EPIC-Bench: A Perception-Centric Benchmark for Fine-Grained Embodied Visual Grounding in Vision-Language Models

Anonymous ACL submission

Abstract

While large vision-language models (VLMs) are increasingly adopted as the perceptual backbone for embodied agents, their fundamental vision-language grounding (VLG) capability in embodied scenario remains unexplored. We present **EPIC-Bench**, **E**mbodied **P**erceptI**o**n Ben**C**hmark, a fine-grained grounding benchmark designed to systematically evaluate the VLG capabilities of VLMs in real-world embodied environments. EPIC-Bench comprise 6.6k meticulously human-annotated tuples (Image, Text, Mask), spans three core stages of the embodied interaction pipeline: Target Localization, Navigation, and Manipulation. With 23 fine-grained tasks, multi-dimension evaluation settings and evaluation over 80 leading VLMs, EPIC-bench provides a diagnostic analysis.

1 Introduction

Vision-Language Models (VLMs) have emerged as a promising foundation for embodied intelligence due to their strong multimodal perception and language understanding capabilities. Recent embodied AI systems employ VLMs in different roles, including end-to-end Vision-Language-Action (VLA) policy learning (Zitkovich et al., 2023; Kim et al., 2024; Intelligence et al., 2025), high-level planning and reasoning (Team et al., 2025b; Figure, 2024), and perception modules for grounding language instructions in embodied scenes before downstream execution (Duan et al., 2024; Deng et al., 2025; Zhong et al., 2026). This modular perception-action paradigm demonstrates strong generalization potential.

However, these studies generally assume that VLMs possess sufficient vision-language priors and grounding capabilities for action execution. To examine this assumption, we conduct a VLM-based grasping experiment and identify two key observations. First, VLMs struggle with compositional semantic following: replacing simple object (e.g., *cookie box*) with compositional descriptions (e.g., *the red object pointed to by the scissors*) leads to a substantial decline in localization accuracy. Second, coarse localization degrades action execution: compared with bounding-box grounding, fine-grained mask-level localization significantly improves grasp success in cluttered scenes and for objects with irregular geometries. These findings indicate that the limited grounding capability of VLMs remains a major bottleneck for reliable embodied action execution, motivating our focus on fine-grained vision-language grounding in embodied setting.

Existing embodied perception benchmarks mainly adopt question-answering (QA) or multiple-choice (MCQ) formats, which are vulnerable to language shortcuts and emphasize recognition over localization (Chen et al., 2024). Meanwhile, traditional visual grounding benchmarks are not designed for embodied settings and offer limited diagnostic granularity (Tong et al., 2024). For example, when a model fails to ground an instruction such as *place the largest red circular object to the right of the table*, existing benchmarks rarely identify whether the failure arises from deficiencies in color discrimination, shape understanding, spatial reasoning, size comparison, or other embodied-specific grounding capabilities.

To address this gap, we introduce **EPIC-Bench**, a fine-grained grounding benchmark for embodied perception. Specifically, EPIC-Bench is designed around grounding requirements across the embodied interaction pipeline, from target object localization upon instruction, to navigation-oriented reasoning, and downstream task-oriented manipulation. It contains 6.6k meticulously human-annotated (Image, Text, Mask) tuples spanning 23 tasks for fine-grained diagnostic analysis. We adopt mask grounding as the primary evaluation protocol and complement it with three additional metrics that account for target multiplicity, ambi-

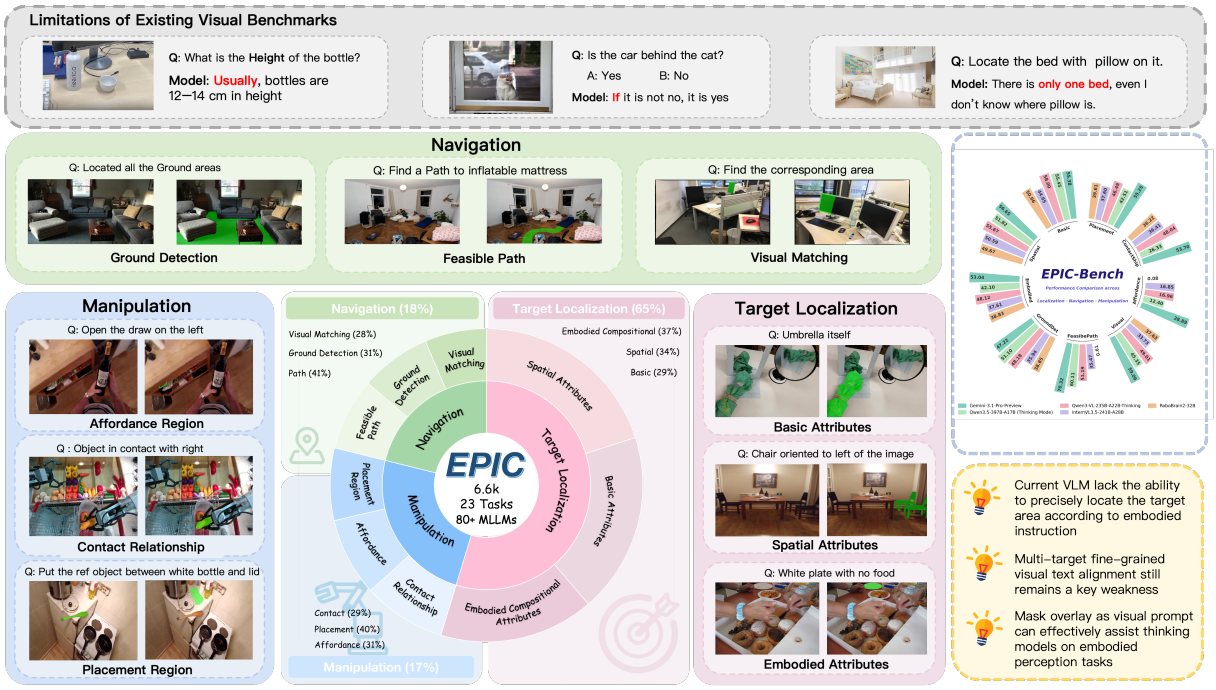


Figure 1: Overview of EPIC-Bench. Existing embodied QA and MCQ style benchmarks suffer from language shortcut and emphasize understanding rather than vision-language grounding. EPIC-Bench evaluates fine-grained embodied visual perception through mask-grounded tasks spanning target localization, navigation-oriented perception, and manipulation-oriented perception. With 6,661 human-annotated samples across 23 tasks and multi-dimensional evaluation, EPIC-Bench supports a more diagnostic analysis.

guity, and manipulation feasibility, enabling multi-dimensional evaluation. Based on EPIC-Bench, we evaluate over 80 representative VLMs and derive three key findings:

VLMs struggle with complex and fine-grained embodied grounding. Most models remain substantially below human performance under complex target settings and show notable weaknesses in affordance regions, contact relationships, part-whole reasoning, material discrimination, and orientation-related spatial grounding tasks.

Visual prompting improves fine-grained embodied grounding. Providing reference object overlays as visual prompts substantially improves fine-grained localization. However, gains vary across models and prompt types, with strong reasoning models benefiting more from mask overlays.

Reasoning improves performance, but inconsistently. While the strongest-performing models are generally reasoning-enabled, the effectiveness of reasoning varies across model scales and embodied grounding tasks. In certain cases, explicit reasoning even leads to performance degradation, suggesting that reasoning strategies must be carefully designed for embodied applications.

Overall, this work introduces EPIC-Bench, a systematic benchmark for evaluating the fine-grained vision-language grounding capabilities of VLMs. Our findings reveal that current limitations in grounding and embodied perception remain key bottlenecks for reliable embodied intelligence.

2 Related Work

In this section, we review related benchmarks from two perspectives, embodied AI benchmarks and vision-language grounding benchmarks, to further motivate our work. Table 1 summarizes representative benchmarks and highlights the distinctions between existing benchmarks and EPIC-Bench.

Vision-Language Grounding for Embodied Applications. Many embodied systems explicitly or implicitly rely on intermediate vision-language grounding (VLG). Existing works use VLMs to localize language-specified targets and support downstream planning, grasp prediction, or action generation (Stone et al., 2023; Duan et al., 2024; Deng et al., 2025; Zhong et al., 2026). Other approaches assume access to target masks and perform grasp keypoint estimation or manipulation planning (Tziafas and Kasaei, 2024; Liu et al., 2024a; Huang et al., 2024), while (Huang et al., 2025) predicts

Benchmark / Dataset	GT Type	Target Localization			Navigation			Manipulation			Data Domain	Manual	Multi View	Size
		BA	SRA	ECA	GD	FP	VM	AR	CR	PR				
EmbSpatial-Bench(Du et al., 2024)	MCQ	✓	✓								Ind.			3.6k
RoboSpatial-Home(Song et al., 2026a)	Point+Binary	✓	✓							✓	Ind.	✓		350
RefSpatial-Bench(Zhou et al., 2026)	Point/Mask	✓	✓							✓	Gen./Ind.	✓		200
OmniSpatial(Jia et al., 2026)	MCQ	✓	✓	✓†				✓			Gen./Ind.	✓		8.4k
OpenEQA (Majumdar et al., 2024)	QA	✓	✓	✓				✓			Ind./Ego.	✓	✓	1.6k
ERQA(Team et al., 2025c)	MCQ	✓	✓	✓			✓	✓	✓	✓	Ind./Ego./Robo	✓	✓	400
RefCOCO(Yu et al., 2016a)	BBox	✓	✓								Gen.	✓‡		7.6k
RefCOCOm(Wang et al., 2024)	mask	✓	✓	✓†							Gen.	✓‡		70k
FineCops-Ref(Liu et al., 2024b)	BBox	✓	✓								Gen.	✓‡		9.6k
Ref-Adv(Dong et al., 2026)	BBox	✓	✓								Gen.	✓‡		5k
EPIC-Bench	**Mask+Count**	✓	✓	✓	✓	✓	✓	✓	✓	✓	**Gen./Ind./Robo./Ego.**	✓	✓	**6.6k**

Table 1: Comparison of perception benchmarks/datasets. † denotes partial support; ‡ denotes datasets that are automatically generated with manual selection. EPIC-Bench is a manually annotated benchmark with mask-level grounding annotations.

target and placement regions for downstream VLA execution. Such modular designs may improve generalization by separating semantic grounding from physical control. Recent studies further emphasize the importance of grounding-aware representation learning (Song et al., 2026b; Zhong et al., 2026).

Embodied Benchmarks. Several benchmarks have been proposed to evaluate visual perception in embodied scenarios. OpenEQA (Majumdar et al., 2024) and ERQA (Team et al., 2025c) focus on embodied understanding and reasoning, while VSI-Bench (Yang et al., 2025), OmniSpatial (Jia et al., 2026), and EmbSpatial (Du et al., 2024) evaluate spatial reasoning capabilities. However, these benchmarks mainly adopt visual question answering or multiple-choice protocols, emphasizing scene understanding and reasoning rather than vision-language grounding. Similarly, BLINK (Fu et al., 2024) evaluates visual perception but still relies on MCQ-based evaluation. RoboSpatial (Song et al., 2026a) and RefSpatial (Zhou et al., 2026) introduce point-selection tasks for spatial reasoning and free-space perception, yet grounding remains only a limited component of their evaluation.

Visual-Language Grounding Benchmarks. Visual-language grounding is a fundamental task for evaluating vision-language alignment. RefCOCO (Yu et al., 2016a) is the most widely used grounding benchmark, while RefCOCOm (Wang et al., 2024) introduces part-level annotations for finer-grained understanding. Subsequent works such as FineCops-Ref (Liu et al., 2024b) and REF-ADV (Dong et al., 2026) further increase grounding difficulty through more detailed and challenging language descriptions. However, these benchmarks focus on generic visual scenes and do not explicitly evaluate grounding under embodied interaction constraints. EPIC-Bench addresses this gap with a finer-grained task taxonomy and evaluation protocol for diagnostic analysis of embodied grounding capabilities. Detailed comparisons are provided in Table 1.

Model		Loc.	Overall			Cond.	
			Box	Mask	GT	Box	Mask
Gemini-3	Simple	80.0	54.0	75.0	90.0	57.5	85.0
	Complex	45.5	40.0	44.0	90.0	63.0	95.0
GPT-5.2	Simple	36.0	28.0	34.0	90.0	47.4	70.6
	Complex	14.0	24.0	26.0	90.0	66.7	100

Table 2: Localization and grasp success rates under different settings, where *Cond.* denotes grasp success rate conditioned on successful localization. Results are averaged over 10 tasks per setting, with 10 trials for each task.

