WHAT A ROBOT HARNESS ASKS ITS VLM: A DEPLOYMENT-GROUNDED BENCHMARK

Anonymous authors
Paper under double-blind review

ABSTRACT

The release of GPT-6 Astra has intensified interest in embodied systems built around vision–language models (VLMs) and robot-specific harnesses. In such systems, however, VLMs are still commonly evaluated as standalone models rather than through the interfaces at which they operate. Even existing benchmarks grounded in real-world embodied settings typically measure either end-to-end system performance or isolated capabilities such as spatial reasoning, abstracting away the interface through which a robot software stack queries its VLM. Their results therefore provide limited guidance for model selection and diagnosis in deployed embodied systems. We introduce a deployment-grounded benchmark whose task taxonomy and evaluation instances both derive from actual calls logged in successful episodes of a binocular-camera manipulation harness, rather than tasks constructed for evaluation. The benchmark covers eight request families and preserves the task-relevant processed views and output contracts used in deployment. It contains semantically aligned Chinese–English question pairs balanced across request families. We evaluate representative VLMs under each item's output contract, keeping acceptance separate from whether the decision itself is correct. Across four matched specialist–base pairs, no specialist has a higher overall request-score point estimate than its base. Embodied-R1.5-8B provides the clearest reversal, trailing Qwen3-VL-8B by 4.96 points despite leading it on three published embodied dimensions.

1 INTRODUCTION

The release of GPT-6 Astra has drawn attention to a modular way of building embodied systems (OpenAI, 2026; Su et al., 2026). In this design, a software harness selects what the vision–language model (VLM) sees, asks it for a decision, and passes its answer to another robot module (Liu et al., 2024; Chen et al., 2026). The harness may ask whether a bottle has a cap before requesting its location, or which objects still need to be placed in a bin. For the robot, the useful answer is one that gets this specific decision right and gives the next module what it needs.

Existing evaluations look at related questions. Vision–language–action evaluations measure how well an action-producing system performs robot tasks (Brohan et al., 2023; Kim et al., 2024). Robot-grounded question-answering and spatial evaluations test what a VLM understands about scenes or trajectories (Sermanet et al., 2023; Chen et al., 2025; Xie et al., 2026; Chen et al., 2024). But a running harness may send a cropped view instead of a full frame. It may ask for a location only if the object exists. It may need a complete list of unfinished goals. A model can score well on spatial questions and still answer these requests incorrectly. Those scores alone do not tell us how it will respond to the calls the robot actually makes.

We evaluate robot-issued VLM requests as the unit of analysis (Figure 1). Each benchmark item retains the visual payload and decision from a production call and adds a reference answer and an explicit answer contract. We replay the same requests across models and measure both contract acceptance and reference-relative correctness.

Our initial public release contains 1,200 requests, one from each of 1,200 distinct successful episodes of a binocular-camera harness operating in tabletop-manipulation and washing-machine scenes. Each request has aligned Chinese and English prompts that share its image and reference. The release covers eight request families, with 150 requests per family. All model evaluations on

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

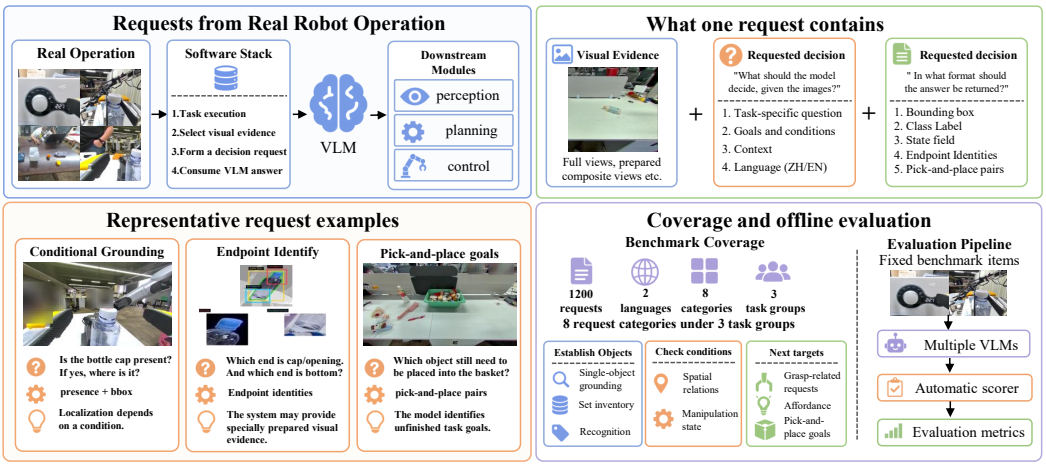


Figure 1: Overview of the deployment-grounded benchmark. Top left, the robot harness selects visual evidence, queries a VLM, and passes its answer to downstream perception, planning, or control modules. Top right, each request specifies the visual evidence, the decision to make, and the required answer format. Bottom left, examples illustrate conditional grounding, endpoint identification from prepared views, and the identification of unfinished pick-and-place goals. Bottom right, the benchmark contains 1,200 requests with paired Chinese and English prompts, spanning eight request families and three workflow roles. Offline evaluation applies a shared automatic scorer to responses from multiple VLMs.

our benchmark in this paper use these same 1,200 requests. We plan to release further requests in stages after company confidentiality review and de-identification, with the aim of expanding the benchmark beyond 10,000 items.

Across all four matched pairs, specialists have lower overall request-score point estimates than their corresponding bases. Especially, Embodied-R1.5-8B provides the clearest reversal: it leads Qwen3-VL-8B on three published embodied dimensions (Yuan et al., 2026), yet trails the same base by 4.96 points on our requests. The other pairs show different strengths and deficits across workflow jobs.

We make the robot–VLM boundary measurable through a benchmark of robot-issued calls, matched specialist–base comparisons, and request-level diagnostics. The benchmark retains each call's visual evidence, decision, and answer requirements. The matched comparisons test whether published specialist advantages carry over to these calls. Analyses by workflow family and contract acceptance identify the requests on which models struggle.

2 Related Work

VLMs in modular robot systems. Vision–language models enter embodied systems through several routes. Vision–language–action models such as RT-2 and OpenVLA map visual observations and language instructions toward robot actions (Brohan et al., 2023; Kim et al., 2024). Other systems ground a general-purpose multimodal model in embodied observations and feedback: PaLM-E interleaves visual, state, and language inputs, Inner Monologue uses environment feedback for language-model planning, and RoboFlamingo adapts an open VLM to manipulation trajectories (Driess et al., 2023; Huang et al., 2022; Li et al., 2023). Modular systems also connect a general-purpose model to perception, skills, and controllers. SayCan grounds language-model choices in available robot affordances, while OK-Robot demonstrates the integration of open models with a mobile-manipulation stack (Ahn et al., 2022; Liu et al., 2024). Recent work such as Show-Harness makes the modular design explicit by exposing semantic action units and feedback to a VLM while leaving embodiment-specific grounding to a harness (Chen et al., 2026). These systems establish the robot–VLM boundary as a meaningful system object: the harness selects visual evidence, asks a concrete question, and passes the answer to another component. If VLMs are evaluated at this boundary, the request made by the running system becomes part of what should be measured.

Benchmarks for embodied and robot-grounded reasoning. Existing benchmarks cover several useful layers of embodied competence. OpenEQA asks open-vocabulary questions about physical environments (Majumdar et al., 2024); EmbodiedBench and RoboBench cover broad instruction, perception, spatial reasoning, planning, affordance, and failure-analysis tasks (Yang et al., 2025; Luo et al., 2025); and VIMA studies general manipulation under multimodal prompts (Jiang et al., 2022). RoboVQA and Robo2VLM derive large-scale questions from robot or human–robot trajectories, while RoboVista collects questions from decision points across multiple robot applications (Sermanet et al., 2023; Chen et al., 2025; Xie et al., 2026). ManipBench probes low-level manipulation reasoning and reports a relationship with physical outcomes (Zhao et al., 2025), RoboProcessBench targets process-aware state and temporal understanding from manipulation traces (Xia et al., 2026), and HRIBench studies VLM perception and latency for human–robot interaction (Shi et al., 2025). SpatialVLM is a representative effort to improve VLM spatial reasoning through large-scale spatial supervision (Chen et al., 2024). These works show the value of grounding evaluation in embodied scenes, traces, spatial supervision, and subproblems. They do not all target the same object: an image may come from a robot while the question is written after the fact, or a broad capability task may abstract away the processed view and answer form supplied by a deployed stack. Our distinction is therefore the provenance of the request itself. We derive both the task families and candidate instances from calls that a running harness had already issued to its VLM.

Structured outputs and related evaluations. The boundary between a foundation model and a robot controller has also motivated evaluation of input–output contracts. The Embodied Agent Interface benchmark standardizes input–output specifications for language-model modules such as goal interpretation, subgoal decomposition, action sequencing, and transition modeling, and reports error types beyond final task success (Li et al., 2024). Code-as-Monitor couples visual constraints to model-generated monitoring logic (Zhou et al., 2025), while SO-Bench treats structural validity as a first-class dimension of multimodal evaluation (Feng et al., 2025). These works motivate separating whether an answer can be consumed from whether it is correct. Our evaluation object is narrower and more operational: it is not an abstract API defined for benchmarking, but a set of request contracts already used by one deployed manipulation stack. A conditional box, a state label over a processed view, and a remaining-goal list are different jobs because downstream code consumes them differently.

Positioning of this work. Our benchmark evaluates requests issued by a running manipulation harness. It retains the visual payload, decision, and answer requirements at the VLM interface, so the same requests can be replayed across models. This lets us compare matched checkpoints and locate differences by request family and answer acceptance.

