door being closed, or one food item followed by both items in a basket. Missing grasp/control records score zero; for the longer tasks, necessary preceding relations are confirmed only on official success.

robosuite uses cube lifting with seeds 99/50, cube stacking and restacking with seed 14, and spill wiping with seed 48. Lifting scores recorded grasp and native success; grasp history is missing. Stacking uses the fixed local reward q: $q > 0.125$ supports a grasp/lift/stack condition, $q \geq 0.5$ supports the top cube being lifted, and native success supplies the third item. A restacking zero can mask intermediate events under the two-cube airborne rule; these items remain unknown and receive zero credit. Wiping checks 25%, 50%, 75%, and complete removal. Current records verify the first three only on native complete success, rather than interpreting total reward as a wiped fraction. Consequently, Fable's 100% complete-task success and 80% auxiliary score are compatible: missing intermediate evidence lowers the latter.

Table A27: BEHAVIOR-1K control evidence. Each item records one type of control event and scores once per case; terminal BDDL goal success is evaluated separately.

Item	Positive evidence
Localization	Finite coordinates or pose from public RGB-D localization or grasp sampling.
Candidate	A nonempty pregrasp or grasp candidate set.
Arrival	Measured arrival from approach or recovery feedback for the same arm.
Holding	An explicit in-hand check or holding confirmation inside checked placement.
Held motion	Associated measured motion of at least 5 cm while holding, or movement with a passed holding check inside checked placement.
Release	Holding, active gripper opening, then no longer holding; or confirmed release inside checked placement.

The five tasks are outfit_a_basic_toolbox, can_meat, tidying_bedroom, putting_shoes_on_rack, and chop_an_onion. Records are linked by arm, name, and temporal order; physical-instance IDs are not consistently recorded. Object identity and inside/ontop relations are checked separately by the terminal goal evaluator. Held motion uses a 5 cm threshold. Control-event records cover four tasks for each of four models and the toolbox task for each of the other three. All retain the five-task denominator. Scoring uses the evaluated episodes up to their first end signal and requires direct evidence for each item. Fable's nine verified items yield $100 \times 9/(5 \times 6) = 30\%$, versus Astra's four items and 13.33%. RoboWits is evaluated through complete-task success.

E Execution and Multi-Round Review Comparisons

E.1 Manipulation under the RGB-Expanded Protocol

Protocol and denominators. Baseline uses Table 2's source-specific observations and native tools. RGB-expanded adds current RGB at each decision. Both conditions use matched inference budgets, interaction allowances, and native execution-step limits. The comparison covers ten suites and 162 tasks: CALVIN, CLIPort, RLBench, RoboCasa, RoboCasa365, RoboTwin 2.0, RoboWits, VIMA, VLABench, and ManiSkill. Table 3b displays three suites; Table A23 gives all results. Common excludes ManiSkill because Fable's baseline records are missing, retaining nine suites and 153 tasks. LIBERO-PRO, robosuite, and BEHAVIOR-1K reuse 20 current records in both CSVs, bringing the domain total to 182. Missing baseline Fable ManiSkill and RoboCasa/365 stage records appear as dashes; the domain success aggregate credits the missing ManiSkill records with zero successes.

Observation and interaction settings. Baseline and RGB-expanded share the per-source inference and interaction budgets in Table A28, together with the same native execution-step limits. Baseline uses each source's native observations and tools; RGB-expanded provides the current scene's RGB views at every model decision. VIMA also supplies task reference images. Text interaction history and public state/perception tools remain available. Both conditions use the same agent loop: one code response may execute several actions before the next model decision.

Request and token accounting. The audited 7,378 expanded Astra/Sol/Fable requests across ten ordinary groups use temperature zero, with neither explicit reasoning effort nor a separate thinking budget. Of these, 7,372 request 3,600 response tokens; six use smaller limits as the case budget runs low. The configurations use a 3,600-token response ceiling and a 12-minute request timeout. Reported reasoning usage is recorded separately and counted once in the case total, avoiding duplication when included in completion usage; absent usage details remain unknown. Verifier allowance is one more than the model-round limit.

Table A28: Shared execution ceilings for Baseline and RGB-expanded. Times are minutes; token budgets are thousands. A round is a model decision, Tools counts interface calls, and Checks counts verifier calls. The limits apply to all seven agents within each source.

