PLAN-BENCH: CAN EXISTING VLMS GO WELL BEYOND PERCEPTION TO REASON AND PLAN IN EMBODIED WORLDS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language Models (VLMs) have achieved strong performance in visual perception and spatial reasoning, leading to growing interest in using them as high-level planners for embodied AI. However, physical tasks require models to integrate spatial information, reason over temporal states, and generate multi-step plans under various constraints, raising the question of whether current VLMs can reliably support such reasoning and planning independently of execution. We introduce **Plan-Bench**, a benchmark for evaluating the high-level reasoning and multi-step planning capabilities of VLMs relevant to embodied worlds, while isolating the influence of action execution. It covers three complementary dimensions, spanning 16 subclasses and 699 curated samples. Extensive evaluation of 28 VLMs reveals room for improvement, with even the strongest model achieving only 75.0% accuracy. We further find that model scaling generally improves these capabilities. Ablation studies also indicate larger models show increasing potential as high-level planners. Overall, Plan-Bench provides a lightweight probe for evaluating VLMs' reasoning and planning capabilities relevant to embodied environments and assessing their potential as decision makers for physical tasks.

1 INTRODUCTION

Recent advances in vision-language models (VLMs) (OpenAI, 2026, Qwen Team, 2026a, Diao et al., 2026c) have substantially strengthened their capabilities in visual perception and spatial reasoning, leading to growing interest in using them as high-level planners for embodied AI (Zhang et al., 2026, Liu et al., 2026, Yang et al., 2026), particularly in complex physical tasks. In such settings, models often need to integrate spatial information, reason over evolving temporal states, and generate multi-step plans under various constraints. These demands expose an important and underexplored challenge: can existing VLMs go well beyond perception to reason and plan in embodied worlds?

Notably, existing benchmarks have not systematically evaluated these capabilities of VLMs for embodied tasks. As shown in Figure 1, multimodal reasoning benchmarks (Fu et al., 2026, Yao et al., 2026) primarily focus on complex visual information, with limited evaluation of the capabilities required for reasoning and planning in embodied environments. In contrast, embodied benchmarks (Fan et al., 2026, Han et al., 2026) mainly evaluate complete embodied tasks, mixing errors from high-level decisions and low-level execution. LongAct (Zhu et al., 2026) attempts to reduce the influence of low-level control and better assess high-level decision making, but evaluates agents rather than VLMs. It therefore remains unclear whether VLMs can perform reliable reasoning and planning in embodied scenarios when high-level decisions are separated from low-level execution.

To bridge this gap, we introduce Plan-Bench, a benchmark designed to evaluate VLMs' high-level reasoning and multi-step planning relevant to embodied tasks. As illustrated on the right side of Figure 1, Plan-Bench systematically covers three dimensions: Spatial Cognition, Temporal Reasoning, and Multi-Step Planning, spanning 16 subclasses and 699 samples. It is constructed through a three-stage pipeline consisting of task definition, dual-source visual data acquisition, and manual QA construction with quality assurance, resulting in a diverse and carefully curated benchmark dataset. Plan-Bench includes both single-choice and open-ended questions, and provides answer hints for samples with multiple valid answers to guide more reliable LLM-as-judge evaluation.

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.