3 Grounding Failures in VLM-based Grasping

First, we conduct a pilot study to reveal the limitations of current VLMs in fine-grained and complex grounding for embodied tasks, thereby motivating the development of EPIC-Bench.

Experiment Setup. Specifically, we construct a grasping pipeline in PyBullet by integrating VLMs (DeepMind, 2025; OpenAI, 2025b) with an off-the-shelf grasping policy (Kumra et al., 2020). Given visual observations and language instructions, the VLM first localizes the target object, after which the predicted localization is used to guide grasp execution. Grounding quality is measured by IoU between predicted localization and ground-truth masks. Localization is considered successful

when IoU exceeds 0.25; in multi-object settings, predicted target count must additionally match the ground truth. We report both overall grasp success and conditional success under correct localization. Even when localization fails (e.g., low IoU or incorrect target count), grasp execution is still performed to analyze error propagation.

Instruction complexity. First, we vary the complexity of language instructions to evaluate grounding robustness in grasping tasks. Specifically, we replace simple category descriptions (e.g., *cracker box*) with compositional expressions (e.g., *the red object pointed to by the scissors*) while keeping the target configuration unchanged. As shown in Table 2, both localization accuracy and overall grasp success decline substantially under compositional instructions, whereas the ground-truth localization baseline remains stable. These results indicate that current VLMs still struggle with embodied grounding, and that grounding errors directly propagate to downstream manipulation performance.

Localization granularity. Next, we investigate whether fine-grained localization improves downstream execution by comparing predicted boxes with refined mask-level localization. As shown in Table 2, refined localization consistently improves both overall and conditional grasp performance, suggesting that precise grounding reduces irrelevant visual interference during grasp generation.

Overall, these findings indicate that current VLMs remain limited in fine-grained embodied grounding, while improved grounding granularity can enhance downstream grasp performance. This motivates the need for a more systematic evaluation of fine-grained grounding capabilities.

4 The EPIC-Bench

Overview. EPIC-Bench contains 6,661 manually curated annotations for mask-level embodied visual grounding, covering 3 categories, 9 subcategories, and 23 fine-grained tasks (Figure 2). It spans perception requirements from target localization to navigation and manipulation, and includes both single-image and multi-image tasks. Benchmark examples are provided in Appendix A.

4.1 Task Taxonomy

4.1.1 Target Localization (TL)

Target Localization forms the foundation of embodied navigation and manipulation by evaluating whether a model can accurately ground targets from fine-grained language descriptions. Given an image and a textual instruction, the model must predict both the locations and the number of objects satisfying the description, where the number of valid targets may range from zero to multiple instances. To systematically analyze different sources of grounding difficulty in embodied settings, we organize the tasks into three groups according to the nature of the required grounding attributes:

Basic Attributes (BA). This group evaluates grounding based on fundamental visual attributes, including **i.** object category, **ii.** color, **iii.** geometry, **iv.** material, **v.** relative attributes, and **vi.** projection. Task instances are designed to prevent target localization based solely on irrelevant attribute cues.

Spatial-Related Attributes (SRA). This group evaluates grounding based on spatial attributes, including orientation-aware spatial reasoning. Tasks are divided into conventional and human-related spatial descriptions according to the reference subject, with instructions explicitly specifying the spatial reference frame (e.g., camera-centric or subject-centric) to reduce ambiguity.

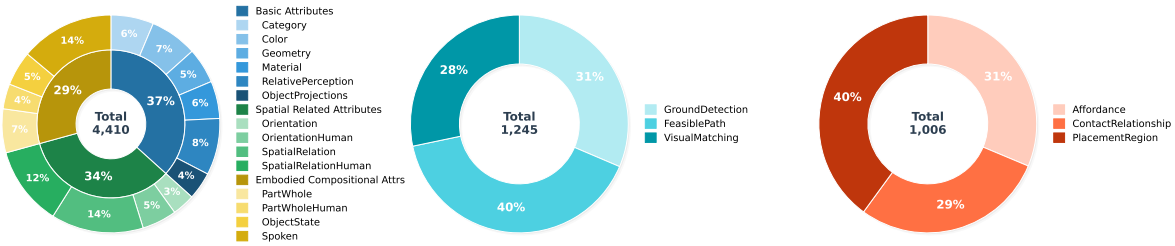
Embodied Compositional Attributes (ECA). This group evaluates attribute recognition abilities that are closely related to embodied tasks. It includes four tasks: **i.** part–whole relationships: identifying object components; **ii.** human-related part–whole relationships: localizing human anatomical regions for safe interaction; **iii.** target state differentiation: distinguishing objects based on their functional or physical states; and **iv.** spoken descriptions: grounding target objects referred to using natural everyday expressions.

4.1.2 Navigation (NAV)

This category evaluates visual perception capabilities required for navigation toward target locations. It consists of three subcategories:

Ground Detection (GD). This task evaluates a model's ability to identify traversable ground regions. Given an input image, the model is required to localize all ground surfaces for movement.

Feasible Path (FP). This task evaluates a model's ability to ground feasible navigation paths under two settings: **i.** egocentric and **ii.** exocentric. Given an image with an overlay indicating the target region, the model must generate a valid path either from the camera position to the target or between two specified targets. The predicted path must remain within traversable areas.



(a) Target Localization (b) Navigation (c) Manipulation

Figure 2: Statistics of EPIC-Bench across three primary task categories. The distribution reflects our design goal of covering both attribute-level grounding and downstream perception requirements for navigation and manipulation.

Visual Matching (VM). This task evaluates an embodied agent's ability to maintain visual correspondence across viewpoints during movement. Given multi-view images of the same environment and a target overlay in one image, the model is required to localize the same target in another view.

4.1.3 Manipulation (MAN)

This category evaluates visual perception capabilities required for embodied manipulation. It consists of three subcategories:

Affordance Region (AR). This task evaluates the ability to identify task-specific operation and tool-use regions. Given an image, a task description, and annotated reference objects, the model must localize the target operational region.

Contact Relationship (CR). This task evaluates understanding of object contact relationships, including gripper–object interactions during manipulation. Given an image and annotated reference objects, the model must localize targets under different contact conditions, including single-object contact, multi-object contact, and all objects involved in contact interactions.

Placement Region (PR). This task evaluates perception and reasoning for placement after grasping. Given an image, an overlay and description of the manipulated object, and a textual description of the placement target, the model must localize the placement region and determine whether the placement is feasible under the given context.

4.2 Benchmark Construction

We design a systematic data curation pipeline consisting of three stages: data filtering, annotation, and quality control, as illustrated in Figure 3.

Data Sources and Filtering. We collect candidate images from over 20 public datasets (Appendix B) spanning diverse domains, including generic scenes, indoor environments, egocentric views, and robot-centric imagery to ensure diversity and embodied relevance.

For the TL category, selected images must contain multiple objects with similar attributes to serve as distractors, preventing shortcut localization based solely on category cues. For NAV and MAN, we prioritize images with complex backgrounds and realistic interaction scenarios to better reflect embodied environments. We first apply Qwen3-VL-32B for preliminary filtering, followed by manual verification to ensure compliance with task-specific requirements.

Annotation Pipeline. To improve annotation efficiency while maintaining annotation quality, we incorporate SAM3 (Carion et al., 2025) into the annotation pipeline. Annotators first select task-relevant images and create corresponding textual descriptions, after which SAM3 generates initial mask proposals. These proposals are then manually refined and corrected to produce high-fidelity mask-level annotations. For certain task categories, we further collect auxiliary labels, including target counts and binary annotations such as precise-count indicators and placement feasibility judgments.

Quality Control. To ensure annotation quality and consistency, we conduct at least two rounds of inspection to revise ambiguous descriptions, inconsistent annotations, and shortcut-prone samples. Through this rigorous curation process, EPIC-Bench provides diverse, high-quality annotations for embodied perception tasks.

5 Empirical analysis

Evaluation Setup. To systematically evaluate embodied perception, we benchmark over 80 VLMs on EPIC-Bench, including proprietary, open-source, and embodied foundation models. Results for locally deployed models are averaged over six runs, while proprietary models are evaluated

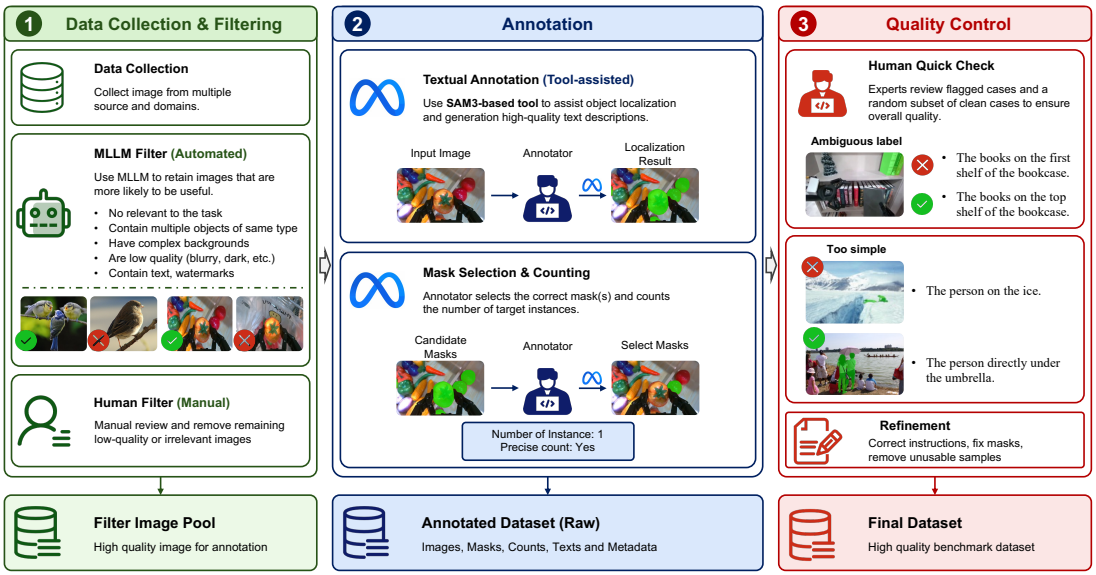


Figure 3: **Data construction pipeline of EPIC-Bench.** Section 4.1.3 describes each step in detail. The key steps include: (1) Select images with distractor instances or complex backgrounds, (2) Annotate using SAM3-assisted segmentation tool with manual refinement, (3) Eliminate or revise ambiguous or overly simple samples.

over two runs due to their high empirical stability. Prompt templates are provided in Appendix C.

Evaluation Metrics. Our evaluation framework adopts task-specific metrics. As summarized in Table 4, the overall score consists of Localization, Counting, Path, and Feasibility scores, with each task using an appropriate subset. Sample scores are computed as weighted averages of selected metrics, and the final benchmark score is averaged over all samples. Details are provided below.

Localization Score. This metric evaluates a model's grounding precision. Although our framework supports fine-grained mask-level evaluation, current VLMs typically produce bounding-box outputs. We therefore prompt models to predict Bboxes and convert each prediction into a solid rectangular mask for IoU computation against ground-truth (GT) masks.

Counting Score. This metric evaluates vision-language alignment through target counting. Let P and G denote the predicted and ground-truth target counts, respectively. The scoring strategy is determined by a binary *precise count* label. If exact counting is required (True), we use binary accuracy: $S_{\text{count}} = \begin{cases} 1 & P = G \\ 0 & \text{otherwise} \end{cases}$.

Otherwise (False), we apply a soft penalty based on counting error: $S_{\text{count}} = \max\left(1 - \frac{|P-G|}{G}, 0\right)$

Path Score. This metric evaluates a model's ability to generate physically feasible navigation paths. Models are required to output at least three consecutive trajectory points constrained within traversable regions. Detailed scoring formulations are provided in Appendix D.

Feasibility Score. This metric evaluates a model's ability to judge the feasibility of embodied actions in a binary format. For example, in placement region tasks, the model determines whether a proposed action is feasible by reasoning over visual cues such as object shape, size, and stability.

Reasoning Models Lead, but Differ in Strengths. Table 3 presents results for representative models, with the full leaderboard provided in the Appendix. Reasoning-enabled VLMs generally dominate the leaderboard, though their advantages vary across categories. For example, Qwen3-VL-235B-Thinking performs better on conventional target localization, while Gemini-3-Pro achieves stronger performance on embodied grounding tasks, suggesting different strengths in grounding and embodied perception. Existing embodied foundation models show no clear grounding advantage, and some exhibit weaker instruction-following ability, potentially due to smaller parameter budgets or optimization trade-offs that prioritize other objectives over fine-grained grounding.

Beyond the main benchmark results, we further analyze embodied vision-language grounding from three perspectives to understand its key limitations.