3 Deployment Requests as the Evaluation Unit

The robot–VLM boundary becomes concrete when the harness calls its model. At that point, the stack has selected the visual evidence and determined what decision it needs to continue. It also specifies the form of a reply that another module can use. We use this call as the starting point for an evaluation item: the transmitted view and requested decision are retained, while a reference answer and scoring contract make the item replayable across models (Figure 1).

The three production calls in Figure 1 show why these choices matter. In conditional grounding, the harness first asks whether a cap exists. If it does, the model returns its box; otherwise the production request requires `<FAILED>`. For endpoint identification, the harness sends an isolated bottle with both ends enlarged, so the prepared image defines what evidence the model receives. For remaining goals, the planner needs the set of unfinished pick-and-place goals, not a single plausible object.

These details determine what a replayed request tests. Replacing the endpoint composite with a full frame changes the input; reducing the goal list to one object changes the requested output. In our comparisons, every model receives the same released image and prompt. We score their responses against the same reference and answer contract, keeping the tested decision fixed across models.

Table 1 groups the eight observed request families by their role in the workflow: establishing objects, checking conditions, and determining next targets. These families describe jobs the harness delegates to the VLM. For example, affordance and grasp-related calls can both involve locations,

Table 1: Eight request families observed in the harness, grouped by workflow role. Each original sampling stratum contains 150 requests. The families describe jobs in the system rather than independent abilities.

Family	Representative decision	Answer form
Establish objects		
Single-object grounding	Is the cap present? If so, where is it?	Box / no target
Set inventory	Which requested objects are visible, absent, or unknown, and where are the visible ones?	Indexed object list
Recognition / attributes	What object, shape, panel state, or visible property is shown?	Class / attributes
Check conditions		
Spatial relations	Does object A have the specified relation to B?	Relation result
Manipulation state	Is the bottle upright, or is the expected item held?	State label
Determine next targets		
Grasp-related requests	Where should the robot grasp the named target?	Target and localization
Affordance	Is the requested interaction feasible, and where?	Target / unavailable
Pick-and-place goals	Which placement goals remain?	Goal list / done

but one asks for a feasible interaction point and the other for a grasp declaration or localization. The families define the original sampling strata used in Section 4; Section 5 specifies how item-level answers are scored.

4 Benchmark Construction

4.1 Sampling Robot-Issued Requests

We build the benchmark from calls logged by a binocular-camera harness in tabletop-manipulation and washing-machine scenes. We first remove duplicate captures and calls with ambiguous request types. We then sample 150 calls from each of eight original sampling strata, favoring variety in targets, views, and answer forms. Each selected call comes from a successful episode at a turn whose resulting action the stack accepted, and no episode contributes more than one call. This produces 1,200 requests from 1,200 distinct episodes. Appendix A.1 gives the full selection procedure.

Each selected request retains the visual payload sent to the VLM. Across the release, 915 payloads are full-frame views, 237 are processed crops, and 48 are processed composites. We retain the production prompt or lightly clarify its required output fields and coordinate convention, while keeping the target, precondition, and decision unchanged. We provide each request in aligned Chinese and English forms with the same image and reference answer. The 1,200 requests therefore yield 2,400 prompts. Some answer contracts require exact target names, so those names remain unchanged in both language forms.

4.2 References Grounded in the Released View

Each sampled request comes from a successful episode and has an associated reply from the VLM deployed in the harness. The released image, prompt, and output contract can admit multiple valid replies. Using the deployed model's reply as the reference would bias the comparison toward that model's choices. We therefore annotate references from the released image and prompt, including permitted alternatives and required failure branches.

Two proposal models answer each released image–prompt pair independently, without the historical reply or episode context. A request goes to human adjudication when the proposals disagree, agree only in part, or have no agreement score (743 requests). The reviewer may consult the episode to resolve ambiguity, but a retained reference must be supported by the released view; otherwise the item is revised or removed.

We then check the references in two ways. For 200 sampled requests, a second annotator reviews the image, prompt, contract, and recorded reference, and accepts 98% before the two annotators resolve

Table 2: Representative answer formats and scoring rules across request families.

Request family	Accepted answer	Scoring rule
Single-object grounding	Legal box. Conditional: Boolean existence + legal box or null.	**IoU** ≥ 0.5. Conditional: correct existence decision; localize present targets.
Set inventory	All reference indices, permitted statuses, and status-compatible boxes.	**Mean per-index correctness** of labels, statuses, and visible-object localization (IoU ≥ 0.5).
Spatial relations	Condition fields or the requested combined tag.	**Exact match**, averaged over required fields; exact match for a combined tag.
Manipulation state	Required state fields or both endpoint labels.	**Exact match** on all required fields.
Grasp-related requests	Box, declaration, and optional point, as required.	Contract-specific declaration check + **IoU** ≥ 0.5; joint point checks where required.
Affordance	Legal point or a finished response.	**Distance** ≤ 60 to one reference point, or correct finished branch.
Pick-and-place goals	List of pick/place label pairs; empty list allowed.	**Set F1** after prescribed label normalization.
Recognition / attributes	String, state, fields, or box.	Contract-specific normalized text, label, field, or localization score.

Note: IoU thresholds apply to the box representation accepted by the item's contract. Point distances are measured in normalized 0–1000 coordinates.

disagreements together. Separately, 200 selected references are executed on the physical robot, and all succeed. Appendix A.2 gives the sampling and agreement criteria.

5 Evaluation Protocol

5.1 Answer Contracts and Scoring

The construction above fixes the image, prompt, and reference for each request. Comparing models on those requests still requires one shared procedure for reading a response and scoring it against that reference. The evaluator applies that procedure to every response: it extracts an answer, checks the answer against the item's contract, and assigns a reference-relative score. Extraction and key normalization are deterministic and shared across models.

The check decides acceptance. It requires the fields, types, and geometric form specified by the contract; some contracts also require reference structure, such as the full index set for inventory. Acceptance is not correctness: an accepted response can still score zero. It is also distinct from compliance with the production parser in Figure 1.

The score then depends on the requested decision (Table 2). Grounding uses box intersection over union (IoU), inventory averages correctness over reference indices, affordance tests point proximity, and pick-and-place uses set F1 over normalized task-label pairs; the remaining contracts check labels, fields, or localization. These scores describe different response properties, so a family score is not a common unit of visual or reasoning ability. Appendix A.3 gives thresholds, variants, and extraction rules.

5.2 Request Score and Acceptance

Each response has two outcomes: whether the evaluator accepts it, and how correct the accepted content is relative to the reference. We summarize both across requests. For request i and language $\ell \in \{zh, en\}$, let $a_{i\ell} \in \{0, 1\}$ indicate acceptance and $s_{i\ell} \in [0, 1]$ the score, with a rejected response scored zero. For $N = 1{,}200$ requests, we report three quantities on a 0–100 scale: the mean request

score S, the acceptance rate V, and the mean score on accepted responses $S_{|V}$,

$$S = \frac{100}{2N} \sum_{i,\ell} s_{i\ell}, \qquad V = \frac{100}{2N} \sum_{i,\ell} a_{i\ell}, \qquad S_{|V} = 100 \frac{\sum_{i,\ell} s_{i\ell}}{\sum_{i,\ell} a_{i\ell}}. \tag{1}$$

S is the primary outcome. Because the eight strata have equal size and the two languages have equal weight, S is also the mean of the eight stratum scores.[1] Rejected responses contribute zero, so $S = VS_{|V}/100$. $S_{|V}$ uses each model's own accepted responses, so it does not compare models on a common set. Neither V nor $S_{|V}$ separates formatting failures from errors of perception or reasoning.

5.3 Models, Matched Pairs, and Uncertainty

The outcomes above are computed for 15 VLMs on the same requests. The comparison centers on four embodied specialists and the bases released with them: Embodied-R1.5-8B (Qwen3-VL-8B-Instruct), RynnBrain1.1-9B (Qwen3.5-9B), UnifoLM-ER-1-4B (Qwen3-VL-4B-Instruct), and MiMo-Embodied-7B (MiMo-VL-7B-SFT). Four additional open Qwen models and InternVL3.5-8B broaden the general-purpose baselines. We include Qwen3.7-plus and GPT-6-astra as closed-model reference points in the overall comparison. Family-level and diagnostic analyses focus on the 13 open-weight models. Appendix B.1 lists lineages and public comparison sources.

Every model receives the same 2,400 prompts, with greedy or temperature-zero decoding and a 1,536-token output limit. Thinking is disabled where supported, and GPT-6-astra runs at reasoning effort low. Appendix B.2 specifies the remaining inference controls.

Uncertainty for a matched pair is a paired bootstrap over the 1,200 episodes, with 10,000 draws. Each draw keeps an episode's two language forms together and keeps both models' responses on those prompts together. Overall comparisons preserve the eight strata; family comparisons resample within a stratum. These intervals condition on the benchmark and exclude reference-annotation uncertainty. Family comparisons are exploratory and unadjusted for multiplicity.

6 Results on Robot-Issued Requests

We compare models on the request score S using the shared extraction and scoring procedure in Section 5. We first report overall and family-level scores, then test whether the specialist–base orderings reported on public benchmarks persist on these requests. We also examine sensitivity to scoring choices and prompt language.