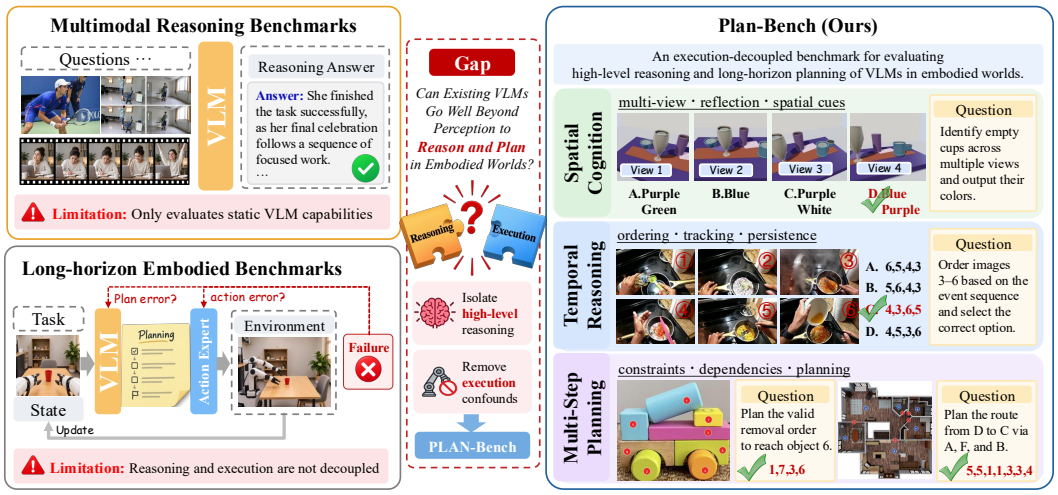

Figure 1: Overview of Plan-Bench. The left illustrates the limitations of existing benchmarks, which either evaluate static VLM capabilities or couple high-level reasoning with low-level execution. The right presents the composition of Plan-Bench, which evaluates execution-decoupled high-level reasoning and planning across Spatial Cognition, Temporal Reasoning, and Multi-Step Planning.

We evaluate Plan-Bench on 28 VLMs. The results show that current VLMs remain limited in reasoning and planning for embodied scenarios. The best-performing model, Qwen3.8-Max (Qwen Team, 2026a), achieves only 75.0% accuracy. We further observe that performance generally improves with model scaling, with some variation across model families. Our ablation studies show that larger models exhibit greater potential as high-level planners, and that hallucination remains one of the major sources of errors in VLMs. Overall, Plan-Bench provides a principled testbed for evaluating reasoning and multi-step planning in VLMs and a lightweight framework for assessing their potential as high-level embodied planners. Our main contributions are summarized as follows:

- **Designing a reasoning and planning benchmark for VLMs in embodied worlds.** Plan-Bench is an easy-to-deploy benchmark spanning 16 subclasses and 699 samples, enabling direct evaluation of VLMs' high-level capabilities independently of low-level execution.
- **Providing a comprehensive evaluation of existing VLMs.** We benchmark 28 representative VLMs and reveal their capabilities in high-level reasoning and multi-step planning.
- **Revealing current limitations and potential.** Evaluation across proprietary and open-source models shows that reasoning and planning in embodied scenarios remain challenging, while larger models exhibit greater potential as high-level planners for embodied tasks.

2 Related Work

2.1 Multimodal Reasoning Benchmarks

Existing benchmarks have gradually evolved from perception toward complex reasoning. MME (Fu et al., 2026) evaluates both perceptual and cognitive abilities, while LENS (Yao et al., 2026) organizes multimodal evaluation from understanding to reasoning. Recent benchmarks such as VisuLogic (Xu et al., 2026b), MIRA (Zhou et al., 2026), and PuzzleWorld (Li et al., 2026a) further explore vision-centric, chain-of-thought, and open-ended multi-step reasoning. Beyond single-image reasoning, MuirBench (Wang et al., 2025a), MMIU (Meng et al., 2025), EgoPlan-Bench2 (Qiu et al., 2026), among others (Li et al., 2026b, Meng et al., 2025) extend evaluation to multiple images or frames, covering multi-view relations, temporal changes, and long-range visual evidence. These benchmarks broaden the scope and complexity of multimodal reasoning. However, existing benchmarks rarely target the reasoning and planning demands of embodied worlds, with their evaluated capabilities remaining centered on reasoning complexity. Plan-Bench addresses this gap by evaluating these capabilities across Spatial Cognition, Temporal Reasoning, and Multi-step Planning.

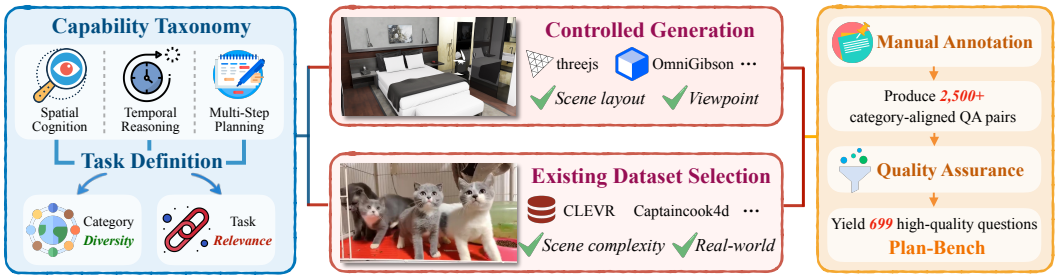

Figure 2: Overview of the Plan-Bench construction pipeline, including capability-driven task definition, dual-source data acquisition, manual annotation, and quality assurance.