Model	Average	Target Local.			Navigation			Manipulation		
		BA	SRA	ECA	GD	FP	VM	AR	CR	PR.
Proprietary Models										
Gemini-3-Pro†	**54.81**	56.81	54.50	**53.90**	48.82	**70.82**	**60.08**	**30.40**	**51.78**	**53.42**
Claude-Sonnet-4.6†	43.24	45.43	45.49	36.47	40.16	68.46	42.77	9.790	46.21	43.81
o3†	36.05	39.37	40.96	32.86	36.24	30.64	40.73	5.49	32.43	43.64
GPT-5.5	50.16	**57.78**	**55.37**	45.19	**50.97**	38.57	51.65	24.62	49.67	48.70
Doubao-Seed-1.8†	46.32	55.29	49.68	46.10	46.02	53.43	45.90	10.94	14.36	40.79
Qwen3.6-Plus†	45.48	47.66	50.17	41.07	44.47	62.12	41.57	11.95	47.84	41.59
HunYuan-T1-Vision†	43.00	51.24	47.60	39.67	42.63	50.08	33.54	6.630	38.95	34.49
Open-source Models										
Qwen3-VL-235B-A22B†	**50.93**	**58.00**	**55.67**	**48.12**	**48.18**	**51.19**	**49.01**	**16.96**	**48.04**	**46.49**
InternVL3.5-241B-A28B	40.75	44.65	50.59	37.61	35.94	35.47	33.73	16.85	36.41	37.60
MiMo-VL-7B-RL-2508	34.65	36.65	42.31	34.92	18.31	36.90	37.45	6.510	30.37	32.86
Gemma-3-27B-IT	27.17	26.93	33.66	26.68	11.62	32.17	26.33	1.650	31.26	32.18
GLM-4.6V†	42.84	50.19	47.71	37.89	42.46	51.11	36.55	8.320	34.14	39.79
Step3-VL-10B	32.40	42.22	40.10	31.88	17.96	24.03	17.77	5.940	27.22	27.56
LLaVA-NeXT-72B	20.40	26.54	29.42	19.93	14.18	6.340	6.190	2.990	12.75	18.63
Embodied Foundation Models										
Pelican1.0-VL-72B	35.29	44.08	42.11	34.79	**44.91**	0.000	**38.13**	**7.320**	22.68	37.96
RoboBrain2-32B	**39.32**	**50.66**	**49.67**	**38.83**	38.85	0.610	37.63	0.080	**36.22**	**39.61**
RoboBrain2.5-8B-MT	29.83	31.20	37.41	26.76	33.52	**30.89**	27.84	2.650	25.45	24.34
RynnBrain-Plan-8B	23.59	23.53	32.53	26.08	2.120	19.57	21.41	0.110	18.79	31.02
VeBrain	29.27	33.63	35.98	27.69	29.56	20.04	27.58	2.850	19.25	32.30
Cosmos-Reason1-7B	23.25	29.74	26.64	19.67	24.68	3.030	25.44	5.370	22.12	32.75

Table 3: Overall performance of representative VLMs on EPIC-Bench. **Green bold** indicates the overall best result across all models. **Bold** and underline indicate the best and second-best results within each model category, respectively. † denotes models evaluated with thinking mode enabled. Comprehensive results for all evaluated models are provided in Tab. 7 of Appendix. E.

Task Type	Score Composition
TL-All	$(1-\alpha)$ Localization Score + α Counting Score
NAV-Ground Detection	Localization Score
NAV-Feasible Path	Path Score
NAV-Visual Matching	$(1-\alpha)$ Localization Score + α Counting Score
MAN-Affordance Region	Localization Score
MAN-Contact Relationship	$(1-\alpha)$ Localization Score + α Counting Score
MAN-Placement Region	$(1-\alpha)$ Localization Score + α Feasibility Score

Table 4: Task-specific score compositions. The weighting factor α defaults to 0.5. Exceptions: (1) For TL, $\alpha = 0.3$ if the GT precise count label is False or $\alpha = 0.1$ if the textual instruction contains explicit numerical information. (2) For MAN-CR, $\alpha = 0.3$ if the GT precise count label is False.

Target Number Generalization Remains Limited. Preliminary analysis shows that models frequently struggle to predict the correct number of targets. We therefore evaluate counting performance using samples with Precise Count label. As shown in Figure 4, models achieve near-human accuracy for single-target cases, but performance drops substantially in zero-target and multi-target settings. We hypothesize that this limitation contributes to lower localization performance. One possible explanation is the bias of existing grounding datasets, where most samples assume one-to-one correspondence between text and targets, encouraging hallucinated or incomplete predictions.

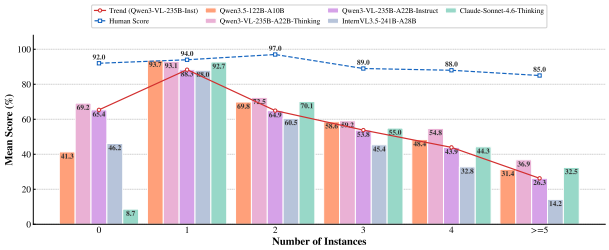


Figure 4: Counting accuracy across different target counts on Contact Relationship and Target Localization tasks. Model performance drops substantially in zero-target and multi-target settings compared with single-target cases.

Fine-Grained Grounding Remains Challenging. In addition, models consistently underperform on affordance-region and part–whole tasks. Case study in Figure 5 suggests that models favour object-level localization, whereas these tasks require part-level grounding. Although models perform relatively well on standard spatial relations, performance drops noticeably on orientation-

sensitive grounding, indicating limited understanding of viewpoint-dependent spatial cues.

Figure 5: Case studies on Part-Whole and Affordance Region tasks. Ground-truth targets are shown as red masks, while model predictions are visualized as green boxes. Gemini-3.1-Pro demonstrates qualitatively stronger grounding performance than Qwen3-VL-235B, yet still exhibits incomplete and inaccurate localization in challenging cases.

Next, we conduct ablation studies to evaluate the effectiveness of overlay-based visual prompts for embodied grounding.

Coarse Visual Prompts Improve Fine-Grained Grounding. EPIC-Bench introduces mask overlays in selected navigation and manipulation tasks. As shown in Figure 6, both bbox and mask overlays improve performance over the no-overlay setting, highlighting the effectiveness of visual prompting for embodied grounding. Reasoning models benefit more from mask overlays, while weaker models tend to perform better with simpler bbox overlays.

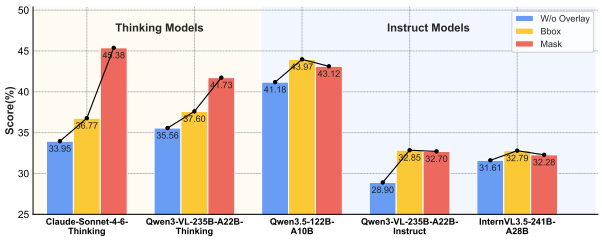
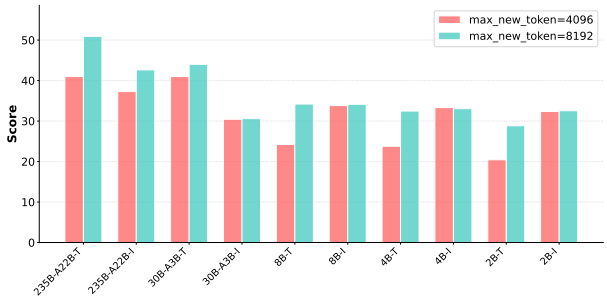
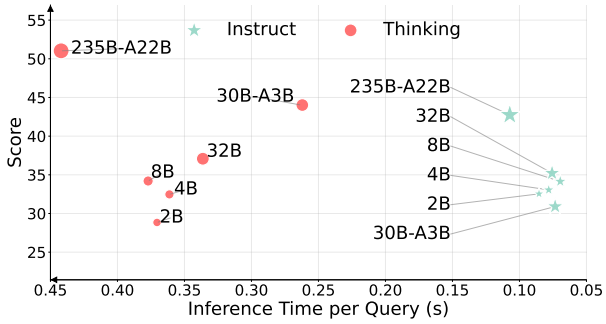


Figure 6: Effect of visual prompting on embodied grounding. Introducing visual overlays consistently improves performance, with stronger reasoning models benefiting more from mask overlays and smaller models preferring simpler bbox prompts.

Impact of Reasoning Length Constraints. Finally, we analyze the Qwen3-VL series under different maximum answer lengths to study reasoning efficiency. As shown in Figure 7a, reasoning models generally outperform or match non-reasoning counterparts when generation budgets are sufficient, with larger models benefiting the most. Under restricted output lengths, however, smaller reasoning models may underperform due to repetitive reasoning loops that exhaust the generation budget, whereas larger models remain more robust to truncation.



(a) Impact of maximum new token length on model performance. Suffix T denotes Thinking, and I denotes Instruct.



(b) Performance-Efficiency trade-off.

Figure 7: Efficiency assessment of Qwen3-VL model family.

Performance-Efficiency Trade-off. Figure 7b illustrates the relationship between average inference time and benchmark performance (8×A100 GPUs), providing practical guidance for embodied system deployment. Although increasing model scale and extending reasoning length can improve performance, both introduce additional inference costs, particularly in latency. Consequently, larger reasoning models are not always the optimal choice in practice. Instead, selecting an appropriate balance between reasoning capability and computational efficiency may lead to more favorable deployment trade-offs.

6 Conclusion

We introduce EPIC-Bench, a benchmark for systematically evaluating embodied vision-language grounding in VLMs and addressing limitations of existing evaluation protocols. Through comprehensive evaluation and analysis across representative models, we reveal key challenges in embodied grounding and provide insights for future model development and embodied applications.

7 Limitations

EPIC-Bench is built upon mask-level annotations and supports fine-grained grounding evaluation. However, the primary focus of this work is to evaluate the native vision-language grounding (VLG) capability of current VLMs under a unified output protocol. Since most existing VLMs do not natively support mask prediction, we adopt bounding-box prediction as the evaluation interface to ensure fair comparison and broad compatibility across models. Consequently, the reported results do not reflect the performance of models with native mask outputs or composite systems that combine VLMs with external segmentation modules (e.g., VLM+SAM pipelines). Extending evaluation to native mask grounding and more diverse grounding interfaces remains an important direction for future work.

In addition, EPIC-Bench focuses on 2D vision-language grounding. Although 2D observations remain the most widely adopted visual input format and provide rich semantic information for embodied tasks, 3D scene representations offer additional spatial and geometric cues that may benefit embodied grounding. Evaluating grounding ability under 3D and multi-view settings is beyond the current scope of EPIC-Bench and will be explored in future extensions.

Finally, the conclusions presented in this work are primarily derived from behavioral evaluation results. While benchmark outcomes reveal capability differences and common failure patterns, they do not fully explain the underlying causes. Future work may further investigate embodied grounding through internal representation analysis, attention and reasoning visualization, and mechanistic understanding of VLM decision processes.

References

AgiBot-World-Contributors, Qingwen Bu, Jisong Cai, and et al. 2025. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *Preprint*, arXiv:2503.06669.

Anthropic. 2025a. Introducing claude 4. https://www.anthropic.com/news/claude-4.

Anthropic. 2025b. Introducing claude haiku 4.5. https://www.anthropic.com/news/claude-haiku-4-5.

Anthropic. 2026. Introducing claude sonnet 4.6. https://www.anthropic.com/news/claude-sonnet-4-6.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, and et al. 2025a. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, and et al. 2025b. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, and et al. 2021. ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. 2013. OpenSurfaces: A richly annotated catalog of surface appearance. *ACM Trans. on Graphics (SIGGRAPH)*, 32(4).

Bytedance Seed. 2025a. Seed1.6 tech introduction. https://seed.bytedance.com/en/seed1_6.

Bytedance Seed. 2025b. Seed1.8 model card:towards generalized real-world agency. https://lf3-static.bytednsdoc.com/obj/eden-cn/lapzild-tss/ljhwZthlaukjlkulzlp/research/Seed-1.8-Modelcard.pdf.

Mingyu Cao, Huajie Tan, Yuheng Ji, and et al. 2025. Robobrain 2.0 technical report. *arXiv preprint arXiv:2507.02029*.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, and 1 others. 2025. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*.

Angel Chang, Angela Dai, Thomas Funkhouser, and et al. 2017. Matterport3d: Learning from rgb-d data in indoor environments. *Preprint*, arXiv:1709.06158.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, and et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Angela Dai, Angel X. Chang, Manolis Savva, and et al. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. *Preprint*, arXiv:1702.04405.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, and et al. 2018. Scaling egocentric vision: The epic-kitchens dataset. *Preprint*, arXiv:1804.02748.