6.1 Performance Across Models and Workflow Jobs

On these requests, S spans a wide range (Table 3). GPT-6-astra reaches 73.86, followed by Qwen3.7-plus (71.21) and Qwen3.6-27B (69.48), while the two MiMo checkpoints score below 24. InternVL3.5-8B scores 53.52, adding a general-purpose baseline outside the Qwen series. Acceptance changes which model looks strongest. Qwen3.5-397B-A17B has the highest score on its accepted responses ($S_{|V} = 82.44$), but a smaller proportion of its responses is accepted by the evaluator than for Qwen3.7-plus, yielding a lower overall S. $S_{|V}$ remains a score on each model's own accepted subset.

Family scores show jobs that this aggregate ranking hides (Figure 2). Among the 13 open-weight models, the best pick-and-place score is 24.8 under exact-label set F1, the best inventory score is 61.0, and the best grasp score is 45.0. Single-object grounding, manipulation state, and affordance exceed 90 for several models. Qwen3.6-27B shows the same spread within one model: 96.7 on single-object grounding, 61.0 on inventory, and 24.8 on pick-and-place. The lower scores fall on contracts that require complete sets, declarations, or goal lists for a downstream module. The contrast marks harder request types; it does not separate perception, decision, and answer-representation errors. Appendix C.1 tabulates the values alongside a constant-reference baseline.

[1] Family results use the original sampling strata. A later metadata-only relabeling of seven items (Appendix A.1) changes no image, prompt, reference, or score.

Table 3: Overall results for all 15 models on 2,400 prompts each, ordered by S. S is the request score, V is scorer acceptance, and $S_{|V}$ is the score on each model's own accepted subset, all on a 0–100 scale. GPT-6-astra runs at reasoning effort low. Family-level and diagnostic tables cover the 13 open-weight models.

Model	V	$\boldsymbol{S}$	$S_{\|V}$
GPT-6-astra	93.88	73.86	78.67
Qwen3.7-plus	90.50	71.21	78.69
Qwen3.6-27B	86.83	69.48	80.01
Qwen3-VL-8B-Instruct	88.46	67.73	76.56
Qwen3.5-397B-A17B	80.42	66.30	82.44
Qwen3-VL-4B-Instruct	89.12	64.54	72.42
Qwen3.5-9B	78.79	63.37	80.43
Embodied-R1.5-8B	85.29	62.77	73.59
RynnBrain1.1-9B	75.33	59.85	79.45
Qwen3.5-4B	75.17	58.05	77.23
InternVL3.5-8B	84.21	53.52	63.56
Qwen3.5-2B	72.75	51.11	70.25
UnifoLM-ER-1-4B	55.04	37.00	67.23
MiMo-VL-7B-SFT	62.25	23.87	38.35
MiMo-Embodied-7B	61.96	22.74	36.71

6.2 A Public Advantage Reverses on Robot Requests

The ranking above compares models by absolute S. For the four matched pairs, we ask how specialist–base comparisons on published benchmarks carry over to these requests. Embodied-R1.5-8B leads its Qwen3-VL-8B base on all three published dimension means (Yuan et al., 2026). On robot-issued requests, this ordering reverses: the base leads by 4.96 request-score points, scoring 67.73 versus 62.77 for the specialist (Table 4).

Table 4: Embodied-R1.5-8B and its Qwen3-VL-8B base. The first three rows are dimension means reported in Yuan et al. (2026); the final row is S on our requests. Scores are comparable within rows, not across evaluations. Bold marks the higher score in each row.

Evaluation	**Qwen3-VL-8B**	**Embodied-R1.5-8B**
Public: planning	54.2	**65.3**
Public: pointing	59.2	**72.8**
Public: cognition	67.0	**70.2**
Robot-issued requests: S	**67.73**	62.77

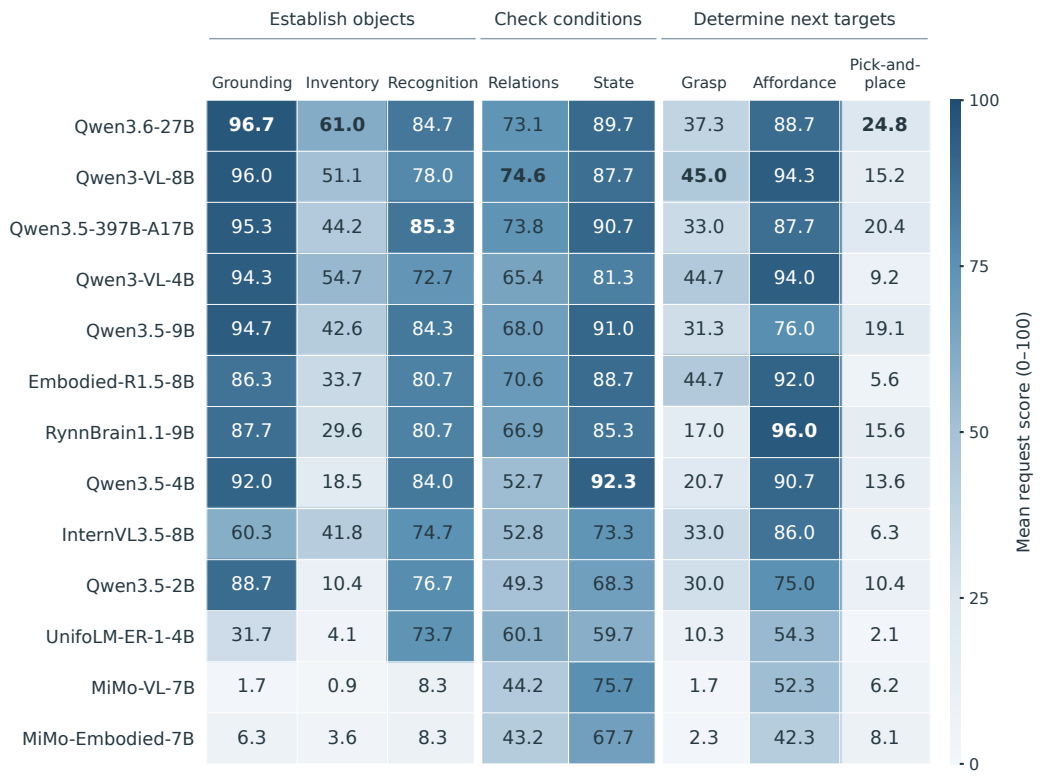


Figure 2: Request scores by request family for the 13 open-weight models. Each cell averages 150 episodes and both language forms; rows follow overall S. Bold marks the highest displayed score in each column. Colors show scores under each family's scoring rule. Model names are shortened; full names appear in Appendix B.1.

The other three pairs have different profiles. UnifoLM-ER-1-4B scores 37.00 versus 64.54 for its base; MiMo-Embodied-7B scores 22.74 versus 23.87; and RynnBrain1.1-9B scores 59.85 versus 63.37. RynnBrain's overall score also hides a lead on one request family.

6.3 Specialist Strengths Depend on the Request Family

That lead is on affordance. RynnBrain1.1-9B exceeds Qwen3.5-9B by 20.0 points on affordance requests and trails by 14.3 on grasp-related requests (Figure 3). Both families involve locations. Affordance scores one point against a distance tolerance, whereas grasp variants also require a declaration or box localization. The specialist's relative strength is therefore specific to the requested job.

The other pairs show three further patterns. Embodied-R1.5's largest deficits are on inventory (-17.4), grounding (-9.7), and pick-and-place (-9.6); its recognition, manipulation-state, and grasp intervals include zero. UnifoLM trails on seven of eight families, and the intervals exclude zero on every family except recognition. MiMo differs from its base in both directions. Absolute scores are in Figure 2, and the paired intervals are in Appendix C.2.

6.4 Scoring and Language Checks

Two checks examine sensitivity to scoring choices. A conservative label-equivalence mapping changes overall scores by at most 0.35 points and preserves the ordering of the 13 open-weight models (Appendix D.3). Qwen3-VL-8B retains an approximately 4.30-point advantage over Embodied-R1.5-8B when pick-and-place is excluded from the eight-family mean (Appendix D.2).

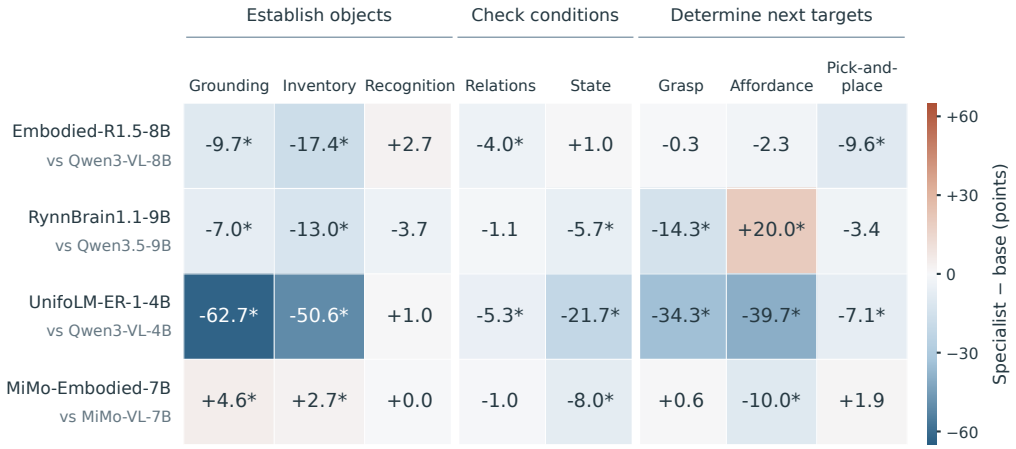


Figure 3: Specialist-minus-base request-score differences on the eight request families. Positive values (orange) favor the specialist; negative values (blue) favor the base. Each cell covers 150 episodes and both language forms. Stars mark unadjusted paired 95% intervals that exclude zero; these family comparisons are exploratory.