2.2 Multi-Step Tasks in Embodied Environments

Recent embodied systems increasingly employ VLMs as decision makers. Harness VLA (Zhang et al., 2026), Goal2Skill (Liu et al., 2026), and HiVLA (Yang et al., 2026), among others (Shi et al., 2025, Li et al., 2025, Yang et al., 2025c), adopt hierarchical frameworks that separate high-level semantic reasoning and planning from low-level action execution. These works demonstrate the potential of VLMs to serve as high-level planners for complex embodied tasks. Meanwhile, several benchmarks have begun to evaluate these capabilities in embodied settings. EmbodiedBench (Yang et al., 2025b) and RoboCerebra (Han et al., 2026) evaluate embodied agents through interactive tasks or hierarchical VLM-VLA systems, where performance jointly reflects reasoning, grounding, and action execution. RoboSPA (Fan et al., 2026) further evaluates VLA models on spatial reasoning and multi-step procedural planning, but still couples reasoning with low-level execution. LongAct (Zhu et al., 2026) attempts to reduce the influence of low-level control to better assess high-level planning, yet evaluates complete embodied agents rather than VLMs alone. Plan-Bench instead decouples embodied action execution and directly evaluates VLMs' high-level reasoning and multi-step planning by constructing a benchmark of 699 questions across 16 subclasses, providing a lightweight and unified interface for isolating the capabilities required by embodied tasks.

3 Plan-Bench

We introduce Plan-Bench, a benchmark designed to systematically evaluate the reasoning and planning capabilities of VLMs in embodied tasks. In this section, we first describe the construction pipeline, then present the details of data distribution, and finally introduce the evaluation protocol.

3.1 Construction Pipeline

To ensure high quality and validity, we adopt a task-driven, multi-stage construction pipeline, as illustrated in Figure 2. The overall process consists of three stages:

Capability Taxonomy and Task Definition. We first organize the capabilities required for reasoning and planning in embodied environments into a taxonomy. The taxonomy comprises three core categories corresponding to Spatial Cognition, Temporal Reasoning, and Multi-Step Planning. Based on this taxonomy, we further define a set of subtasks, balancing category diversity with task relevance to the targeted capabilities. This step provides explicit guidance for data acquisition, determining what types of visual evidence and task conditions should be collected or constructed.

Dual-Source Data Acquisition. We collect visual data from two sources. For subclasses requiring precise control over factors such as scene layout and viewpoint, we employ controlled-generation platforms such as Three.js (Danchilla, 2012) and OmniGibson (Li et al., 2023) to configure task-relevant variables for consistent construction. For those that benefit from more natural visual distributions, we select suitable images from existing datasets such as CLEVR (Johnson et al., 2017) and Captaincook4d (Peddi et al., 2024), among others (Yang et al., 2019, Dai et al., 2026, Ding et al., 2023, Lin et al., 2020), to increase visual variety and capture greater real-world scene complexity.