Ronghao Dang, Jiayan Guo, Bohan Hou, Sicong Leng, Kehan Li, Xin Li, Jiangpin Liu, Yunxuan Mao, Zhikai Wang, Yuqian Yuan, and 1 others. 2026. Rynnbrain: Open embodied foundation models. *arXiv preprint arXiv:2602.14979*.

Sudeep Dasari, Frederik Ebert, Stephen Tian, and et al. 2020. Robonet: Large-scale multi-robot learning. *Preprint*, arXiv:1910.11215.

Google DeepMind. 2025. Gemini 3: A new era of intelligence. https://blog.google/products-and-platforms/products/gemini/gemini-3/.

Shengliang Deng, Mi Yan, Songlin Wei, Haixin Ma, Yuxin Yang, Jiayi Chen, Zhiqi Zhang, Taoyu Yang, Xuheng Zhang, Wenhao Zhang, and 1 others. 2025. Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data. *arXiv preprint arXiv:2505.03233*.

Qihua Dong, Kuo Yang, Lin Ju, Handong Zhao, Yitian Zhang, Yizhou Wang, Huimin Zeng, Jianglin Lu, and Yun Fu. 2026. Ref-adv: Exploring mllm visual reasoning in referring expression tasks. *arXiv preprint arXiv:2602.23898*.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. 2024. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. *Preprint*, arXiv:2406.05756.

Jiafei Duan, Wentao Yuan, Wilbert Pumacay, Yi Ru Wang, Kiana Ehsani, Dieter Fox, and Ranjay Krishna. 2024. Manipulate-anything: Automating real-world robots using vision-language models. *arXiv preprint arXiv:2406.18915*.

AI Figure. 2024. Helix: A vision-language-action model for generalist humanoid control. *Figure AI News*.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer.

GLM Team. 2025. Glm-4.6v: Open source multimodal models with native tool use. https://z.ai/blog/glm-4.6v.

Agrim Gupta, Piotr Dollar, and Ross Girshick. 2019. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

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, and 25 others. 2026. Mimo-embodied: X-embodied foundation model technical report. *Preprint*, arXiv:2511.16518.

Ailin Huang, Chengyuan Yao, Chunrui Han, and et al. 2026. Step3-vl-10b technical report. *Preprint*, arXiv:2601.09668.

Haifeng Huang, Xinyi Chen, Yilun Chen, Hao Li, Xiaoshen Han, Zehan Wang, Tai Wang, Jiangmiao Pang, and Zhou Zhao. 2025. Roboground: Robotic manipulation with grounded vision-language priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22540–22550.

Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Li Fei-Fei. 2024. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. *arXiv preprint arXiv:2409.01652*.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Preprint*, arXiv:1902.09506.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, and 1 others. 2025. pi0.5: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. 2026. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. *Preprint*, arXiv:2506.03135.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, and et al. 2016. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *Preprint*, arXiv:1612.06890.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, and 1 others. 2024. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, and et al. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Preprint*, arXiv:1602.07332.

Sulabh Kumra, Shirin Joshi, and Ferat Sahin. 2020. Antipodal robotic grasping using generative residual convolutional neural network. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9626–9633. IEEE.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024. Llava-next: Stronger llms supercharge multimodal capabilities in the wild.

Yin Li, Miao Liu, and James M Rehg. 2018. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European conference on computer vision (ECCV)*, pages 619–635.

Fangchen Liu, Kuan Fang, Pieter Abbeel, and Sergey Levine. 2024a. Moka: Open-world robotic manipulation through mark-based visual prompting. *arXiv preprint arXiv:2403.03174*.

Junzhuo Liu, Xuzheng Yang, Weiwei Li, and Peng Wang. 2024b. Finecops-ref: A new dataset and task for fine-grained compositional referring expression comprehension. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15440–15457.

Gen Luo, Ganlin Yang, Ziyang Gong, Guanzhou Chen, Haonan Duan, Erfei Cui, Ronglei Tong, Zhi Hou, Tianyi Zhang, Zhe Chen, Shenglong Ye, Lewei Lu, Jingbo Wang, Wenhai Wang, Jifeng Dai, Yu Qiao, Rongrong Ji, and Xizhou Zhu. 2025. Visual embodied brain: Let multimodal large language models see, think, and control in spaces. *Preprint*, arXiv:2506.00123.

Teli Ma, Jia Zheng, Zifan Wang, and et al. 2025. Glover++: Unleashing the potential of affordance learning from human behaviors for robotic manipulation. *arXiv preprint arXiv:2505.11865*.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, and et al. 2024. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16488–16498.

NVIDIA, :, Alisson Azzolini, Junjie Bai, and et al. 2025. Cosmos-reason1: From physical common sense to embodied reasoning. *Preprint*, arXiv:2503.15558.

OpenAI, :, Aaron Hurst, Adam Lerer, and et al. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2025a. Gpt-5.1: A smarter, more conversational chatgpt. https://openai.com/index/gpt-5-1/.

OpenAI. 2025b. Introducing gpt-5.2. https://openai.com/index/introducing-gpt-5-2/.

OpenAI. 2025c. Openai o3 and o4-mini system card. https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf.

OpenAI. 2026a. Introducing gpt-5.4. https://openai.com/index/introducing-gpt-5-4/.

OpenAI. 2026b. Introducing gpt-5.5. https://openai.com/index/introducing-gpt-5-5/.

QwenTeam. 2026. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5.

QwenTeam. 2026. Qwen3.6-27B: Flagship-Level Coding in a 27B Dense Model. https://qwen.ai/blog?id=qwen3.6-27b.

QwenTeam. 2026a. Qwen3.6-35b-a3b: Agentic coding power, now open to all. https://qwen.ai/blog?id=qwen3.6-35b-a3b.

QwenTeam. 2026b. Qwen3.6-plus: Towards real world agents. https://qwen.ai/blog?id=qwen3.6.

Pierre Sermanet, Tianli Ding, Jeffrey Zhao, and et al. 2023. Robovqa: Multimodal long-horizon reasoning for robotics. In *arXiv preprint arXiv:2311.00899*.

Aaditya Singh, Adam Fry, Adam Perelman, and et al. 2025. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. 2026a. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. *Preprint*, arXiv:2411.16537.

Wenxuan Song, Ziyang Zhou, Han Zhao, Jiayi Chen, Pengxiang Ding, Haodong Yan, Yuxin Huang, Feilong Tang, Donglin Wang, and Haoang Li. 2026b. Reconvla: Reconstructive vision-language-action model as effective robot perceiver. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 18549–18557.

Austin Stone, Ted Xiao, Yao Lu, Keerthana Gopalakrishnan, Kuang-Huei Lee, Quan Vuong, Paul Wohlhart, Sean Kirmani, Brianna Zitkovich, Fei Xia, and 1 others. 2023. Open-world object manipulation using pre-trained vision-language models. *arXiv preprint arXiv:2303.00905*.

Markus Suchi, Timothy Patten, and David Fischinger snd et al. 2019. Easylabel: A semi-automatic pixel-wise object annotation tool for creating robotic RGB-D datasets. In *International Conference on Robotics and Automation, ICRA 2019, Montreal, QC, Canada, May 20-24, 2019*, pages 6678–6684.

Huajie Tan, Enshen Zhou, Zhiyu Li, and et al. 2026. Robobrain 2.5: Depth in sight, time in mind. *Preprint*, arXiv:2601.14352.

Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, Kainan Bao, Hao Tian, Hailin Zhang, Gang Wang, Dawei Zhu, Cici, Chenhong He, Bowen Ye, and 55 others. 2025a. Mimo-vl technical report. *Preprint*, arXiv:2506.03569.

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, and 1 others. 2025b. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*.

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, and et al. 2025c. Gemini robotics: Bringing ai into the physical world. *Preprint*, arXiv:2503.20020.

Gemma Team, Aishwarya Kamath, Johan Ferret, and et al. 2025d. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Tencent Hunyuan Team. 2025. Hunyuanvision. https://github.com/Tencent-Hunyuan/HunyuanVision.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9568–9578.

Georgios Tziafas and Hamidreza Kasaei. 2024. Towards open-world grasping with large vision-language models. *arXiv preprint arXiv:2406.18722*.

Johanna Wald, Armen Avetisyan, Nassir Navab, and et al. 2019. Rio: 3d object instance re-localization in changing indoor environments. In *Proceedings IEEE International Conference on Computer Vision (ICCV)*.

Homer Walke, Kevin Black, Abraham Lee, and et al. 2024. Bridgedata v2: A dataset for robot learning at scale. *Preprint*, arXiv:2308.12952.

Weiyun Wang, Zhangwei Gao, Lixin Gu, and et al. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *Preprint*, arXiv:2508.18265.

Wenxuan Wang, Tongtian Yue, Yisi Zhang, Longteng Guo, Xingjian He, Xinlong Wang, and Jing Liu. 2024. Unveiling parts beyond objects: Towards finer-granularity referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12998–13008.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2025. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643.

Xin Yang, Haiyang Mei, Ke Xu, and et al. 2019. Where is my mirror? *Preprint*, arXiv:1908.09101.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016a. Modeling context in referring expressions. *Preprint*, arXiv:1608.00272.

Licheng Yu, Patrick Poirson, Shan Yang, and et al. 2016b. Modeling context in referring expressions. *Preprint*, arXiv:1608.00272.

Yi Zhang, Che Liu, Xiancong Ren, and et al. 2025. Pelican-vl 1.0: A foundation brain model for embodied intelligence. *Preprint*, arXiv:2511.00108.

Yifan Zhong, Xuchuan Huang, Ruochong Li, Ceyao Zhang, Zhang Chen, Tianrui Guan, Fanlian Zeng, Ka Nam Lui, Yuyao Ye, Yitao Liang, and 1 others. 2026. Dexgraspvla: A vision-language-action framework towards general dexterous grasping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 18836–18844.

Bolei Zhou, Hang Zhao, Xavier Puig, and et al. 2018. Semantic understanding of scenes through the ade20k dataset. *Preprint*, arXiv:1608.05442.

Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, and Shanghang Zhang. 2026. Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. *Preprint*, arXiv:2506.04308.

Jinguo Zhu, Weiyun Wang, Zhe Chen, and et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, and 1 others. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR.

A Benchmark Examples

In this section, we present detailed examples from EPIC-Bench to illustrate the input–output content required by tasks across different categories. For certain tasks, we also provide optional auxiliary annotations. These additional labels enable more accurate evaluation by accounting for the specific characteristics of different tasks and data types.

Figure 8, 9 and 10 present examples from the Target Localization category, including the three task types of Basic Attributes, Spatial-Related Attributes, and Embodied Compositional Attributes.

Figure 11 illustrates the Navigation category, which includes the tasks of Ground Detection, Feasible Path Recognition, and Visual Matching.

Figure 12 and 13 present examples from the Manipulation category, covering the tasks of Affordance Region, Contact Relationship, and Placement Region.