For prompt language, 11 of the 13 open-weight models score higher on Chinese prompts on average. InternVL3.5-8B and MiMo-VL-7B-SFT score higher on English prompts, by 3.43 and 1.06 points, respectively. The paired forms mix prompt phrasing with contract-required Chinese target literals, so this difference does not isolate cross-lingual ability (Appendix D.4).

7 Discussion and Conclusion

Our benchmark evaluates VLMs at the robot–VLM interface using requests collected from a running harness, retaining their visual evidence, decisions, and answer requirements. Under the reported evaluation protocol, embodied specialists have lower overall request-score point estimates than their corresponding bases in all four tested pairs. Embodied-R1.5-8B shows a clear reversal of its published advantage over Qwen3-VL-8B. These results show why model selection should include evaluation on the actual images, decisions, and answer requirements of the target robot system.

The results also show why overall rankings need closer examination. A specialist can improve on one request family while performing worse on another, making its strengths relevant to specific parts of the workflow. In practice, developers should compare models on the request types used by their harness and examine both contract acceptance and answer correctness.

The current release contains 1,200 requests from successful episodes of one robot harness, with equal coverage of eight request families. Additional harnesses, failed episodes, and recovery requests would allow us to examine whether the same patterns hold across a wider range of operating conditions. A further step is to run the compared models within the same harness and test whether higher request scores predict greater task success. This would connect the decisions measured by the benchmark to their effects on robot execution.

AI USE STATEMENT

Large language models assisted manuscript editing and figure-generation code. Reference answers were proposed by the models named in Appendix A.2 and reviewed through the procedure described there. The authors are responsible for the manuscript, the reference annotations, and the benchmark artifacts.

ETHICS STATEMENT

Images were reviewed for faces and other information that identifies a person. Visible instances were removed or the item was excluded. Reference answers were reviewed by the authors, who are responsible for the annotations.

REPRODUCIBILITY STATEMENT

Appendix A records how requests were selected, how references were constructed, and how responses are extracted and scored. Appendix B records the evaluated models and the inference settings. The supplementary archive contains 40 requests, five from each family, with the released image, the Chinese and English prompts, and the reference, together with the extractor and the contract scorer. Scoring each reference against itself accepts all 40 items.

That archive is a review sample, not the 1,200-request release. Faces and other personal information were removed from the images, but the original images and prompts can still show the authors' institution, location, or other details that would break double-blind review. The full release is withheld from this anonymous submission for that reason.

REFERENCES

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J. Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. 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.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. URL https://arxiv.org/abs/2307.15818.

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

Kaiyuan Eric Chen, Shuangyu Xie, Zehan Ma, Pannag Sanketi, and Ken Goldberg. Robo2VLM: Improving visual question answering using large-scale robot manipulation data. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/

hash/1f467c3e37abf9f86c78f44c6a27ee7c-Abstract-Datasets_and_Benchmarks_Track.html.

Yanzhe Chen, Zechen Bai, Zhijun Cao, Wenzheng Zeng, Kevin Qinghong Lin, Yiqi Lin, Guoqiang Liang, Kevin Yuchen Ma, Qiming Huang, and Mike Zheng Shou. Show-Harness: Just a VLM agent can play robots. *arXiv preprint arXiv:2609.10522*, 2026. URL https://arxiv.org/abs/2609.10522.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. URL https://arxiv.org/abs/2303.03378.

Di Feng, Kaixin Ma, Feng Nan, Haofeng Chen, Bohan Zhai, David Griffiths, Mingfei Gao, Zhe Gan, Eshan Verma, Yinfei Yang, Zhifeng Chen, and Afshin Dehghan. SO-Bench: A structural output evaluation of multimodal LLMs. *arXiv preprint arXiv:2511.21750*, 2025. URL https://arxiv.org/abs/2511.21750.

Xiaoshuai Hao, Lei Zhou, Zhijian Huang, Zhiwen Hou, Yingbo Tang, Lingfeng Zhang, Guang Li, Zheng Lu, Shuhuai Ren, Xianhui Meng, Yuchen Zhang, Jing Wu, Jinghui Lu, Chenxu Dang, Jiayi Guan, Jianhua Wu, Zhiyi Hou, Hanbing Li, Shumeng Xia, Mingliang Zhou, Yinan Zheng, Zihao Yue, Shuhao Gu, Hao Tian, Yuannan Shen, Jianwei Cui, Wen Zhang, Shaoqing Xu, Bing Wang, Haiyang Sun, Zeyu Zhu, Yuncheng Jiang, Zibin Guo, Chuhong Gong, Chaofan Zhang, Wenbo Ding, Kun Ma, Guang Chen, Rui Cai, Diyun Xiang, Heng Qu, Fuli Luo, Hangjun Ye, and Long Chen. MiMo-Embodied: X-Embodied foundation model technical report. *arXiv preprint arXiv:2511.16518*, 2025. URL https://arxiv.org/abs/2511.16518.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022. URL https://arxiv.org/abs/2207.05608.

Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. 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, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. URL https://arxiv.org/abs/2406.09246.

Kehan Li, Bohan Hou, Minghao Zhu, Tianyi Zhang, Zesen Cheng, Zhikai Wang, Sicong Leng, Xin Li, Xiao Lin, Biying Yao, Minghua Zeng, Jiangpin Liu, Ronghao Dang, Jiayan Guo, Siteng Huang, Haoyu Zhao, Heng Ping, Yaxi Zhao, Tong Zhao, Kexiang Wang, Tong Lu, Shengke Xue, Jiahao Tang, Yulei Wang, Zejing Wang, Jianwei Gao, Shijian Lu, Chengju Liu, Jianfei Yang, Mingxiu Chen, and Deli Zhao. RynnBrain 1.1: Towards more capable and generalizable embodied foundation model. *arXiv preprint arXiv:2607.17977*, 2026. URL https://arxiv.org/abs/2607.17977.

Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Li Erran Li, Ruohan Zhang, Weiyu Liu, Percy Liang, Li Fei-Fei, Jiayuan Mao, and Jiajun Wu. Embodied agent interface: Benchmarking LLMs for embodied decision making. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b631da756d1573c24c9ba9c702fde5a9-Abstract-Datasets_and_Benchmarks_Track.html.

Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, Hang Li, and Tao Kong. Vision-language foundation models as effective robot imitators. *arXiv preprint arXiv:2311.01378*, 2023. URL https://arxiv.org/abs/2311.01378.

Peiqi Liu, Yaswanth Orru, Jay Vakil, Chris Paxton, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. OK-Robot: What really matters in integrating open-knowledge models for robotics. *arXiv preprint arXiv:2401.12202*, 2024. URL https://arxiv.org/abs/2401.12202.

Yulin Luo, Chun-Kai Fan, Menghang Dong, Jiayu Shi, Xiangju Mi, Mengdi Zhao, Bo-Wen Zhang, Cheng Chi, Jiaming Liu, Gaole Dai, Rongyu Zhang, Ruichuan An, Kun Wu, Zhengping Che, Shaoxuan Xie, Guocai Yao, Zhongxia Zhao, Pengwei Wang, Guang Liu, Zhongyuan Wang, Tiejun Huang, and Shanghang Zhang. RoboBench: A comprehensive evaluation benchmark for multimodal large language models as embodied brain. *arXiv preprint arXiv:2510.17801*, 2025. URL https://arxiv.org/abs/2510.17801.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, Mrinal Kalakrishnan, Franziska Meier, Chris Paxton, Alexander Sax, and Aravind Rajeswaran. OpenEQA: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16488–16498, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Majumdar_OpenEQA_Embodied_Question_Answering_in_the_Era_of_Foundation_Models_CVPR_2024_paper.html.

OpenAI. GPT-6 Astra: A new generation of intelligence, 2026. URL https://openai.com/index/gpt-6-astra/.

Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J. Joshi, Pete Florence, Wei Han, Robert Baruch, Yao Lu, Suvir Mirchandani, Peng Xu, Pannag Sanketi, Karol Hausman, Izhak Shafran, Brian Ichter, and Yuan Cao. RoboVQA: Multimodal long-horizon reasoning for robotics. *arXiv preprint arXiv:2311.00899*, 2023. URL https://arxiv.org/abs/2311.00899.

Zhonghao Shi, Enyu Zhao, Nathaniel Dennler, Jingzhen Wang, Xinyang Xu, Kaleen Shrestha, Mengxue Fu, Daniel Seita, and Maja J. Mataric. HRIBench: Benchmarking vision-language models for real-time human perception in human-robot interaction. *arXiv preprint arXiv:2506.20566*, 2025. URL https://arxiv.org/abs/2506.20566.

Jiayi Su, Yixin Zheng, Mi Yan, Li Yi, Zhizheng Zhang, and He Wang. GPT 6 Astra as an embodied policy. Technical report and code, 2026. URL https://github.com/anonymous-report-421/GPT-as-Policy.

Unitree Robotics. UnifoLM-ER-1-4B. Hugging Face model card, 2026. URL https://huggingface.co/unitreerobotics/UnifoLM-ER-1.

Dayu Xia, Yue Shi, Yao Mu, Huiting Ji, Chaofan Ma, Yingjie Zhou, Hua Chen, Yang Liu, Jiezhang Cao, and Guangtao Zhai. RoboProcessBench: Benchmarking process-aware understanding in vision-language robotic manipulation. *arXiv preprint arXiv:2606.13040*, 2026. URL https://arxiv.org/abs/2606.13040.