Source	Rounds	Tools	Checks	Tokens	Trial	Code	Native
CALVIN	40	7,000	41	1,280	120	10	10
CLIPort	24	256	25	960	60	10	10
RLBench	16	256	17	640	60	10	10
RoboCasa	24	256	25	1,120	60	6	10
RoboCasa365	24	256	25	1,120	60	8	10
RoboTwin 2.0	16	256	17	640	60	10	20
RoboWits	24	256	25	640	60	30	20
VIMA-Bench	16	256	17	640	60	10	10
VLABench	16	256	17	640	60	10	10
ManiSkill	32	1,024	33	640	60	10	10
LIBERO-PRO	24	256	25	640	180	60	60
robosuite	24	256	25	640	180	60	60
BEHAVIOR-1K	24	256	25	640	180	60	60

Tool calls, action calls, and physical steps are distinct counters. Trial time includes model waiting, code execution, and feedback within the loop; Code and Native limit one code execution and one native call, respectively. Request timeouts do not add extra time beyond Trial. Initialization, outer watchdogs, and retry backoff have separate clocks, so Trial is not identical to total wall-clock duration.

Table A29: Expanded camera presets, width × height in pixels. The preset inventory covers all 183 assigned cases, including AnymalC-Spin. Multiple views enter as separate images.

Source	Views	Resolution
CALVIN	Static; gripper	200×200; 84×84
CLIPort	Front, left, right; no Oracle camera	640×480
RLBench	Front, wrist, overhead	128×128
RoboCasa/365	Left/right agent view, eye-in-hand	512×512
RoboTwin 2.0	Head, left, right	320×240
RoboWits	Ego, left/right wrist	848×480
VIMA-Bench	Front/top scene views and task reference images	256×128
VLABench	Right, left, forward, Franka wrist	480×480
ManiSkill	Base; additional hand view for PegInsertionSide, PlugCharger, StackCube	128×128
LIBERO-PRO	Agent view, wrist	800×512
robosuite	Agent view, wrist	512×512
BEHAVIOR-1K	Ego, left/right wrist	512×512

VIMA's reference-image count varies by task. The remaining seven ManiSkill inventory cases, including AnymalC-Spin, use one base view. RoboCasa/365 views are rendered scene observations. Request images and process videos are separate records: continuous video is not enabled uniformly for VLABench, LIBERO-PRO, or robosuite, while the later BEHAVIOR revision queue records at 15 fps. Video frame rate does not define model observation frequency.

Table A30: Native execution-step limits shared by both conditions. Values show environment defaults → effective evaluation limits; control frequency is unchanged.

Source	Base → effective	Counter and termination semantics
CLIPort	$5 \rightarrow 10$	Pick-and-place actions
RoboCasa/365	$1000 \rightarrow 2000$	Control steps; field does not truncate episodes in the current wrapper
RoboTwin 2.0	$400 \rightarrow 800$	Native action calls, not all simulator steps
RoboWits	$800 \rightarrow 1600$	Control steps for the checked case
ManiSkill	$50 \rightarrow 100$	Control steps; PickCube termination/boundary checks verify the effective limit
LIBERO-PRO	$4000 \rightarrow 8000$	Wrapper/control counter; field does not truncate episodes
robosuite	$1500 \rightarrow 3000$	Wrapper/control counter; field does not truncate episodes
BEHAVIOR-1K	$500 \rightarrow 1000$	Environment control steps; later collision revision retains 1000
CALVIN, RLBench, VIMA	No finite limit	Checked instances remain bounded by rounds, calls, tokens, and time
VLABench	Unlimited	Simulation-time limit remains unlimited

CALVIN: interaction versus feedback control. The execution analysis reports Astra's average consecutive goals decreasing from 3.10 to 1.10, Sol's from 1.40 to 0.50, and Fable's increasing from 0.90 to 2.90. Astra's median action-interface calls fall from 1,795 to 480.5, and Sol's from 1,247.5 to 734. Their median cumulative prompt tokens rise from approximately 321k and 363k to 1,206k and 1,217k. These totals accumulate prompt tokens over successive requests. Under RGB-expanded, five Astra and eight Sol episodes exhaust the shared per-case token allowance. Sol nevertheless increases full-sequence success from zero to one, illustrating why mean progress and completion need separate reporting.