B Data Sources

In this section, we provide a detailed summary of the source datasets used in EPIC-Bench. As shown

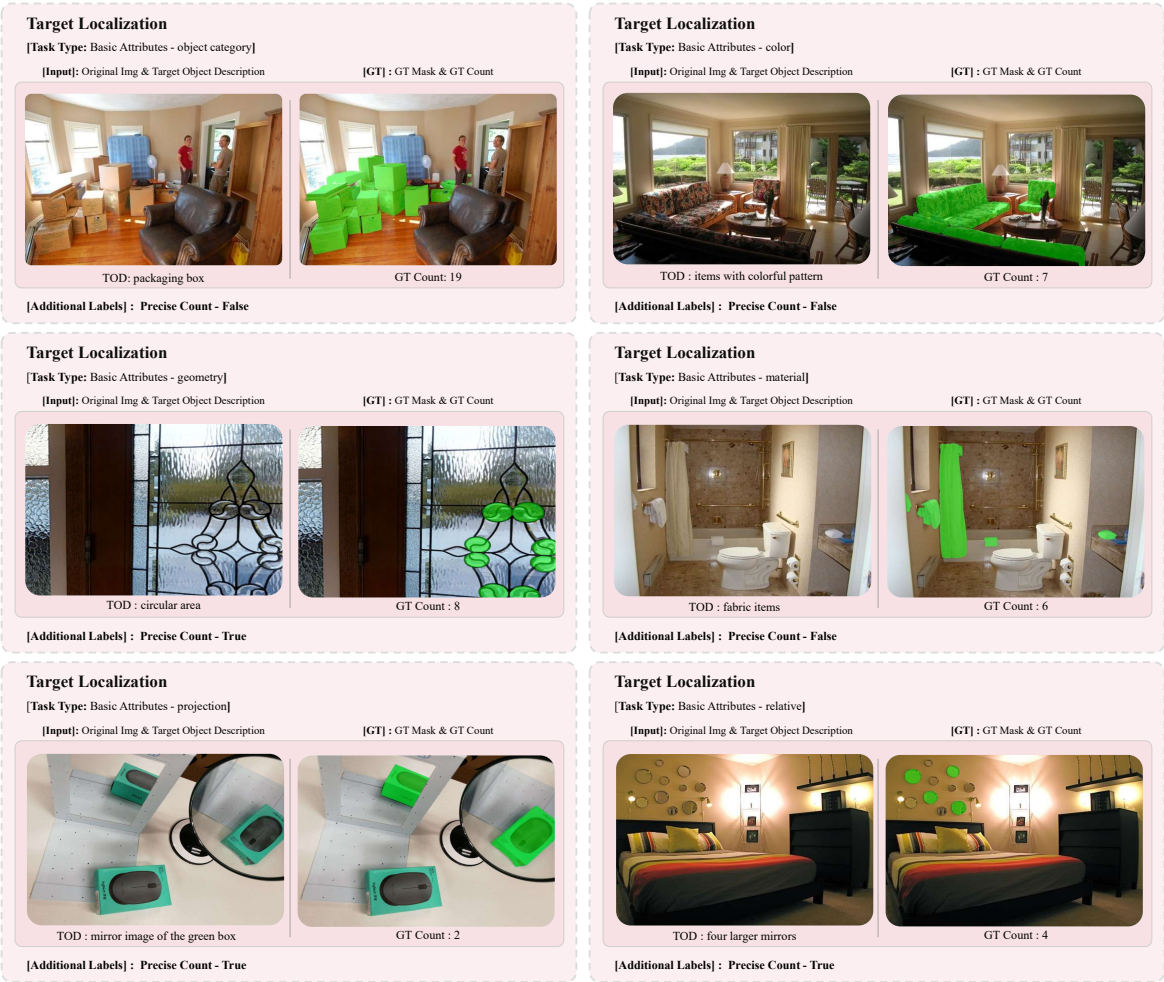


Figure 8: Examples of Target Localization - Basic Attributes task.

in Table ??, we curate images from 25 publicly available datasets spanning four domains: general scenes, indoor environments, robotics, and egocentric views, thereby ensuring the diversity of our dataset.

We obtain all existing datasets from their original sources released by the authors. We refer readers to these sources for the dataset licenses.

C Prompt Templates for Each Sub-task

C.1 Prompt

In this section, we present the prompt templates for each sub-task. We design task-specific prompts for the categories of Target Localization (shown in Figure 14), Navigation (Ground Detection, Feasible Path, Visual Matching) (shown in Figure 15), and Manipulation (Affordance Region, Contact Relationship, Placement Region) categories (shown in Figure 16).

C.2 Response format

In this section, we present the response format specifications. As illustrated in Figure 17, we design structured JSON output formats for each sub-task to ensure consistent and parseable model responses.

D Path Score Definition

In this section, we describe the detailed design of the **Path Score**, which is used to evaluate the quality of predicted navigation paths. The score consists of five components: **Start-End Score**, **Traversable Path Ratio Score**, **Away-From-Start-Area Score**, **Approaching-Goal-Area Score** and **Continuity Score**.

Given a sequence of points predicted by the model, we first apply a **Neighbor Filtering** step to remove redundant points. If the distance between two consecutive points is smaller than $\frac{1}{30}$ of the image diagonal length, the two points are merged. After filtering, each remaining point is assigned a weighted score based on the five components

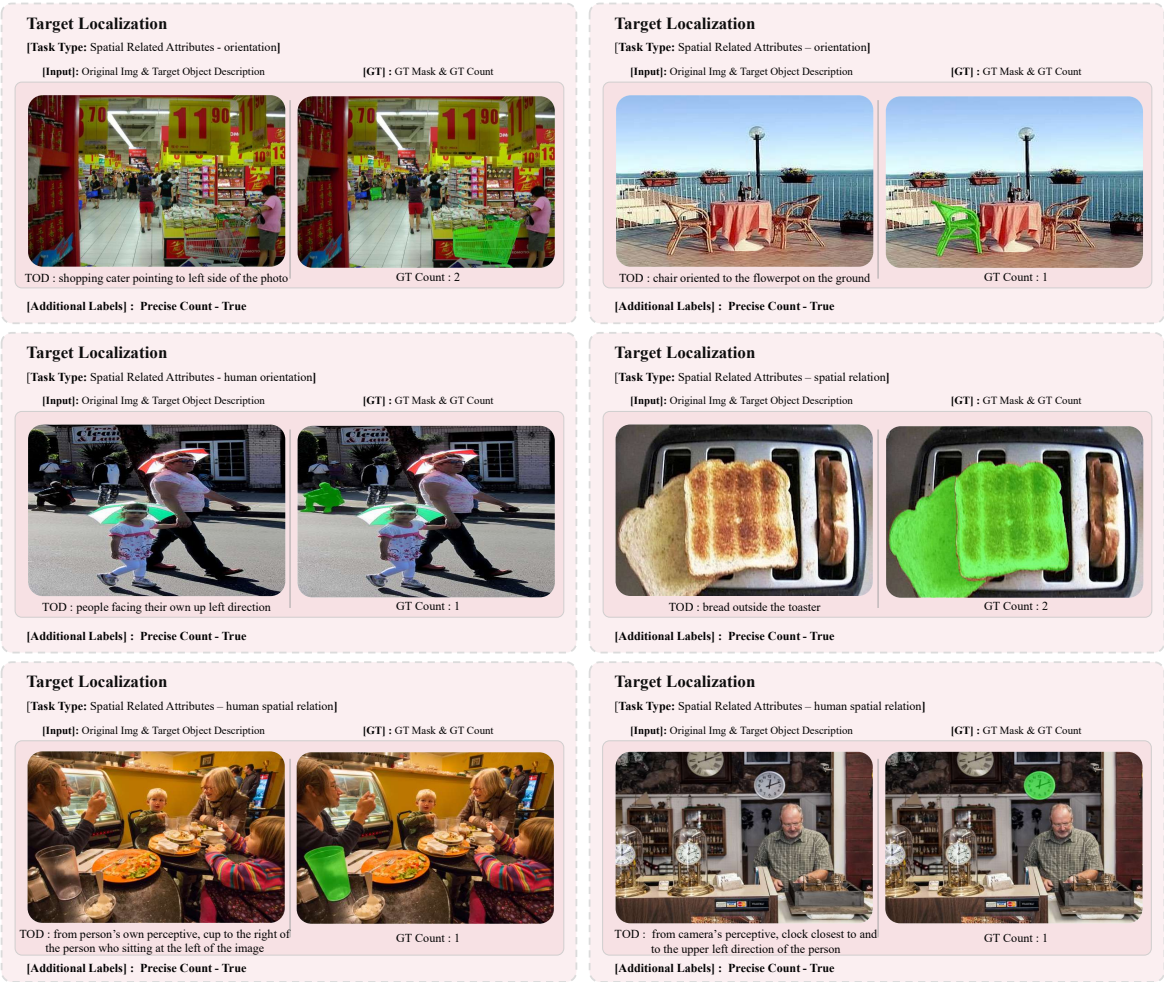


Figure 9: Examples of Target Localization - Spatial Related Attributes task

Task	Benchmark Dataset Source
General	Visual Genome (Krishna et al., 2016), GQA (Hudson and Manning, 2019), ADE20K (Zhou et al., 2018), MSD (Yang et al., 2019) RefCOCO (Yu et al., 2016b), OmniSpatial(Jia et al., 2026), LVIS (Gupta et al., 2019), CLEVR (Johnson et al., 2016)
Indoor	OpenSurfaces (Bell et al., 2013), ARKitScenes (Baruch et al., 2021), ScanNet (Dai et al., 2017), MP3D(Chang et al., 2017), OCID (Suchi et al., 2019) 3RScan (Wald et al., 2019)
Robotic	Agibot-World (AgiBot-World-Contributors et al., 2025), RoboVQA (Sermanet et al., 2023), BridgeData V2 (Walke et al., 2024), RoboNet (Dasari et al., 2020)
Egocentric	EPIC-KITCHENS (Damen et al., 2018), RoboVQA (Sermanet et al., 2023) HOVA-500K (Ma et al., 2025), EGTEA Gaze+ (Li et al., 2018)

Table 5: Dataset source summary in Epic-Bench.

described below.

Start-End Score. The Start-End Score measures how well the predicted path aligns with the target start and end areas. Specifically, we compute the distance between the first predicted point and the start area, as well as the distance between the last predicted point and the goal area $Dis(S, D)$. The distances are normalized by the image diagonal length to ensure scale invariance.

$$\text{Start-End Score} = \max(0, 1 - k \cdot D_{\text{norm}}), \quad (1)$$

$$D_{\text{norm}} = \frac{\text{Distance}}{\text{Image Diagonal Length}}, \quad (2)$$

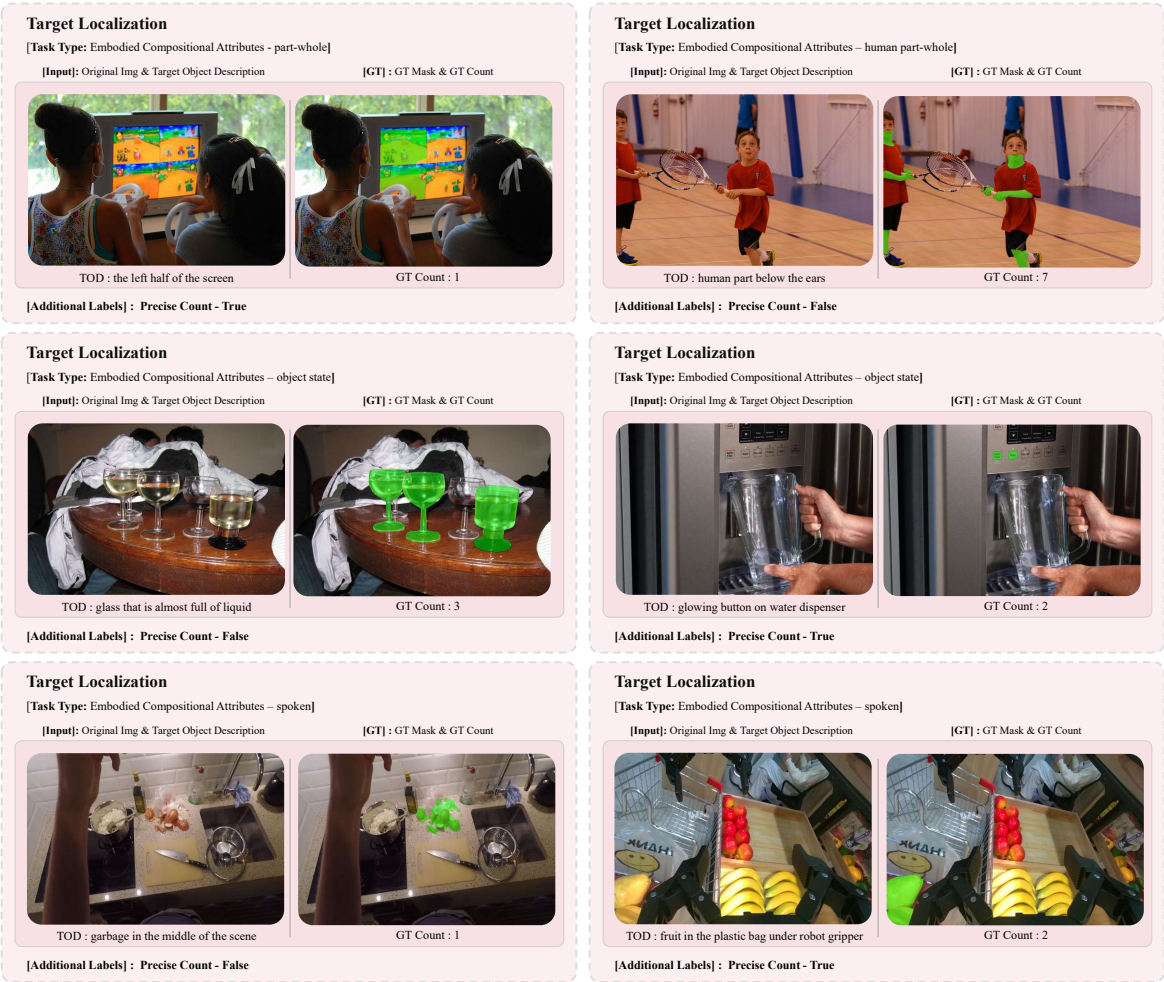


Figure 10: Examples of Target Localization - Embodied Compositional Attributes task

where k is a hyperparameter (default $k = 3.0$).