Shuangyu Xie, Kaiyuan Chen, Ziyang Chen, Simeon Adebola, Yixuan Huang, Zehan Ma, Tianshuang Qiu, Wentao Yuan, Dhruv Shah, Pannag R. Sanketi, and Ken Goldberg. RoboVista: Evaluating vision language models for diverse robot applications. In *Proceedings of Robotics: Science and Systems*, 2026. doi: 10.15607/RSS.2026.XXII.095. URL https://www.roboticsproceedings.org/rss22/p095.html.

Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, Heng Ji, Huan Zhang, and Tong Zhang.

EmbodiedBench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. *arXiv preprint arXiv:2502.09560*, 2025. URL https://arxiv.org/abs/2502.09560.

Yifu Yuan, Yaoting Huang, Xianze Yao, Yutong Li, Shuoheng Zhang, Linqi Han, Pengyi Li, Jiangeng Sun, Wenting Jia, Zhao Zhang, Yuhao Liu, Ruihao Liao, Yucheng Hu, Qiyu Wu, Yuxiao Li, Zibin Dong, Fei Ni, Yan Zheng, Shuyang Gu, Yi Ma, Hongyao Tang, Han Hu, and Jianye Hao. Embodied-R1.5: Evolving physical intelligence via embodied foundation models. *arXiv preprint arXiv:2606.11324*, 2026. URL https://arxiv.org/abs/2606.11324.

Enyu Zhao, Vedant Raval, Hejia Zhang, Jiageng Mao, Zeyu Shangguan, Stefanos Nikolaidis, Yue Wang, and Daniel Seita. ManipBench: Benchmarking vision-language models for low-level robot manipulation. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 3413–3462. PMLR, 2025. URL https://proceedings.mlr.press/v305/zhao25a.html.

Enshen Zhou, Qi Su, Cheng Chi, Zhizheng Zhang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, and He Wang. Code-as-monitor: Constraint-aware visual programming for reactive and proactive robotic failure detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6919–6929, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhou_Code-as-Monitor_Constraint-aware_Visual_Programming_for_Reactive_and_Proactive_Robotic_Failure_CVPR_2025_paper.html.

A BENCHMARK DETAILS

This appendix expands Section 4 and Section 5: how requests were selected and released (A.1), how references were constructed and verified (A.2), and how the evaluator extracts and scores responses (A.3).

A.1 REQUEST SELECTION AND RELEASE COMPOSITION

The release is built from calls logged by the binocular-camera harness. We remove exact and near-duplicate captures and calls with residual or ambiguous request types. A step may have been retried because of the harness or the physical robot, but only one accepted call from each successful episode enters the release. Sampling then draws 150 calls from each of eight strata defined by the requested decision and contract, favoring diversity in answer type, target, visual view, and capture context. It uses neither the evaluated models' responses nor their scores. This produces 1,200 requests from 1,200 episodes. Every released request has a determinate reference on its released view; Appendix A.2 states when an item is revised or removed. Table 5 summarizes the release. The VLM serving the harness during collection was not a Qwen3-VL model; company information-security policy prevents us from naming that checkpoint.

Table 5: Composition of the released benchmark. Visual-payload counts sum to the 1,200 underlying requests.

Released component	**Count**
Distinct successful episodes / underlying requests	1,200
Sampling strata	8
Requests per sampling stratum	150
Chinese and English prompt forms	2,400
Full-frame payloads	915
Processed crops	237
Processed composites	48

Prompts. Each item retains the visual payload sent to the VLM. Prompts are verbatim or clarify the output schema or coordinate convention without changing the target, precondition, or decision, so they preserve the evidence and task without being byte-for-byte replays of every production call.

Paired language forms. The operation stream contains both Chinese and English calls. Each released item has aligned forms that share an image, decision, and reference. Back-translation checks quantifiers, negation, thresholds, and output requirements. Contract-specific target literals remain unchanged, including Chinese literals in English prompts, and the English pick-and-place contract requires Chinese output labels because that is what the downstream grounding module consumed (Appendix D.4).

Metadata relabeling. Canonical metadata relabels seven grasp items as grounding because their job is target localization, giving 157 grounding and 143 grasp items with 150 in each other family. All reported family scores still use the original eight strata of 150; images, prompts, references, and aggregation are unchanged.

A.2 REFERENCE CONSTRUCTION AND VERIFICATION

This subsection records how the references summarized in Section 4 were produced and checked. Two models propose a reference. A human adjudicates when those proposals do not agree. The same adjudicator then audits a sample of the unflagged items, a second annotator verifies a separate sample, and a physical execution checks that the final references can be carried out.

Proposals and adjudication. Qwen3.8 Max and Grok 4.6 independently propose references from the released image and prompt, without the historical service reply or episode context. Human adjudication covers 254 clear disagreements, 183 partial agreements, and 306 requests without an agreement score (743 in total). The remaining 457 requests do not trigger review under this rule. The reviewer may inspect the episode to resolve ambiguity, but a retained reference must be supported by the released view. If it is not, the item is revised or removed. References record permitted answer sets or equivalence classes, including prescribed failure branches, and the reviewer checks that privacy redaction has not changed the decision.

Audit of unflagged references. The adjudicator also sampled 100 of the 457 unflagged requests and agreed with all 100 recorded references. Because the same person conducted both checks, this is not an inter-annotator agreement estimate.

Second-annotator verification. A second annotator reviewed 200 requests sampled without consulting evaluated-model scores: 50 with complete proposal-model agreement and no adjudication, and 150 from the remaining requests. The annotator saw the released image, prompt, contract, and recorded reference, but not the proposal outputs or the first annotator's labels; the task was to verify a displayed reference, not to produce an independent answer. Agreement required full credit under the contract (score 1), the same criterion as the proposal comparison: boxes had to match the existence branch and reach IoU at least 0.5, points had to lie within distance 60 in normalized 0–1000 coordinates, and where multiple answers were legal the displayed answer had to belong to an equivalence class already recorded in the reference. Subthreshold overlap, points beyond the tolerance, and partially matching sets counted as disagreements.

Before joint resolution, the second annotator accepted 98% of the 200 displayed references. The annotators then jointly revised disagreements to obtain final references; no post-resolution agreement rate was computed. The 98% measures reference verification on this sample, not a family-specific error rate, and is separate from the 100-request audit and the physical executions.

Physical execution of references. Final references were executed on the physical robot for 200 requests: 20 where two evaluated models agreed and 180 where they differed. All 200 executions succeeded. This sample is separate from both reference reviews. Because the trials executed references rather than model outputs, they establish executability only; they do not test whether request-score differences predict robot success.

A.3 Answer Extraction and Contract Scoring

This subsection gives the extraction, acceptance, and scoring rules summarized in Section 5.

Extraction. The evaluator extracts an answer block when present, then a fenced response or parseable JSON content, and otherwise retains the response text. It recursively treats `bbox`, `bbox_2d`, `box_2d`, `box2d`, and `bounding_box` as equivalent box keys. Conflicting values under multiple aliases cause rejection. This processing is deterministic and is applied to every model before the contract check.

Acceptance versus production compliance. The check then decides acceptance. It requires the fields, types, and geometric form specified by the contract; some contracts also require reference structure, such as the full index set for inventory. Acceptance is not correctness. For the conditional grounding contract, a reply with a Boolean existence field and a legal box passes the structural check even when the target is absent; a wrong existence decision receives $s = 0$ with $a = 1$. This offline acceptance is also distinct from the production parser in Figure 1. The illustrated production request may instead require a specific failure token, and scorer acceptance does not certify production-parser acceptance.

A rejected response scores zero. An accepted response is then scored under its contract.

Indexed inventory. Let K be the reference index set. A predicted list must contain exactly that set, use an allowed status for each entry, and provide a legal box for visible entries; non-visible

entries must have a null box. For accepted responses, the score is

$$s = \frac{1}{|K|} \sum_{k \in K} \mathbf{1}[\text{label and status match at } k]\, \mathbf{1}[\text{non-visible at } k \text{ or } \text{IoU}_k \geq 0.5]\,.$$

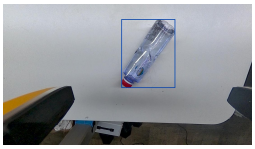
Labels use the contract's text normalization. This is an average over indexed reference objects, not set F1. Duplicate predicted indices retain their final entry in the parsed index mapping, and if both index sets are empty an accepted response scores one.

Pointing and grasp variants. Affordance compares one predicted point with one reference point using Euclidean distance at most 60 in normalized 0–1000 coordinates. The tolerance is a geometric scoring region around that reference, not an estimate of all physically feasible interaction locations. A finished response is accepted structurally and scores according to the reference branch; extra fields are not universally rejected. Grasp subcontracts differ: choose-and-declare requires the reference identity and box overlap, while a free-declaration subcontract requires a nonempty identity string and box overlap. Joint box-and-point requests additionally require point proximity and containment in the reference box. Contracts with unspecified box units accept either the documented normalized or the original-pixel interpretation.

Text and task lists. Text normalization includes Unicode normalization, whitespace handling, and case folding where the contract specifies it. Pick-and-place compares sets of normalized label pairs after removing a trailing `:dof` suffix. Duplicate pairs collapse to one set member; two empty sets receive F1 of one, and one empty set against a nonempty set receives zero. Closed-state, OCR, and relation contracts retain their own field and normalization rules. Appendix D.3 reports a separate label-equivalence diagnostic that leaves the primary score unchanged.