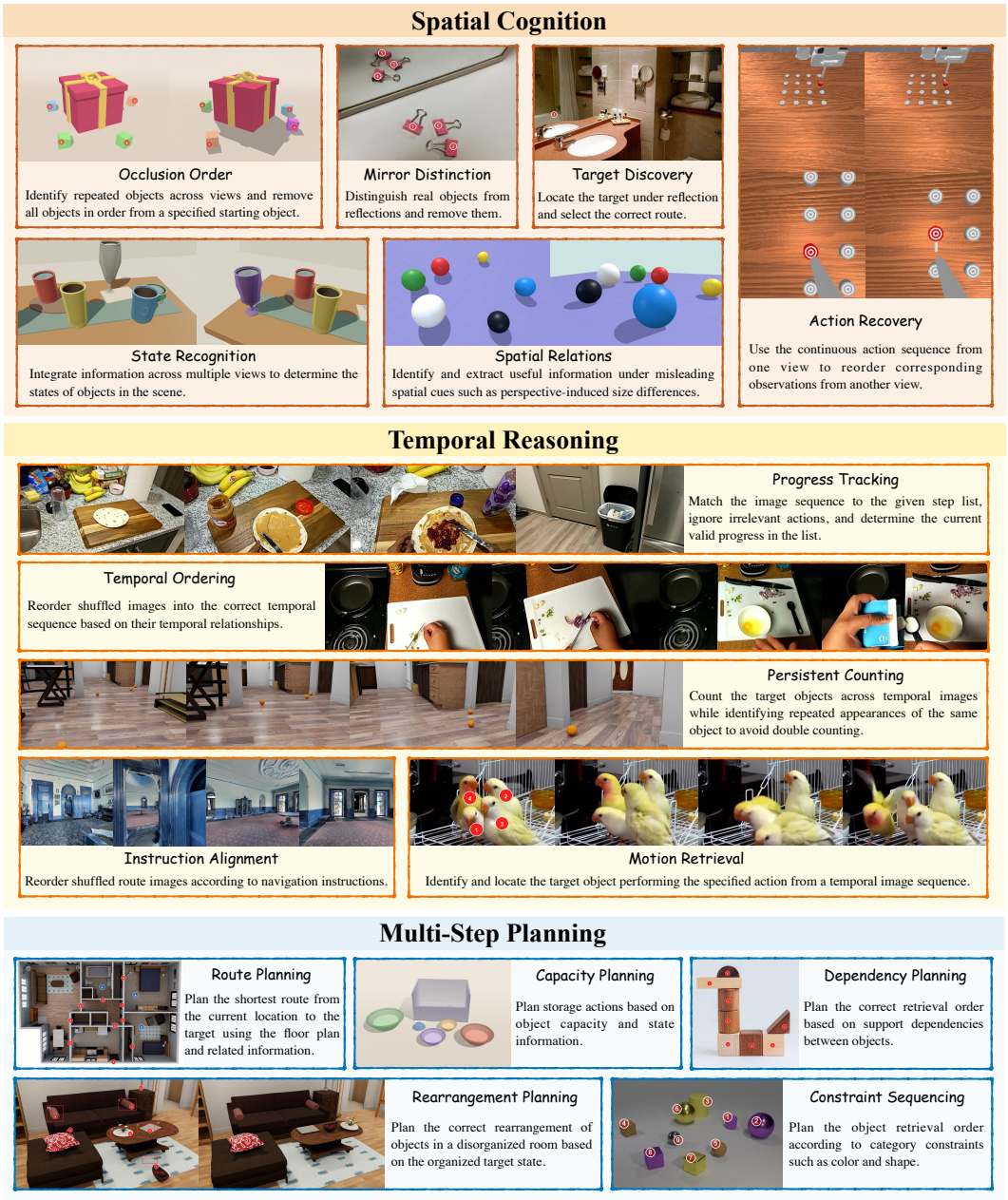

Figure 3: Illustration of representative samples from the three dimensions of Plan-Bench, including Spatial Cognition, Temporal Reasoning, and Multi-Step Planning.

QA Annotation and Quality Assurance. Based on the acquired images, we manually construct 2,500+ questions and answers according to the task specification of each subclass. All samples are then independently reviewed by three annotators who are not involved in the initial annotation process. The review considers answer correctness, image-QA consistency, and relevance to the corresponding task definition. A sample is accepted only when three reviewers agree that it satisfies all criteria. Finally, this process yields 699 high-quality questions to construct Plan-Bench.

3.2 Data Distribution

Plan-Bench comprises 699 instances, including 183 single-choice questions and 516 open-ended questions. Our benchmark covers 16 subtasks across three primary categories, providing a compre-

hensive evaluation of VLMs' reasoning and planning capabilities in diverse embodied task settings. Figure 3 presents representative examples from each category in Plan-Bench. The design motivations and task compositions of these three categories are detailed below:

Spatial Cognition. This category contains 259 instances across six subtasks: *Occlusion Order*, *Mirror Distinction*, *Action Recovery*, *Target Discovery*, *State Recognition*, and *Spatial Relations*. It is designed to reflect embodied settings where VLMs may observe the same scene through multiple cameras or viewpoints and must integrate these observations to form a coherent understanding of the environment. It also considers challenging spatial cues such as mirrors and reflections, requiring VLMs to distinguish reflected content from physical entities and reason correctly about the scene.

Temporal Reasoning. This category contains 210 instances across five subtasks: *Progress Tracking*, *Temporal Ordering*, *Persistent Counting*, *Instruction Alignment*, and *Motion Retrieval*. This category includes many egocentric multi-image sequences, designed to reflect embodied tasks such as cooking and navigation, where VLMs must reason over temporally evolving observations from a first-person perspective. It evaluates whether VLMs can understand temporal dependencies across sequential observations, including ordering events, tracking target actions or states, and identifying recurring objects across frames. These tasks further assess whether models can preserve relevant historical context and maintain consistent reasoning throughout extended embodied interactions.