Traversable Path Ratio Score. The Traversable Path Ratio Score encourages the model to generate paths that lie within regions annotated as feasible and traversable by human annotators. Starting from the second point, each point is connected to its preceding point to form a line segment l_i. The score is determined by the proportion of the segment that lies inside the ground-truth feasible-path mask M_{fp}.

$$\text{Traversable Path Ratio Score}_i = \frac{\text{Length}(l_i \cap M_{fp})}{\text{Length}(l_i)}, \tag{3}$$

where l_i denotes the line segment connecting points p_{i-1} and p_i, and M_{fp} denotes the ground-truth feasible-path Mask.

Away-From-Start-Area Score and **Approaching-Goal-Area Score.** The two scores jointly encourage the model to generate more reasonable navigation trajectories, especially in complex scenes where detours may be necessary. If the evaluation only considers whether the path approaches the destination, the model may favor trivial straight-line predictions. To address this issue, we additionally reward paths that progressively move away from the start area while approaching the goal area.

For each point p_i ($i \geq 2$), we compare its relative distances to the start and goal areas with those of its preceding point p_{i-1}. The **Away-From-Start-Area Score** is assigned in a binary manner: a point receives a score of 1 if it is farther from the start area than its predecessor, and 0 otherwise. Similarly, the **Approaching-Goal-Area Score** is also binary: a point receives a score of 1 if it is closer to the goal area than its predecessor, and 0 otherwise. Let A_i denote the Away-From-Start-Area score, and B_i denote the Approaching-Goal-Area score.

$$A_i = \begin{cases} 1, & d(p_i, S) > d(p_{i-1}, S), \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

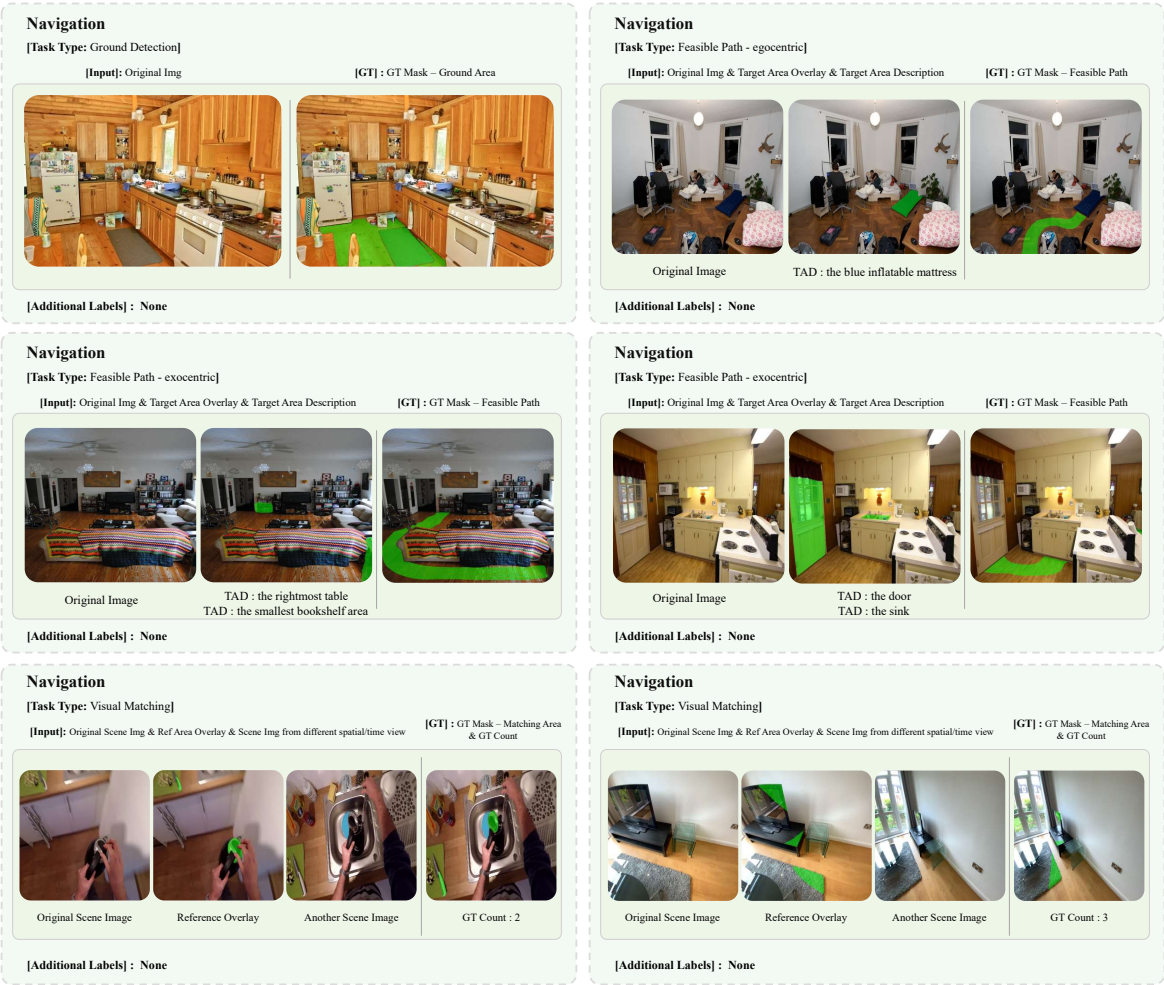


Figure 11: Examples of Navigation tasks

$$B_i = \begin{cases} 1, & d(p_i, G) < d(p_{i-1}, G), \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

where $d(p_i, S)$ denotes the shortest distance between point p_i and start area S, and $d(p_i, G)$ denotes the shortest distance between point p_i and goal area G.

Continuity Score. The continuity Score evaluates whether consecutive points form a coherent and realistic navigation path. Extremely large jumps between consecutive points often indicate unreliable or unrealistic predictions. Therefore, we penalize points whose distance to the preceding point exceeds a predefined threshold.

For each point p_i ($i \geq 2$), the Continuity Score is assigned in a binary manner. If the distance between two consecutive points exceeds a threshold proportional to the start–goal distance, the point is considered discontinuous and receives a score of 0; otherwise, it receives a score of 1.

$$\text{Continuity Score}_i = \begin{cases} 1, & \text{if } d(p_i, p_{i-1}) \leq \frac{2}{3} D_{SG}; \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

where $d(p_i, p_{i-1})$ denotes the Euclidean distance between two consecutive points, and D_{SG} represents the shortest distance between the start area and the destination goal area.

We require the model to output at least **three consecutive points** to represent a valid path. The final score for each point is computed as a weighted combination of the five components, with weights determined by the point type, as shown in Table **??**. The final score of a predicted path is defined as the **average score over all valid points** in the sequence.

E Additional Evaluation Results

In this section, we report the evaluation results of 76 representative VLMs on EPIC-Bench, as shown

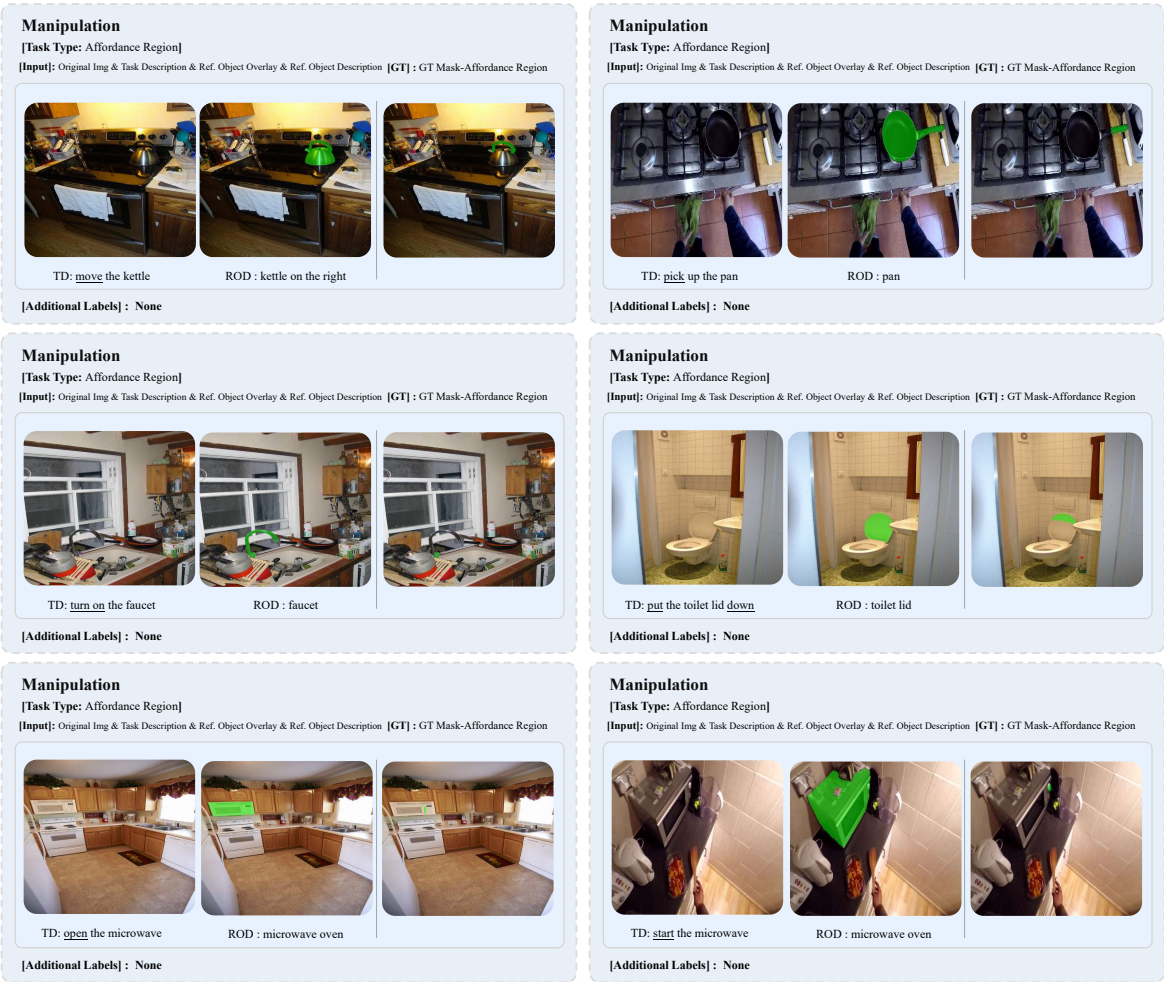


Figure 12: Examples of Manipulation - Affordance Region task

Point Type	Start-End	Traversable Path Ratio	Away-from-Start-Area	Approaching -Goal-Area	Continuity
Start point (first)	1.0	0	0	0	0
End point (last)	0.3	0.4	0.1	0.1	0.1
Intermediate point	0	0.4	0.2	0.2	0.2

Table 6: Weight configuration of the Path Score components.

in Table 7. We organize the models into three main categories:

Proprietary models: Gemini-2.5-Flash-Lite (Comanici et al., 2025), Gemini-2.5-Flash (Comanici et al., 2025), Gemini-2.5-Pro (Comanici et al., 2025), Gemini-3-Flash-Pre (DeepMind, 2025), Gemini-3-pro (DeepMind, 2025), Gemini-3.1-Pro (DeepMind, 2025), Claude-Sonnet-4.6 (Anthropic, 2026), Claude-Haiku-4.5 (Anthropic, 2025b), Claude-Opus-4 (Anthropic, 2025a), qwen3.6-plus (Qwen-Team, 2026b), Qwen3.5-plus (QwenTeam, 2026), Qwen3.6-27B (QwenTeam, 2026), Qwen3.6-35B-A3B (QwenTeam, 2026a), Doubao-Seed-1.8 (Bytedance Seed, 2025b), Doubao-Seed-1.6-Vision (Bytedance Seed, 2025a), HunYuan-T1-Vision (Tencent Hunyuan Team, 2025), HunYuan-Vision-1.5 (Tencent Hunyuan Team, 2025), HunYuan-Turbos-Vision (Tencent Hunyuan Team, 2025), o3 (OpenAI, 2025c), o4-mini (OpenAI, 2025c), GPT-4o (OpenAI et al., 2024), GPT-5-mini (Singh et al., 2025), GPT-5.1 (OpenAI, 2025a), GPT-5.2 (OpenAI, 2025b), GPT-5.4 (OpenAI, 2026a), GPT-5.5 (OpenAI, 2026b).