A.4 Released Examples

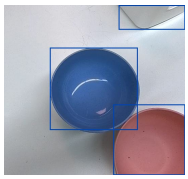
Figure 4 shows one released item from each family. The image is the visual payload sent to the model. Where a blue mark is drawn, it is the reference used for scoring: a box, the indexed boxes of an inventory, or one affordance point. The words under each panel state the decision and the reference. They are not the full production prompt.



(a) Grounding

Request: Locate the water bottle.

Reference: Marked box.



(b) Inventory

Request: List objects, statuses, and boxes.

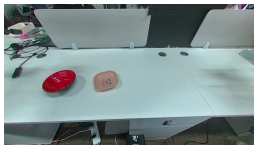
Reference: Blue bowl, pink bowl, and partial white object, each with a status and box.



(c) Recognition

Request: Classify the visible shape.

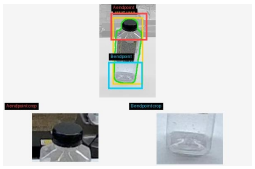
Reference: `bowl`.



(d) Relation

Request: Are the black-red bowl and pink plate both visible and not stacked?

Reference: Both conditions pass.



(e) State

Request: Identify the cap and bottom in the enlarged views.

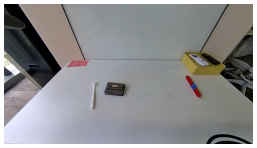
Reference: Cap A; bottom B.



(f) Grasp

Request: Declare a grasp of the pen among the blocks.

Reference: Marked box.



(g) Affordance

Request: Choose a point on the red sausage.

Reference: Marked point.



(h) Pick-and-place

Request: List the remaining pick/place goals.

Reference: Green block, pink object, and yellow-green object, each into the large green bin.

Figure 4: One released request from each family. Blue marks are references drawn for this figure. The manipulation-state panel already contains the endpoint crops and boxes that the harness sent.

B Models and Published Comparisons

B.1 Evaluated Models

We evaluate 15 vision–language models on the same 2,400 prompts. Eight of them form four matched pairs. In each pair, an embodied finetune is compared with the base checkpoint released for that finetune. The pairs are Embodied-R1.5-8B with Qwen3-VL-8B-Instruct, RynnBrain1.1-9B with Qwen3.5-9B, UnifoLM-ER-1-4B with Qwen3-VL-4B-Instruct, and MiMo-Embodied-7B with MiMo-VL-7B-SFT. The released `MiMo-VL-7B-SFT` weights are the 7B-SFT-2508 checkpoint named in the MiMo-Embodied report. We use that checkpoint as the base, so the pair matches the weights named by the specialist's authors.

The other seven models supply context for the overall ranking. They are not one side of a matched pair. Five are additional open-weight models: Qwen3.6-27B, Qwen3.5-397B-A17B, Qwen3.5-4B, Qwen3.5-2B, and InternVL3.5-8B. Two are closed APIs, Qwen3.7-plus and GPT-6-astra, whose providers do not disclose a parameter count.

Table 6 lists every model in the same order as the main results table. The columns do not all use one meaning of "base," so we state the reading rule here. For an embodied finetune, Base is the checkpoint in its matched comparison, such as Qwen3-VL-8B for Embodied-R1.5-8B. For a model that is not a finetune, Base names the public model family, such as Qwen3.6 for Qwen3.6-27B, and does not mean that a second checkpoint was evaluated. Embodied is Yes only for the four finetunes. Scale is the public parameter count; for Qwen3.5-397B-A17B the two numbers are the total and the activated parameters. Weights are Open when the checkpoint can be downloaded. The two APIs are closed, and their scale cells are undisclosed.

The last column names the document from which Appendix B.3 copies published specialist-versus-base scores. It is not a claim that the model was trained on that document's benchmark. Embodied-R1.5-8B and Qwen3-VL-8B-Instruct both point to EmbodiedEvalKit because that source reports the three dimension means for both models. Both MiMo rows point to Tables 2 and 3 of the MiMo-Embodied report. RynnBrain1.1-9B and UnifoLM-ER-1-4B point to their own reports. Qwen3.5-9B and Qwen3-VL-4B-Instruct are left blank because their published numbers appear as the base columns of those specialist reports, not as a separate comparison. The seven context models are also blank because this paper does not pair them with a published base. Those published numbers are reproduced in Appendix B.3. They are copied from the cited tables and are not averaged into one public score.

Table 6: Lineage of the 15 evaluated models, ordered as in Table 3. Scale is the public parameter count; for the mixture-of-experts model, the second number is activated parameters. "Open" means downloadable weights. Qwen3.7-plus and GPT-6-astra are closed APIs whose providers disclose no parameter count. "Embodied" marks an embodied finetune. The last column names the published specialist-versus-base comparison used in this paper.

Model	Base	Embodied	Scale	Weights	Comparison source
GPT-6-astra	—	No	undisclosed	Closed API	—
Qwen3.7-plus	—	No	undisclosed	Closed API	—
Qwen3.6-27B	Qwen3.6	No	27B	Open	—
Qwen3-VL-8B-Instruct	Qwen3-VL	No	8B	Open	EmbodiedEvalKit
Qwen3.5-397B-A17B	Qwen3.5	No	397B / 17B	Open	—
Qwen3-VL-4B-Instruct	Qwen3-VL	No	4B	Open	—
Qwen3.5-9B	Qwen3.5	No	9B	Open	—
Embodied-R1.5-8B	Qwen3-VL-8B	Yes	8B	Open	EmbodiedEvalKit
RynnBrain1.1-9B	Qwen3.5-9B	Yes	9B	Open	RynnBrain 1.1
Qwen3.5-4B	Qwen3.5	No	4B	Open	—
InternVL3.5-8B	InternVL3.5	No	8B	Open	—
Qwen3.5-2B	Qwen3.5	No	2B	Open	—
UnifoLM-ER-1-4B	Qwen3-VL-4B	Yes	4B	Open	ER-1 card
MiMo-VL-7B-SFT	MiMo-VL	No	7B	Open	MiMo Tabs. 2–3
MiMo-Embodied-7B	MiMo-VL-7B-SFT	Yes	7B	Open	MiMo Tabs. 2–3

B.2 Inference Protocol

Every model receives the same 2,400 released prompts. The protocol below is what Section 5 summarizes as greedy or temperature-zero decoding, a configured output-length limit of 1,536, and thinking disabled where the interface supports it. Table 7 puts the local and API settings side by side and reports each interface's parameter name for that limit.

Twelve models ran from local checkpoints and used each checkpoint's official chat template. Three models were requested by the names in the table through APIs: Qwen3.5-397B-A17B, Qwen3.7-plus, and GPT-6-astra. Qwen3.5-397B-A17B has open weights, but this evaluation called it through an API rather than loading a local checkpoint. Local runs use greedy decoding. API runs set temperature 0 through the control those interfaces expose. Where an interface accepts the flag, thinking is turned off with `enable_thinking=False`. GPT-6-astra does not use that flag; it runs at reasoning effort low, as stated in the main text.

The image is the released visual payload, not a new crop made at evaluation time. Local runs convert it to RGB and use model-specific image preprocessing. InternVL3.5-8B uses dynamic tiling into up to twelve 448×448 patches, with an additional thumbnail for multi-patch inputs, and its native chat interface. API runs send the same released image. Before the call, a leading `<image>` placeholder is removed from the prompt text and the image is passed as a separate input. One local model does not use its own image processor. RynnBrain1.1-9B is loaded from its own weights, while its image processor is the Qwen3.5-9B processor.

A call that fails is recorded and is not retried inside that pass. The same rule applies to local checkpoints and to APIs. A failed call therefore remains a missing response for that pass rather than a later, possibly different decode.

Table 7: Inference protocol for all 15 models. API models are Qwen3.5-397B-A17B, Qwen3.7-plus, and GPT-6-astra. Shared limits use each interface's parameter names.

	Local checkpoints (12 models)	API models (3 models)
Interface	Official chat template	API requests
Reasoning control	`enable_thinking=False` where the interface accepts the flag	Qwen: `enable_thinking=False`; GPT-6-astra: reasoning effort low
Decoding	Greedy	Temperature 0
Length limit	1,536 new tokens	1,536 tokens
Image	Released image, converted to RGB	Same released image, sent as-is
Retries	Failed calls recorded, not retried in-pass	Failed calls recorded, not retried in-pass

B.3 Published Comparison Scores

This subsection copies the published specialist-versus-base numbers used to read Section 6.2. They come from the sources named in Table 6. They are not on the same scale as our request score S, and a higher published number does not by itself predict a higher S.

Table 8 reproduces three EmbodiedEvalKit dimension means from the Embodied-R1.5 report (Yuan et al., 2026): planning, pointing, and cognition. The report gives those means for Qwen3-VL-8B-Instruct, Embodied-R1.5-8B, and MiMo-Embodied-7B. It does not include MiMo-VL, so this table has no base row for MiMo. The MiMo base scores are the ones copied separately in Table 9. The column labeled public mean is the unweighted average of the three printed dimensions. The main comparison uses the three dimensions themselves. That average is not a score we compare with S.

Table 8: Published EmbodiedEvalKit dimension means from the Embodied-R1.5 report (Yuan et al., 2026), alongside our request score S. The public mean is the unweighted mean of the three printed dimensions; the two evaluations use different tasks and scales.