Multi-Step Planning. This category contains 230 instances across five subtasks: *Constraint Sequencing*, *Capacity Planning*, *Rearrangement Planning*, *Dependency Planning*, and *Route Planning*. It reflects embodied scenarios where VLMs must first reason about task goals, scene constraints, and action dependencies to produce plans before execution. It evaluates whether VLMs can translate such information into action sequences toward a target. Successful planning may indicate the potential of VLMs to serve as more stable high-level decision makers in physical systems.

3.3 Evaluation Protocol

We evaluate model responses using an LLM-as-judge approach. Many open-ended questions in Plan-Bench admit multiple valid answers due to equivalent object references, alternative orderings, or other semantically equivalent forms. To handle such cases, we pair these samples with answer hints that specify additional valid alternatives or equivalence conditions beyond the ground-truth answer, such as valid order swaps, position exchanges, or alternative object names. DeepSeek-V4-Flash (Xu et al., 2026a) serves as the judge, taking the model output, ground-truth answer, and answer hint as input to determine whether the response should be considered correct under these acceptable alternatives. To validate the reliability of the evaluation process, we randomly sample 200 instances for manual verification, and the judge achieves 100% agreement with human evaluation on these samples. The full judge prompt is provided in the Appendix B.2 for reproducibility.

4 Experiments

In this section, we present an evaluation of VLMs on Plan-Bench to examine their reasoning and planning capabilities in embodied worlds. We first introduce the experimental setup, then report the main results on 28 VLMs, and finally provide ablation studies to further investigate model behaviors.

4.1 Setup

Models. We evaluate Plan-Bench on a broad suite of 28 VLMs, including 23 open-source models and 5 commercial models. The evaluated models are generally grouped into four categories according to model scale and deployment setting: *(i)* lightweight models with around 2B-4B parameters, *(ii)* middle-scale models with 7B–12B parameters, *(iii)* larger open-source models, comprising 27B–31B dense models and Mixture-of-Experts (MoE) models, and *(iv)* commercial frontier models. For open-source models, we evaluate a diverse set of model families with different architectures, including Qwen3.8 (Qwen Team, 2026c), Qwen3.5 (Qwen Team, 2026b), Qwen3-VL (Yang et al., 2025a), Gemma-4 (Team et al., 2026a), Step3-VL (Huang et al., 2026), GLM-4.6V (Hong et al., 2025), InternVL3.5 (Wang et al., 2025b), Ovis (Lu et al., 2025), SenseNova-U1 (Diao et al., 2026c), NEO (Diao et al., 2026a), and MiMo-VL (Yue et al., 2025). For commercial models, we evalu-

Table 1: Comparison of different VLMs on Plan-Bench. We report performance across the three task categories: Spatial Cognition, Temporal Reasoning, and Multi-Step Planning, together with the overall accuracy. The **best** results in each column are bolded. The random baseline is computed by assigning 0% to open-ended questions and 25% to single-choice questions. All values are rounded to the nearest tenth of a percentage point and reported as percentages (%).