Figure A3a connects these aggregate changes to controller construction and final goal completion in sequence 0557.

Budget utilization and interface effects. Three of Fable's four RGB-expanded CALVIN successes finish at rounds 22, 24, and 25, within the shared execution allowance. These episodes use later rounds to continue control after establishing feedback. Figure A3b examines controller construction in PickCube. Tool semantics also matter: the RGB tool compares colors on a 0–255 scale, whereas some Astra/Sol calls use 0–1 values and tight tolerances, producing empty detections. Fable ob-

tains the red region with [255,0,0] and tolerance 80. Correct units determine whether the controller receives a usable target.

Input and execution audit. The audit covers 7,378 Astra/Sol/Fable requests from the ten ordinary benchmark groups: all contain RGB, archived Astra/Sol responses have no length termination, and Fable has 396. RLBench still contains four Astra, two Sol, and four Fable code-execution timeouts. Astra/Sol failures therefore occur even with delivered images and complete responses; code timeouts and interface semantics provide additional execution-level explanations. VIMA improves for all three agents, and Astra improves on RoboTwin while remaining unchanged over the common 153 tasks. The gains vary with the task and how each model uses the additional visual observations.

E.2 Fixed-Input Multi-Round Review

Multi-round review lets agents revisit their initial answers using the same task observations and scoring criteria. Self-review adds reconsideration, Crop review adds image crops, and Tool-assisted review adds specialist visual and geometric measurements. The three conditions permit up to two, two, and six additional responses, respectively, across Geometry, Spatial Reasoning, and Affordance. Table 3a gives the results; the sampling and tools below define its scope.

Shared cases. Astra and Qwen-27B share 46 cases: 19 Geometry, 18 Spatial Reasoning, and nine Affordance. A 72-case pilot covers 24 benchmark/task strata selected by task category and SHA-256 ordering with seed 20260923, independently of scores and answers. It excludes the 60 exact cases from an earlier performance-selected study. One case per stratum forms a 24-case development set; withholding two more cases that share development scenes leaves the 46 evaluation cases. The diagnostic includes recurring scenes and cases inspected during development.

Table A31: Review conditions for the same cases and native scoring interfaces. Budgets are additional to native execution; response limits count output tokens.

Condition	Evidence available during review	Additional response budget
Initial	Original task observations	None
Self-review	Original observations and initial answer	Two responses, up to 1,800 tokens each
Crop review	Self-review evidence plus image crops	Two responses, up to 1,800 tokens each
Tool-assisted	Original RGB, crops, and specialist measurements	Six responses, up to 4,096 tokens each; up to 16 specialist requests

Evidence construction. Tool-assisted review adapts GCA's measurement and reference-frame formalization (Chen et al., 2025b) and SpatialClaw's inspectable execution (Cho et al., 2026). It formulates the measurement before consulting the candidate answer, then uses a persistent Python session. SAM 3 (Carion et al., 2025) supplies regions; SuperPoint (DeTone et al., 2018) and LightGlue (Lindenberger et al., 2023) supply correspondences; GeoCalib (Veicht et al., 2024) estimates calibration. Numerical routines compute relations and estimates. Finite-value and coordinate checks validate the computed outputs. Remaining semantic errors include selecting a whole-object mask for a functional part and ambiguous translation scale.

Observation and execution controls. Specialist tools receive only previously exposed RGB, without predicted depth, learned 3D reconstruction, private calibration, scoring annotations, or extra video frames. Public depth or masks originally supplied by a task remain available to the base model. Review reuses native candidates where available; one Astra Spatial case starts from its public observation because the candidate is missing. Temperature is zero, and total-token and wall-time accounting includes native execution. Incomplete runs may be retried; the first completed result is retained regardless of correctness, yielding one outcome per condition and case.

Outcome interpretation. Geometry counts improved, tied, and worsened results across 24 metric entries from 19 cases; one Qwen-27B crop entry is unavailable. Spatial accuracy retains 18 assigned cases, with missing answers counted as incorrect. Affordance reports box–box IoU and success at IoU≥ 0.5 over nine cases.

F CASE STUDIES

From spatial judgments to usable contacts. Figure A2 illustrates Section 4.3 with six cases from the first evaluation run.