	Ours	**Published EmbodiedEvalKit**			
Model	***S***	Planning	Pointing	Cognition	Mean
Qwen3-VL-8B-Instruct	67.73	54.2	59.2	67.0	60.1
Embodied-R1.5-8B	62.77	65.3	72.8	70.2	69.4
MiMo-Embodied-7B	22.74	54.1	55.1	65.6	58.3

Table 9 copies selected specialist and base columns for the other three pairs. Each row is one published benchmark. The columns are not averaged into a public total. A dagger marks a UnifoLM base number that the UnifoLM source copied from an official report or another public paper, rather than a number measured in that source. An asterisk marks a MiMo base number that the MiMo source describes as a re-evaluation. In this table the specialist is not always ahead. RynnBrain trails Qwen3.5-9B on EmbSpatial. UnifoLM trails Qwen3-VL-4B on VSI-Bench and RealWorldQA. MiMo-Embodied leads MiMo-VL on all five listed benchmarks. On our requests, the paired 95% interval for Qwen3-VL-8B minus Embodied-R1.5 is 3.45–6.49 around the 4.96-point gap. MiMo-Embodied minus its base is -1.13 points, with interval $[-2.46, 0.22]$.

C Complete Family-Level Results

This section gives the numbers behind the two family figures in Section 6. The overall table covers all 15 models. The family table, heatmap, and diagnostic tables in Appendix D use the same set of 13 open-weight models.

Table 9: Selected published specialist and base scores, copied from the cited tables. Specialist is listed first. A dagger marks a UnifoLM base number copied by that source from an official report or another public paper. An asterisk marks a MiMo base number that the source describes as a re-evaluation.

Pair	Benchmark	Specialist	Base
RynnBrain 1.1-9B vs. Qwen3.5-9B (Li et al., 2026)			
	VSI-Bench	74.9	61.0
	EmbSpatial	81.9	83.0
	RynnBrain-Grounding	85.4	33.0
	RynnBrain-Spatial	67.9	28.4
	RynnBrain-Affordance	90.7	60.5
	RynnBrain-Trajectory	70.0	45.0
UnifoLM-ER-1-4B vs. Qwen3-VL-4B (Unitree Robotics, 2026)			
	RoboVQA	62.4	47.7†
	Where2Place	82.0	63.0†
	RefSpatial-Bench	61.7	46.6†
	EmbSpatial	88.9	79.6†
	VSI-Bench	54.2	59.3†
	RealWorldQA	69.8	71.0†
MiMo-Embodied-7B vs. MiMo-VL (Hao et al., 2025)			
	Where2Place	63.60	29.60*
	RoboVQA	61.99	35.27*
	EmbSpatial	76.24	71.29*
	RefSpatial	48.00	29.79*
	VSI-Bench	48.49	42.10*

C.1 Family Scores

Table 10 is the numerical form of Figure 2. Each cell is the mean request score for one model on one original sampling stratum: 150 requests, with the Chinese and English forms averaged together. Rows follow overall S. A column is not a common unit of ability. Grounding, inventory, grasp, affordance, and pick-and-place use different contracts, so a high score in one column and a high score in another are not the same kind of success.

The main text reads three low ceilings and one within-model spread from this table. Among the 13 open-weight models, the highest pick-and-place score is 24.8, the highest inventory score is 61.0, and the highest grasp score is 45.0. Qwen3.6-27B scores 96.7 on single-object grounding, 61.0 on inventory, and 24.8 on pick-and-place. Several models exceed 90 on single-object grounding, manipulation state, and affordance. The low columns are the contracts that ask for a complete set, a declaration, or a goal list.

The last row is not another evaluated model. It is a majority-class constant predictor: on each family it emits the most frequent reference answer and is scored with that family's contract. The row shows how far a column sits above a constant answer. Relation, state, and recognition have higher constant scores because a frequent reference answer exists; grounding, inventory, grasp, and affordance do not.

C.2 Matched Family Differences

Table 11 is the numerical form of Figure 3. Each cell is the specialist minus its matched base on one family, in request-score points. A positive number means the specialist scored higher. The brackets are 95% paired bootstrap intervals. Each draw resamples the 150 episodes in that family and keeps an episode's Chinese and English scores together, with 10,000 draws. A bold difference indicates an interval that excludes zero. Four pairs and eight families give 32 comparisons. They are exploratory, they are not adjusted for testing all 32, and they do not include uncertainty from reference annotation.

Table 10: Family scores for the 13 open-weight models. Each cell averages 150 requests and both language forms. The last row is a constant predictor, not an evaluated model.

Model	Ground.	Invent.	Relat.	State	Grasp	Afford.	Pick/ place	Recogn.
Qwen3.6-27B	96.7	61.0	73.1	89.7	37.3	88.7	24.8	84.7
Qwen3-VL-8B-Instruct	96.0	51.1	74.6	87.7	45.0	94.3	15.2	78.0
Qwen3.5-397B-A17B	95.3	44.2	73.8	90.7	33.0	87.7	20.4	85.3
Qwen3-VL-4B-Instruct	94.3	54.7	65.4	81.3	44.7	94.0	9.2	72.7
Qwen3.5-9B	94.7	42.6	68.0	91.0	31.3	76.0	19.1	84.3
Embodied-R1.5-8B	86.3	33.7	70.6	88.7	44.7	92.0	5.6	80.7
RynnBrain1.1-9B	87.7	29.6	66.9	85.3	17.0	96.0	15.6	80.7
Qwen3.5-4B	92.0	18.5	52.7	92.3	20.7	90.7	13.6	84.0
InternVL3.5-8B	60.3	41.8	52.8	73.3	33.0	86.0	6.3	74.7
Qwen3.5-2B	88.7	10.4	49.3	68.3	30.0	75.0	10.4	76.7
UnifoLM-ER-1-4B	31.7	4.1	60.1	59.7	10.3	54.3	2.1	73.7
MiMo-VL-7B-SFT	1.7	0.9	44.2	75.7	1.7	52.3	6.2	8.3
MiMo-Embodied-7B	6.3	3.6	43.2	67.7	2.3	42.3	8.1	8.3
Majority-class constant	2.0	2.0	37.6	35.3	2.0	3.3	7.3	46.0

The columns are the four patterns in the main text. Embodied-R1.5 trails its base by 17.4 on inventory, 9.7 on grounding, and 9.6 on pick-and-place. Its recognition, manipulation-state, and grasp intervals include zero. RynnBrain leads by 20.0 on affordance and trails by 14.3 on grasp. UnifoLM trails on seven families; only its recognition interval includes zero. MiMo is higher on grounding and inventory and lower on manipulation state and affordance, so that pair differs in both directions.

Table 11: Specialist minus base, in request-score points. Each difference is followed by its 95% paired bootstrap interval on the line below. Bold differences have intervals excluding zero.

Family	**R1.5** vs Q3-VL-8B	**Rynn** vs Q3.5-9B	**ER1** vs Q3-VL-4B	**MiMo** vs MiMo-VL
Grounding	**-9.7** [-14.3, -5.3]	**-7.0** [-11.3, -3.0]	**-62.7** [-69.0, -56.3]	**+4.6** [+1.7, +8.0]
Inventory	**-17.4** [-22.7, -12.2]	**-13.0** [-19.1, -6.9]	**-50.6** [-57.1, -43.9]	**+2.7** [+1.0, +4.6]
Recognition	+2.7 [-2.3, +7.7]	-3.7 [-7.3, +0.0]	+1.0 [-4.0, +6.0]	+0.0 [-2.7, +2.7]
Relations	**-4.0** [-7.4, -0.9]	-1.1 [-5.3, +2.9]	**-5.3** [-9.7, -0.9]	-1.0 [-4.1, +2.2]
State	+1.0 [-3.7, +5.7]	**-5.7** [-9.7, -1.7]	**-21.7** [-29.7, -13.7]	**-8.0** [-13.7, -2.0]
Grasp	-0.3 [-5.3, +5.0]	**-14.3** [-20.7, -8.3]	**-34.3** [-41.0, -27.3]	+0.6 [-1.0, +2.3]
Affordance	-2.3 [-5.3, +0.3]	**+20.0** [+15.3, +25.0]	**-39.7** [-44.7, -34.7]	**-10.0** [-16.0, -4.3]
Pick-and-place	**-9.6** [-12.4, -7.0]	-3.4 [-7.7, +1.1]	**-7.1** [-10.2, -4.5]	+1.9 [-0.4, +4.3]

D SCORE AND INTERFACE DIAGNOSTICS

Section 6.4 checks whether the reported comparisons are sensitive to scoring choices and prompt language. This section records those checks. It also records how acceptance and scorer outcomes

are distributed, because Section 6.1 shows that acceptance can change which model looks strongest. None of these tables replaces a primary score or a matched-pair interval. The rank, scorer-outcome, paired-language, and label-equivalence tables cover the same 13 open-weight models as the family analysis.

D.1 Acceptance Ranks and Scorer Outcomes

The first two tables describe acceptance for those 13 models. Table 12 ranks them separately by scorer acceptance V and by the score on accepted responses $S_{|V}$. Because $S_{|V}$ uses a different accepted subset for each model, the ranks are descriptive rather than measures of independent abilities. Table 13 then splits each model's 2,400 responses into four scorer outcomes: invalid, valid with zero score, partial credit, and full credit. Invalid responses fail contract-specific acceptance, including reference-index requirements where applicable. Valid-zero and partial-credit responses pass acceptance but fall short of the reference. These categories do not separate formatting errors from semantic errors, nor either from physical causes.

Table 12: Acceptance V and score on accepted responses $S_{|V}$ for 13 open-weight models, with ranks computed within this set. Rows follow the overall request score S; $\Delta\text{rank} = \text{rank}(V) - \text{rank}(S_{|V})$.