Open-source models: Qwen3.6-35B-

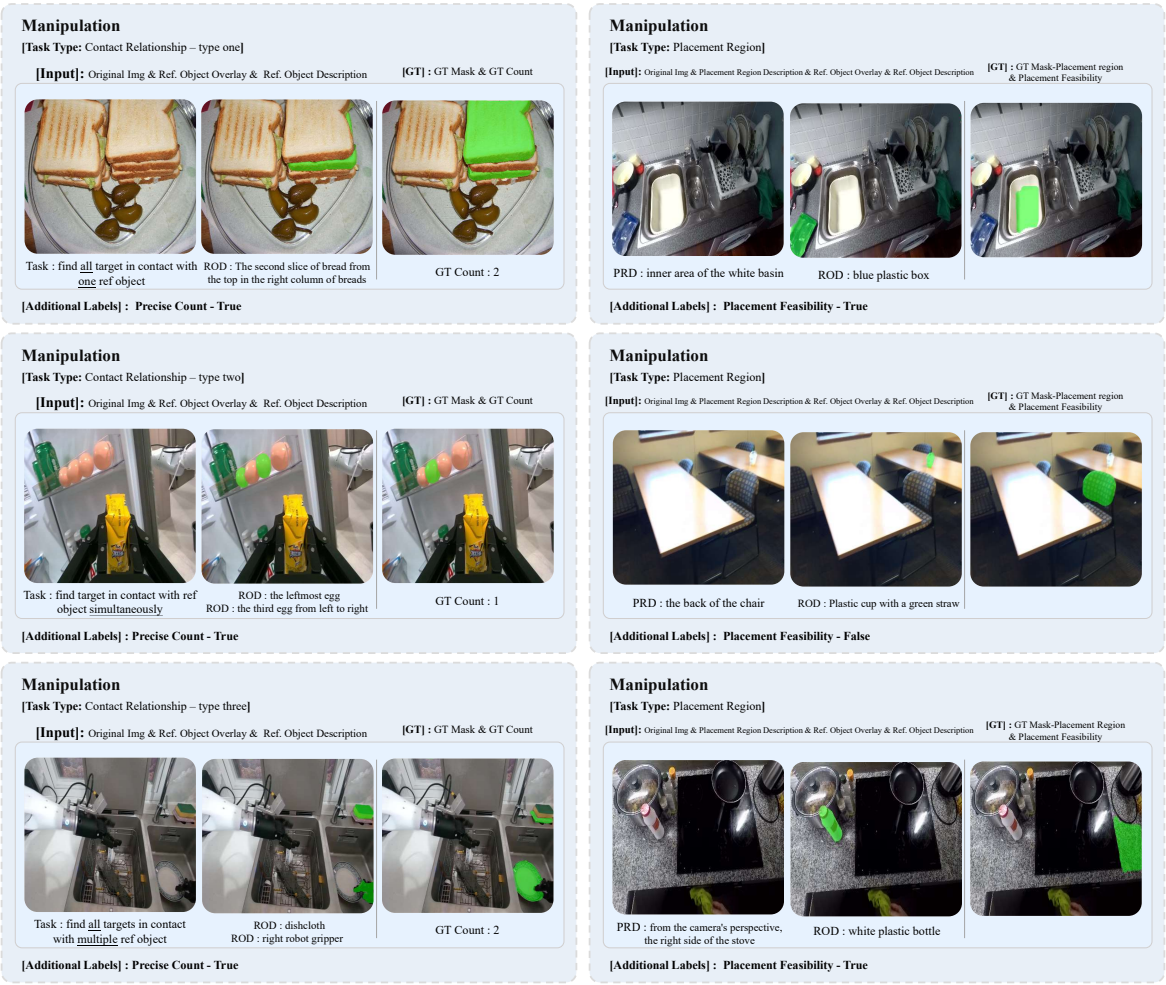


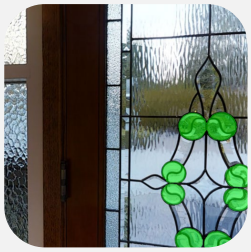
Figure 13: Examples of Manipulation - Contact Relationship and Placement Region tasks

A3B (QwenTeam, 2026a), Qwen3.6-27B (QwenTeam, 2026), Qwen3.5 (QwenTeam, 2026), Qwen3-VL (Bai et al., 2025a), Qwen2.5-VL-72B (Bai et al., 2025b), InternVL3.5 (Wang et al., 2025), InternVL3-78B (Zhu et al., 2025), InternVL3-38B (Zhu et al., 2025), InternVL3-14B (Zhu et al., 2025), InternVL3-8B (Zhu et al., 2025), Gemma-3 (Team et al., 2025d), MiMo-VL-7B (Team et al., 2025a), GLM-4.6V (GLM Team, 2025), Step3-VL-10B (Huang et al., 2026), LLaVA-NeXT-72B (Li et al., 2024).

Embodied foundation models: RoboBrain2 (Cao et al., 2025), MiMo-Embodied-7B (Hao et al., 2026), RoboBrain2.5 (Tan et al., 2026), Pelican1.0-VL-72B (Zhang et al., 2025), RynnBrain (Dang et al., 2026), VeBrain (Luo et al., 2025), Cosmos-Reason1-7B (NVIDIA et al., 2025).

Our evaluation covers a diverse range of model scales, from large models to lightweight variants, and includes both open-source and proprietary systems. We hope these results will provide a comprehensive reference for researchers when selecting VLMs for embodied perception tasks.

Target Localization



[System Prompt] You are a target localization assistant.
Task

You will receive
1. **Image** - An RGB image of the scene.
2. **Target Object Description** - The detailed description of the target object to be localized.

Based on the image and the target object description, you need to localize all the target objects in the image. Notice, there might be multiple or none target objects in the image.

The number of the objects in your answer should be accurate and consistent with the target object description.
Your localization answer must be precise, and try to maximize the Intersection over Union (IOU) with the Ground Truth as much as possible.
+Response format [C.2→TL]

Figure 14: **Prompt for Target Localization.** Target Localization requires the model to identify all target objects matching the textual description.

Ground Detection

[System Prompt] You are a ground detection assistant.
Task

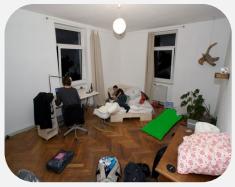
You will receive
1. **Image** - An RGB image of the scene.

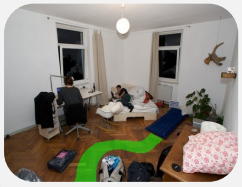
You need to detect all the ground areas in the image.

Your localization answer must be precise, and try to maximize the Intersection over Union (IOU) with the Ground Truth as much as possible.

You should return multiple bounding boxes if the ground areas are separated by non-ground areas.
+Response format [C.2→GD]

Feasible Path





[System Prompt]You are a path-finding assistant.
Task

You will receive
1. **Two Images** - The first image is the original image from the **first/third** perspective, the second image is the overlay image of the target area on the first image.
2. **Target Area Description** - The detailed description of the target **area/areas**.

Based on the scene image and the overlay image of the target area and the target area description, you need to find the feasible path **to reach the target area/from one target area to another target area** in the scene image.

The path should start from **the place where the picture is taken/any one target area**.
The path you find should be the shortest path, but the path in your answer must all be on the ground area.
The path you find must be a path that exists, though it might be visually blocked. Visually blocked means the path you find may not be seen from the image with the current perspective, but the actual feasible path exists.
+Response format [C.2→FP]

Visual Matching

[System Prompt] You can match areas across different perspectives.
Task

You will receive
1. **Three Images** - The first image is the original scene from one perspective, the second image is the overlay image of the reference areas in the first image, and the third is the image of the original scene from a different perspective.

Based on the first and second images, you need to locate and count the targets in image 3 matching references in images 1-2.

Your matching answer of the target areas must be precise, and try to maximize the Intersection over Union (IOU) with the Ground Truth of targets areas as much as possible.
+Response format [C.2→VM]

Figure 15: **Prompts for Navigation.** Navigation category includes three sub-tasks: Ground Detection (identifying traversable ground regions), Feasible Path (planning routes to target objects in egocentric view or between two targets in exocentric view), and Visual Matching (localizing the same reference area across different viewpoints).

Affordance Region

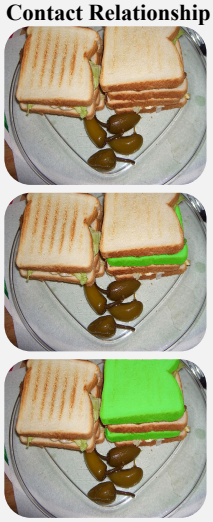
[System Prompt] You are an affordance region detection assistant.
Task

You will receive
1. **Two Images** - The first image is the original image, the second image is the overlay image of the reference object on the first image.
2. **Reference Object Description** - The detailed description of the reference object.
3. **Task Description** - The detailed description of the task description.

You need to localize the Affordance Region of the reference object that can be used to complete the task in the first image.

Your answer must be precise, and try to maximize the Intersection over Union (IOU) with the Ground Truth as much as possible.
+Response format [C.2→AR]

Contact Relationship



[System Prompt] You are a contact relationship detection assistant.
Task

You will receive
1. **Two Images** - The first one is the image of the scene, the second one is the overlay image of the reference object on the scene image.
2. **Reference Object Description** - The detailed description of the reference **object/objects.**

Based on the scene image and the overlay image and the reference object description, you need to find all the target objects that are **DIRECTLY in contact with the reference object/DIRECTLY and SIMULTANEOUSLY in contact with ALL reference objects/DIRECTLY in contact with ANY reference objects.**

You must return the number of target objects.
The target objects you return must NOT include any object that is excessively larger than the reference **object/objects.**
Your localization answer must be precise, and try to maximize the Intersection over Union (IOU) with the Ground Truth as much as possible.
+Response format [C.2→CR]

Placement Region

[System Prompt] You are a placement suitability assistant.
Task

You will receive
1. **Two Images** - The first image is the original scene image, the second image is the overlay image of the reference object on the first image.
2. **Reference Object Description** - The detailed description of the reference object.
3. **Placement Region Description** - The detailed description of the placement region.

You need to localize the Placement Region in the scene image based on Placement Region Description. And return whether the placement region can be used for the reference object based on the size, stability, or other reasonable factors in the scene image and overlay image.

Your localization answer must be precise, and try to maximize the Intersection over Union (IOU) with the Ground Truth as much as possible.
+Response format [C.2→PR]

Figure 16: **Prompts for Manipulation.** Affordance Region (localizing operable regions of a reference object), Contact Relationship (identifying objects in contact with single/multiple reference objects under three contact conditions), and Placement Region (localizing placement areas and determining placement feasibility).

Target Localization (TL), Contact Relationship (CR), Visual Matching (VM)

You need to return the answer in the following JSON format:

```
{
    "number_of_objects": int,
    "bounding_boxes": [(x1, y1, x2, y2), (x3, y3, x4, y4), ...],
}
```

Each bounding box is in the format of (x1, y1, x2, y2), where (x1, y1) is the top-left corner of the bounding box, and (x2, y2) is the bottom-right corner of the bounding box.

Affordance Region (AR), Ground Detection (GD)

You need to return the answer in the following JSON format:

```
{
    "bounding_boxes": [(x1, y1, x2, y2), (x3, y3, x4, y4), ...],
}
```

Each bounding box is in the format of (x1, y1, x2, y2), where (x1, y1) is the top-left corner of the bounding box, and (x2, y2) is the bottom-right corner of the bounding box.

Placement Region (PR)

You need to return the answer in the following JSON format:

```
{
    "bounding_boxes": [(x1, y1, x2, y2), (x3, y3, x4, y4), ...],
    "placement_feasibility": Boolean,
}
```

Each bounding box is in the format of (x1, y1, x2, y2), where (x1, y1) is the top-left corner of the bounding box, and (x2, y2) is the bottom-right corner of the bounding box.

Feasible Path (FP)

You need to return a list of multiple points to represent the path in the following JSON format:

```
{
    "points": [(x1, y1), (x2, y2),(x3, y3),...,(xn, yn)],
}
```

where x is the x coordinate of the point, y is the y coordinate of the point, and n >= 3 is the number of points in the path.

Figure 17: **Response format for each sub-task.** For **Target Localization (TL)**, **Contact Relationship (CR)**, and **Visual Matching (VM)**, models are required to output bounding boxes and the count of target objects. **Affordance Region (AR)** and **Ground Detection (GD)**, models are required to output the corresponding bounding boxes. For **Placement Region (PR)**, models are required to output the placement region bounding box along with a binary judgment on placement feasibility. For **Feasible Path (FP)**, models are required to output a sequence of points representing the path.