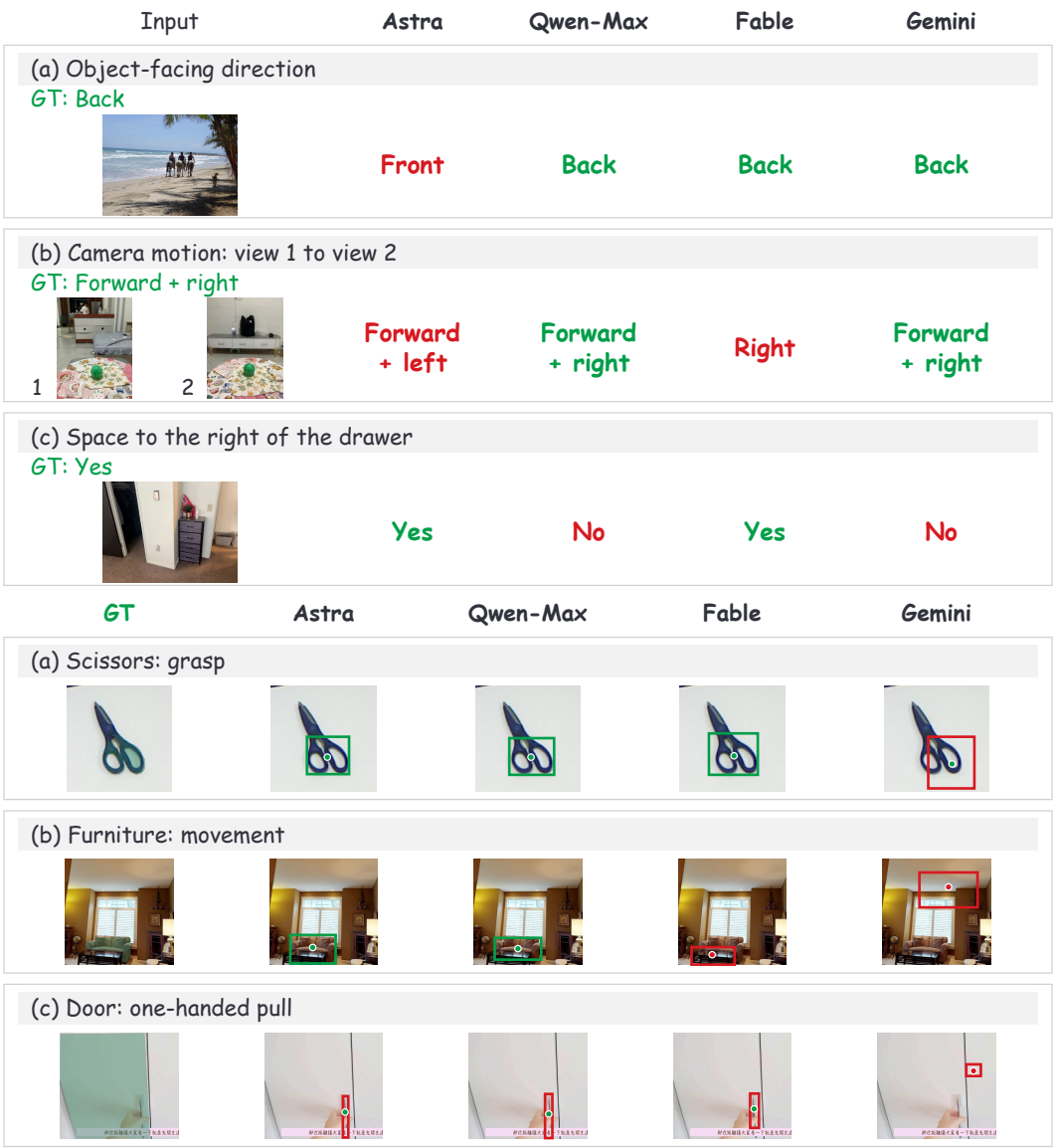

Figure A2: Spatial judgments from the first evaluation run (top: 3DSRBench, MindCube, RoboSpatial) and affordance localization (bottom: UMD, ReasonAff, RAGNet). Top questions concern the dog's facing direction, camera motion, and whether the plant fits beside the drawer; green answers match GT. Below, green points lie inside the GT mask and green boxes reach box IoU$\geq$ 0.5; red marks fail their respective criteria.