Model	V	$S_{\|V}$	Rank V	Rank $S_{\|V}$	Δrank
Qwen3.6-27B	86.83	80.01	3	3	0
Qwen3-VL-8B-Instruct	88.46	76.56	2	6	-4
Qwen3.5-397B-A17B	80.42	82.44	6	1	+5
Qwen3-VL-4B-Instruct	89.12	72.42	1	8	-7
Qwen3.5-9B	78.79	80.43	7	2	+5
Embodied-R1.5-8B	85.29	73.59	4	7	-3
RynnBrain1.1-9B	75.33	79.45	8	4	+4
Qwen3.5-4B	75.17	77.23	9	5	+4
InternVL3.5-8B	84.21	63.56	5	11	-6
Qwen3.5-2B	72.75	70.25	10	9	+1
UnifoLM-ER-1-4B	55.04	67.23	13	10	+3
MiMo-VL-7B-SFT	62.25	38.35	11	12	-1
MiMo-Embodied-7B	61.96	36.71	12	13	-1

Table 14 takes the invalid responses from the two Qwen-based matched pairs and counts them by request family. Those pairs contribute 1,731 invalid responses, and inventory and grasp-related requests account for most of them. The table cannot say whether an invalid response would otherwise have contained a correct decision.

D.2 Excluding Pick-and-Place

Pick-and-place is scored by exact-label set F1, and Embodied-R1.5 trails its base on that family. The 4.96-point reversal could therefore depend on this one column. This calculation removes the column from the equal-weighted mean of the eight families. Because those families have equal weight, excluding family f gives $S_{-f} = (8S - S_f)/7$. Using the displayed overall and pick-and-place scores for Qwen3-VL-8B and Embodied-R1.5-8B, the base-minus-specialist gap becomes approximately

$$\frac{8(67.73 - 62.77) - (15.2 - 5.6)}{7} = 4.30 \text{ points.}$$

The base remains ahead after pick-and-place is removed. The calculation does not change the primary eight-family score. It also does not establish significance for the reduced set; an interval would require request-level scores.

Table 13: Scorer outcomes out of 2,400 responses per model. Outcomes are reference-relative and do not identify visual or physical error causes.

Model	Invalid	Valid zero	Partial	Full
Qwen3.6-27B	316	286	206	1592
Qwen3-VL-8B-Instruct	277	392	156	1575
Qwen3.5-397B-A17B	470	211	195	1524
Qwen3-VL-4B-Instruct	261	512	116	1511
Qwen3.5-9B	509	241	187	1463
Embodied-R1.5-8B	353	503	68	1476
RynnBrain1.1-9B	592	285	118	1405
Qwen3.5-4B	596	308	150	1346
InternVL3.5-8B	379	629	168	1224
Qwen3.5-2B	654	422	135	1189
UnifoLM-ER-1-4B	1079	398	50	873
MiMo-VL-7B-SFT	906	892	41	561
MiMo-Embodied-7B	913	896	63	528

Table 14: Invalid responses by request family for Qwen3-VL-8B, Embodied-R1.5-8B, Qwen3.5-9B, and RynnBrain1.1-9B. Each family contains 1,200 responses across these four models and two language forms; the rate is the within-family invalid rate.

Family	Qwen3-VL 8B	Embodied R1.5-8B	Qwen3.5 9B	RynnBrain 1.1-9B	Total	Rate
Single-object grounding	8	23	13	36	80	6.7%
Set inventory	121	174	149	197	641	53.4%
Recognition / attributes	22	22	22	20	86	7.2%
Spatial relations	55	66	67	76	264	22.0%
Manipulation state	0	0	0	0	0	0.0%
Grasp-related requests	67	56	186	225	534	44.5%
Affordance	4	11	70	5	90	7.5%
Pick-and-place	0	1	2	33	36	3.0%
Total	277	353	509	592	1,731	–

D.3 Pick-and-Place Label Equivalence

The preceding calculation removes the family. This one keeps every family and changes only the label comparison, to see whether exact string matching is what moves the scores. The primary pick-and-place score is exact-label set F1. Table 15 applies the conservative equivalence table shipped with the release. The table deletes the particle *de* and a trailing toy-noun suffix, merges the closed morph pairs specified by the release, and refuses unsafe merges such as a light-green bowl with a green bowl or a black pen-holder with a black pen. Stored references are unchanged; the mapping is applied only during comparison. Overall scores change by at most 0.35 points, and the ordering of the 13 open-weight models is preserved.

Table 15: Pick-and-place scores under exact label match versus the conservative equivalence table. The S columns replace only the pick-and-place family in the mean over the eight strata; the primary tables keep the exact-label score.

	Pick-and-place			**Overall S**		**Rank**	
Model	Exact	Equiv.	Δ	Exact	Equiv.	Exact	Equiv.
Qwen3.6-27B	24.8	26.3	+1.5	69.48	69.67	1	1
Qwen3-VL-8B-Instruct	15.2	16.3	+1.2	67.73	67.87	2	2
Qwen3.5-397B-A17B	20.4	23.3	+2.8	66.30	66.65	3	3
Qwen3-VL-4B-Instruct	9.2	10.3	+1.1	64.54	64.68	4	4
Qwen3.5-9B	19.1	20.6	+1.6	63.37	63.57	5	5
Embodied-R1.5-8B	5.6	5.8	+0.2	62.77	62.80	6	6
RynnBrain1.1-9B	15.6	17.5	+1.8	59.85	60.08	7	7
Qwen3.5-4B	13.6	15.6	+2.0	58.05	58.30	8	8
InternVL3.5-8B	6.3	6.9	+0.6	53.52	53.61	9	9
Qwen3.5-2B	10.4	11.4	+0.9	51.11	51.23	10	10
UnifoLM-ER-1-4B	2.1	2.8	+0.6	37.00	37.08	11	11
MiMo-VL-7B-SFT	6.2	6.6	+0.4	23.87	23.92	12	12
MiMo-Embodied-7B	8.1	8.5	+0.4	22.74	22.80	13	13

D.4 Paired Chinese–English Forms

Each request has a Chinese form and an English form. The two forms share the image, the decision, and the reference. Table 16 reports the paired difference $S_{zh} - S_{en}$ for the 13 open-weight models. The intervals are 95% bootstrap intervals over 10,000 family-stratified draws of question pairs. They are exploratory: unadjusted for testing multiple models, and conditional on the observed items. Eleven models have positive Chinese-minus-English point estimates. The largest Chinese advantages are Qwen3.5-9B at 5.98 points, Qwen3-VL-4B at 4.35, and MiMo-Embodied-7B at 3.86. InternVL3.5-8B has a -3.43-point difference (95% interval: $[-5.65, -1.19]$), and MiMo-VL-7B-SFT has a -1.06-point difference ($[-2.69, 0.56]$).

Dependence on family and output vocabulary. The difference is concentrated in particular families. For Qwen3-VL-8B, the overall 2.91-point difference decomposes into 14.67 points on inventory, 6.00 on grasp, and 3.76 on pick-and-place, against -2.44 on relations and -2.00 on manipulation state, with single-object grounding exactly tied. The families that favor Chinese are those whose references or required output labels are Chinese strings; families scored on geometry or enumerations show no consistent direction. Qwen3.5-2B shows both patterns at once, with a 22.00-point English advantage on affordance and a 15.56-point Chinese advantage on inventory.

Dependence on acceptance. Restricting to pairs accepted in both languages shrinks the difference by model-specific amounts: Qwen3.5-9B falls from 5.98 to 2.47 points on 898 jointly accepted pairs, Qwen3-VL-8B from 2.91 to 2.14 on 1,036 pairs, and Qwen3.5-2B reverses sign, from $+1.90$ to -0.48 on 807 pairs. These are selected-subset comparisons; because acceptance is contract-dependent and changes the subset, they do not isolate formatting effects from perception.

Table 16: Paired Chinese-minus-English score differences in request-score points, with exploratory 95% bootstrap intervals over 10,000 family-stratified draws of question pairs, and the number of requests on which the Chinese form scores higher, equal, or lower.

Model	$S_{zh} - S_{en}$	95% interval	ZH win / tie / EN win
Qwen3.6-27B	+2.40	[+0.94, +3.87]	95 / 1026 / 79
Qwen3-VL-8B-Instruct	+2.91	[+1.57, +4.29]	103 / 1058 / 39
Qwen3.5-397B-A17B	+1.11	[-0.26, +2.49]	79 / 1048 / 73
Qwen3-VL-4B-Instruct	+4.35	[+2.81, +5.91]	105 / 1050 / 45
Qwen3.5-9B	+5.98	[+4.35, +7.65]	145 / 996 / 59
Embodied-R1.5-8B	+1.91	[+0.45, +3.39]	73 / 1090 / 37
RynnBrain1.1-9B	+1.60	[+0.15, +3.08]	80 / 1059 / 61
Qwen3.5-4B	+3.22	[+1.53, +4.90]	121 / 1008 / 71
InternVL3.5-8B	-3.43	[-5.65, -1.19]	110 / 934 / 156
Qwen3.5-2B	+1.90	[+0.01, +3.84]	136 / 973 / 91
UnifoLM-ER-1-4B	+1.66	[-0.65, +4.02]	168 / 913 / 119
MiMo-VL-7B-SFT	-1.06	[-2.69, +0.56]	59 / 1074 / 67
MiMo-Embodied-7B	+3.86	[+1.93, +5.86]	117 / 1011 / 72

Why this is not a cross-lingual measurement. English prompts keep the Chinese target literals the harness used, and the English pick-and-place contract still requires Chinese labels. A monolingual rewrite would no longer be the request the robot issued, so these pairs are a consistency diagnostic.