Model	Average	Target Local.			Navigation			Manipulation		
		Bas.	Spa.	Emb.	Gro.	Fea.	Vis.	Aff.	Con.	Pla.
Proprietary Models										
Gemini-2.5-Flash-Lite	34.77	39.06	39.53	31.51	35.97	40.21	33.02	2.580	30.17	32.38
Gemini-2.5-Flash	36.33	38.46	39.16	34.05	27.28	51.98	40.79	2.880	37.15	36.28
Gemini-2.5-Pro†	37.63	42.86	41.62	36.43	9.630	46.29	42.70	8.720	33.94	42.78
Gemini-3-Flash-Pre	37.04	46.40	31.48	32.29	48.98	22.89	47.25	11.91	42.29	48.09
Gemini-3-Flash-Pre†	44.69	50.26	41.70	37.79	**51.11**	52.63	49.92	22.62	46.51	50.89
Gemini-3-pro†	**54.81**	56.81	54.50	**53.90**	48.82	**70.82**	**60.08**	**30.40**	51.78	53.42
Gemini-3.1-Pro†	54.72	55.38	**56.65**	53.04	47.22	70.32	59.06	28.89	**53.70**	**55.25**
Claude-Sonnet-4.6†	43.24	45.43	45.49	36.47	40.16	68.46	42.77	9.790	46.21	43.81
Claude-Sonnet-4.6	31.34	43.84	32.16	21.94	5.250	39.56	40.90	0.540	35.90	35.77
Claude-Haiku-4.5†	22.03	25.55	24.15	17.59	3.920	33.82	36.34	0.400	9.810	30.37
Claude-Haiku-4.5	30.94	31.96	34.50	28.78	24.76	40.91	35.88	1.890	29.88	33.18
Claude-Opus-4	25.71	26.60	31.22	22.47	12.44	27.30	32.57	1.420	27.27	34.81
qwen3.6-plus†	45.48	47.66	50.17	41.07	44.47	62.12	41.57	11.95	47.84	41.59
o3†	36.05	39.37	40.96	32.86	36.24	30.64	40.73	5.490	32.43	43.64
o4-mini†	32.83	36.75	38.02	32.73	18.27	20.49	45.63	3.540	33.58	38.75
GPT-4o	29.52	27.42	35.99	26.69	26.95	32.45	36.36	2.180	33.22	34.50
GPT-5-mini†	35.21	38.70	39.43	30.24	24.36	38.53	45.97	3.860	41.32	38.78
GPT-5.1	34.81	35.31	39.66	32.05	36.62	38.62	44.69	4.340	26.44	38.63
GPT-5.2	36.82	38.26	41.53	32.02	33.06	48.72	44.10	5.010	35.25	37.26
GPT-5.4	38.95	43.78	43.90	33.84	32.24	43.12	44.35	8.670	39.55	37.30
GPT-5.5	50.16	**57.78**	55.37	45.19	50.97	38.57	51.65	24.62	49.67	48.70
qwen3.6-plus†	45.48	47.66	50.17	41.07	44.47	62.12	41.57	11.95	47.84	41.59
Qwen3.5-plus†	47.10	55.09	51.90	41.82	50.85	57.89	45.21	13.82	22.34	42.46
Qwen3.6-27B†	36.31	46.33	45.05	36.31	11.61	35.10	28.07	10.76	21.53	26.78
Qwen3.6-35B-A3B†	35.07	48.17	43.99	36.61	14.41	33.97	21.30	2.840	17.05	15.90
Doubao-Seed-1.8†	46.32	55.29	49.68	46.10	46.02	53.43	45.90	10.94	14.36	40.79
Doubao-Seed-1.6-Vision†	46.80	54.64	49.92	42.27	41.64	58.74	42.66	9.520	47.48	41.62
HunYuan-T1-Vision†	43.00	51.24	47.60	39.67	42.63	50.08	33.54	6.630	38.95	34.49
HunYuan-Vision-1.5	32.55	38.37	39.60	29.67	31.72	17.07	35.71	6.380	32.88	29.25
HunYuan-Turbos-Vision	24.61	24.72	33.33	25.58	4.490	17.88	36.14	0.670	22.02	26.69
Open-source Models										
Qwen3.6-35B-A3B	38.52	46.40	46.19	40.01	3.900	39.16	39.42	5.410	34.19	34.52
Qwen3.6-27B	45.38	51.04	48.02	41.46	49.99	59.15	43.45	10.61	38.69	37.35
Qwen3.5-397B-A17B†	47.47	55.45	51.82	42.10	51.10	60.11	45.35	12.40	26.33	42.51
Qwen3.5-397B-A17B	45.16	50.56	50.06	43.45	**51.94**	50.61	41.46	8.310	34.51	37.08
Qwen3.5-122B-A10B†	47.24	53.61	49.13	41.16	51.42	**63.47**	45.19	12.31	46.04	39.77
Qwen3.5-122B-A10B	44.54	50.77	49.28	42.31	49.65	50.50	42.62	8.300	36.91	32.16
Qwen3.5-35B-A3B†	45.58	53.51	49.18	39.03	43.09	61.17	41.18	10.75	41.70	38.05
Qwen3.5-35B-A3B	41.32	46.76	45.82	38.19	48.00	45.47	37.01	10.01	36.94	32.36
Qwen3.5-27B†	45.67	54.15	50.03	40.61	46.96	39.43	46.77	13.15	**49.56**	39.84
Qwen3.5-27B	44.71	49.10	48.55	42.23	49.74	53.50	44.10	11.93	37.22	36.42
Qwen3.5-9B†	38.86	44.60	43.38	33.87	31.26	49.06	40.22	6.070	37.64	34.92
Qwen3.5-9B	35.45	40.49	42.04	31.90	38.07	36.81	33.93	6.090	26.70	28.25
Qwen3.5-4B†	35.59	40.83	42.34	31.24	24.09	45.21	35.33	5.650	24.79	33.90
Qwen3.5-4B	33.27	39.84	40.87	30.04	33.70	29.66	32.54	5.560	17.60	26.49
Qwen3.5-2B†	28.49	33.36	35.06	27.34	14.58	25.31	25.13	3.390	30.10	27.01

Continued on next page

Model	Average	Target Local.			Navigation			Manipulation		
		Bas.	Spa.	Emb.	Gro.	Fea.	Vis.	Aff.	Con.	Pla.
Qwen3.5-2B	23.12	26.93	32.91	25.11	7.840	3.720	22.43	0.370	20.64	24.09
Qwen3.5-0.8B†	19.27	21.89	28.72	19.82	0.770	9.260	13.02	0.950	18.88	22.36
Qwen3.5-0.8B	4.890	2.580	4.620	2.050	0.150	4.940	11.33	0.570	13.12	20.67
Qwen3-VL-235B-A22B†	**50.93**	**58.00**	**55.67**	**48.12**	48.18	51.19	**49.01**	16.96	48.04	**46.49**
Qwen3-VL-235B-A22B	42.64	50.12	45.86	39.34	50.96	43.14	38.69	12.00	34.55	35.71
Qwen3-VL-30B-A3B†	44.02	52.69	49.37	40.87	42.34	41.65	36.20	10.62	38.78	40.72
Qwen3-VL-30B-A3B	30.60	28.31	34.27	25.04	49.37	37.09	30.12	9.330	31.49	34.07
Qwen3-VL-32B†	37.21	39.94	41.26	36.05	31.03	42.04	39.07	4.840	35.23	40.03
Qwen3-VL-32B	35.21	37.49	39.82	33.21	44.38	36.74	36.57	4.720	30.62	30.49
Qwen3-VL-8B†	35.55	39.92	41.66	33.89	21.68	38.51	34.94	5.220	32.05	37.01
Qwen3-VL-8B	34.07	39.48	42.54	32.10	38.11	26.67	31.84	5.190	14.63	30.91
Qwen3-VL-4B†	32.83	38.42	41.30	31.97	18.29	29.47	30.95	3.910	22.15	31.83
Qwen3-VL-4B	33.24	39.95	41.55	33.78	40.39	18.45	30.17	3.200	11.30	27.00
Qwen3-VL-2B†	31.82	39.67	40.75	30.08	15.36	24.02	22.59	6.070	28.24	28.97
Qwen3-VL-2B	32.53	38.43	42.87	29.08	37.55	9.780	28.55	7.990	29.56	29.59
Qwen2.5-VL-72B	42.51	47.54	49.79	40.92	49.45	42.49	37.08	14.77	21.21	35.90
InternVL3.5-241B-A28B	40.75	44.65	50.59	37.61	35.94	35.47	33.73	16.85	36.41	37.60
InternVL3.5-38B	42.54	48.69	52.60	39.95	43.21	35.01	28.07	**17.22**	34.05	35.83
InternVL3.5-30B-A3B	29.71	30.96	39.11	27.03	45.43	24.11	17.24	8.410	18.75	25.30
InternVL3.5-14B	38.92	43.95	49.96	37.13	46.11	31.82	24.57	11.23	25.98	28.66
InternVL3.5-8B	33.58	40.67	45.40	32.05	32.15	22.77	15.94	10.79	17.42	25.57
InternVL3-78B	36.04	40.68	44.56	32.78	35.98	30.55	34.93	6.900	29.07	31.69
InternVL3-38B	28.50	34.88	28.85	23.22	36.69	22.31	34.88	3.990	30.43	30.44
InternVL3-14B	30.17	31.98	38.44	25.41	32.29	29.53	27.64	3.250	27.49	31.19
InternVL3-8B	30.23	35.51	40.72	26.69	33.38	21.53	23.16	3.830	16.08	26.07
Gemma-3-27B-IT	27.17	26.93	33.66	26.68	11.62	32.17	26.33	1.650	31.26	32.18
Gemma-3-12B-IT	26.21	25.55	32.44	24.59	13.71	36.27	23.39	1.300	27.48	31.43
Gemma-3-4B-IT	23.50	21.79	31.48	22.31	8.040	27.53	22.04	0.970	28.80	29.59
MiMo-VL-7B-SFT-2508	31.18	33.01	37.17	32.98	17.59	29.14	36.89	4.910	23.20	32.82
MiMo-VL-7B-RL-2508	34.65	36.65	42.31	34.92	18.31	36.90	37.45	6.510	30.37	32.86
GLM-4.6V†	42.84	50.19	47.71	37.89	42.46	51.11	36.55	8.320	34.14	39.79
Step3-VL-10B	32.40	42.22	40.10	31.88	17.96	24.03	17.77	5.940	27.22	27.56
LLaVA-NeXT-72B	20.40	26.54	29.42	19.93	14.18	6.340	6.190	2.990	12.75	18.63
Embodied Foundation Models										
RoboBrain2-32B	**39.32**	**50.66**	**49.67**	**38.83**	38.85	0.610	37.63	0.080	**36.22**	**39.61**
RoboBrain2-7B	30.09	35.93	36.01	30.52	42.92	0.000	24.31	2.290	31.58	32.95
MiMo-Embodied-7B	37.16	41.22	44.17	38.74	24.67	**32.22**	**40.23**	**8.710**	28.23	33.88
RoboBrain2.5-8B-NV	23.81	24.81	32.88	24.97	22.57	1.660	27.98	1.130	24.98	24.03
RoboBrain2.5-8B-MT	29.83	31.20	37.41	26.76	33.52	30.89	27.84	2.650	25.45	24.34
Pelican1.0-VL-72B	35.29	44.08	42.11	34.79	**44.91**	0.000	38.13	7.320	22.68	37.96
RynnBrain-8B	22.34	28.31	33.28	24.81	7.980	5.950	15.04	0.130	13.84	11.26
RynnBrain-CoP-8B	16.83	12.48	21.78	16.08	21.51	22.51	13.58	0.280	19.62	18.09
RynnBrain-Plan-8B	23.59	23.53	32.53	26.08	2.120	19.57	21.41	0.110	18.79	31.02
RynnBrain-2B	12.05	12.84	16.49	17.07	0.000	7.100	13.35	0.010	8.730	12.45
VeBrain	29.27	33.63	35.98	27.69	29.56	20.04	27.58	2.850	19.25	32.30
Cosmos-Reason1-7B	23.25	29.74	26.64	19.67	24.68	3.030	25.44	5.370	22.12	32.75

Continued on next page

Model	Average	Target Local.			Navigation			Manipulation		
		Bas.	Spa.	Emb.	Gro.	Fea.	Vis.	Aff.	Con.	Pla.

Table 7: The full evaluation results for all **89 VLM variants** on EPIC-Bench. **Green bold** indicates the overall best result across all models. **Bold** and underline indicate the best and second-best results within each model category, respectively. † denotes models evaluated with thinking mode enabled.