① *Spatial references.* In these runs, Astra correctly judges placement compatibility but misidentifies object-facing direction and camera translation. ② *Contact versus coverage.* Astra identifies usable tool and furniture regions. For pulling, three models place valid contacts on the handle, but their boxes cover only a fraction of the reference door region.

Controller construction across observation settings. Figure A3 links controller construction to terminal outcomes in CALVIN sequence 0557 and ManiSkill PickCube; Appendix E.1 gives the execution settings.

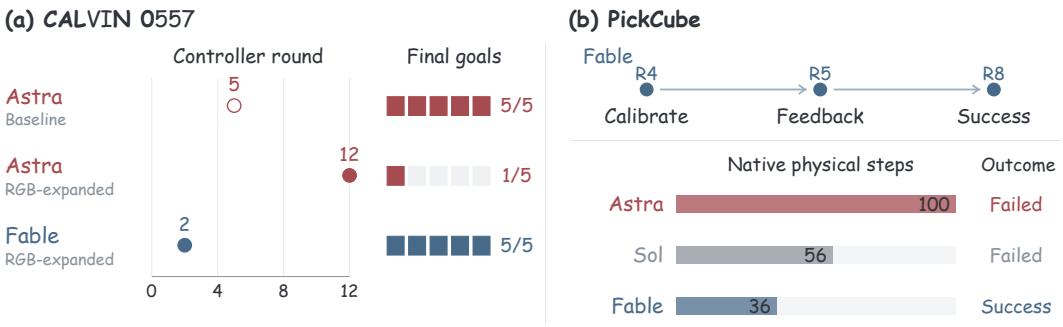

Figure A3: Controller construction and task outcomes. (a) CALVIN dots mark controller construction rounds; filled tiles count final consecutive goals. (b) PickCube shows Fable's ordered calibration, feedback-control, and success events, plus all three agents' physical-step totals and outcomes under RGB-expanded execution. R denotes model rounds; physical steps count environment updates.

CALVIN: Astra's feedback controller emerges later. Baseline Astra builds control from tool-center-point error at round 5; RGB-expanded Astra builds it at round 12. Fable constructs a state-reading movement loop at round 2. Earlier controller construction coincides with longer completed goal sequences in these cases.

PickCube: fewer steps, better organized control. Fable calibrates joint response before constructing reusable numerical feedback control, then succeeds in 36 native steps. Astra and Sol use 100 and 56 steps across 25 and 32 responses without success; their trajectories emphasize direct joint increments and cross-round queries.

Motion and task completion. Figure A4 contrasts visible motion with verified task completion.

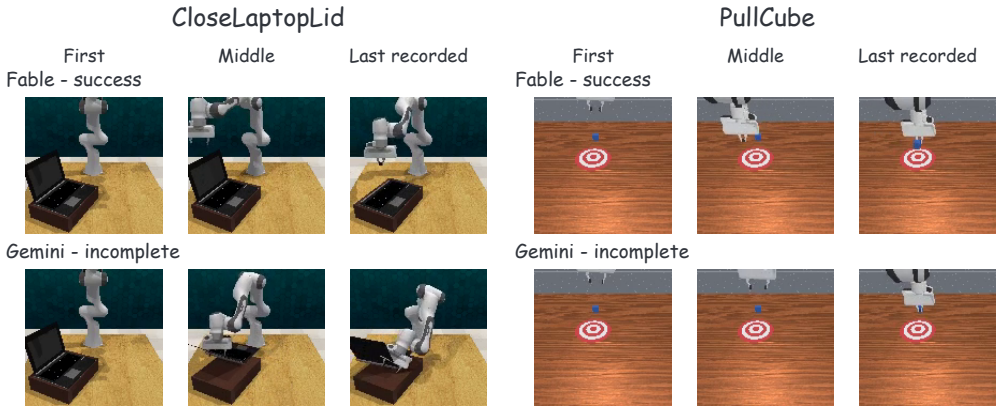

Figure A4: Archived execution contrasts under matched task settings and budgets. CloseLaptopLid uses RLBench's front camera; PullCube uses ManiSkill's base camera. Native records determine success. These earlier episodes are separate from the protocol comparison.