Model	Spatial Cognition	Temporal Reasoning	Multi-Step Planning	Overall (699)
Random Baseline	10.4	4.2	4.3	6.5
▼ Commercial Models				
Qwen3.8-Max (Qwen Team, 2026a)	**66.4**	78.1	**81.7**	**75.0**
Kimi-K3 (Team et al., 2026b)	69.5	**78.6**	63.5	70.2
GPT-5.6-sol (OpenAI, 2026)	56.4	69.5	73.0	65.8
Claude-Opus-4.8 (Anthropic, 2026)	42.5	49.1	63.5	51.4
Gemini-3.1-Pro-Preview (Google, 2026)	39.8	57.1	43.9	46.4
▼ Large Open-Weight Models				
Qwen3.8-27B (Qwen Team, 2026c)	**48.7**	**65.2**	**74.8**	**62.2**
Qwen3.5-27B (Qwen Team, 2026b)	39.0	65.7	66.1	55.9
Qwen3.5-35B-A3B (Qwen Team, 2026b)	34.8	62.4	54.4	49.5
Gemma-4-31B-it (Team et al., 2026a)	30.5	51.0	51.3	43.5
Qwen3-VL-30B-A3B (Yang et al., 2025a)	22.4	21.9	15.7	20.0
InternVL3.5-30B-A3B (Wang et al., 2025b)	14.7	25.2	13.9	17.6
▼ Medium Open-Weight Models				
Qwen3.5-9B (Qwen Team, 2026b)	**29.7**	**59.5**	**61.7**	**49.2**
Step3-VL-10B (Huang et al., 2026)	25.1	43.8	40.9	35.9
Gemma-4-12B-it (Team et al., 2026a)	24.7	31.9	29.6	28.5
GLM-4.6V-Flash (Hong et al., 2025)	23.9	35.2	23.0	27.0
InternVL3.5-8B (Wang et al., 2025b)	19.3	28.1	12.2	19.6
Ovis2.5-9B (Lu et al., 2025)	20.9	24.3	13.9	19.6
NEO-9B-SFT (Diao et al., 2026a)	23.6	19.1	12.6	18.6
MiMo-VL-7B-SFT (Yue et al., 2025)	19.7	21.4	13.9	18.3
SenseNova-U1 (Diao et al., 2026c)	17.4	15.2	20.9	17.9
Qwen3-VL-8B-Instruct (Yang et al., 2025a)	17.8	19.5	16.5	17.9
NEO-ov (Diao et al., 2026b)	18.5	17.6	12.6	16.3
Ovis2-8B (Lu et al., 2025)	15.1	19.1	13.0	15.6
▼ Small Open-Weight Models				
Qwen3.5-4B (Qwen Team, 2026b)	**27.0**	**52.4**	**53.0**	**43.2**
Qwen3.5-2B (Qwen Team, 2026b)	15.1	24.3	22.2	20.2
Qwen3-VL-2B-Instruct (Yang et al., 2025a)	14.3	17.1	8.3	13.2
InternVL3.5-2B (Wang et al., 2025b)	13.9	18.6	5.7	12.6
NEO-2B-SFT (Diao et al., 2026a)	10.4	7.1	7.0	8.3

ate GPT-5.6-sol (OpenAI, 2026), Kimi-K3 (Team et al., 2026b), Gemini-3.1-pro-preview (Google, 2026), Claude-Opus-4.8 (Anthropic, 2026), and Qwen3.8-Max (Qwen Team, 2026a).

Evaluation. All open-source models are evaluated using VLMEvalKit (Duan et al., 2024) with the default decoding settings from their official repositories, running in BF16 precision on NVIDIA A800 GPUs. Proprietary models are evaluated through their official APIs under the same protocol. Further details on the evaluation prompts and model configurations are provided in the Appendix.

4.2 Main Results

Current VLMs show limited reasoning and planning capabilities toward embodied tasks. As shown in Table 1, current VLMs exhibit room for improvement on Plan-Bench. Among the 28 evaluated models, Qwen3.8-Max achieves the highest overall accuracy of 75.0%, followed by Kimi-K3 at 70.2%. All remaining models fall below 70%, including GPT-5.6-sol at 65.8%. Among the MoE, 7–12B, and 2–4B groups, the best-performing models are Qwen3.5-35B-A3B, Qwen3.5-9B,

	Spatial Cognition						Temporal Reasoning					Long-Horizon Planning				
	OO	MD	AR	TD	SR	SpR	PT	TO	PC	IA	MR	CS	CP	RP	DP	Route
Qwen3.8-Max	60.0	74.3	54.0	42.5	70.0	93.2	93.3	92.5	38.0	90.0	88.0	92.0	86.7	58.0	84.0	90.0
Kimi-K3	66.0	71.4	48.0	50.0	90.0	95.5	96.7	90.0	48.0	95.0	76.0	68.0	93.3	10.0	80.0	78.0
Qwen3.8-27B	46.0	54.3	38.0	32.5	57.5	65.9	100.0	75.0	34.0	67.5	66.0	82.0	90.0	66.0	80.0	62.0
Gemma-4-31B-it	22.0	37.1	28.0	27.5	35.0	36.4	93.3	75.0	24.0	17.5	60.0	68.0	56.7	20.0	68.0	46.0
Step3-VL-10B	10.0	45.7	30.0	7.5	35.0	27.3	80.0	55.0	12.0	35.0	52.0	58.0	66.7	12.0	50.0	28.0
GLM-4.6V-Flash	16.0	40.0	16.0	10.0	17.5	47.7	70.0	37.5	24.0	17.5	38.0	38.0	43.3	6.0	6.0	30.0
Ovis2.5-9B	20.0	17.1	12.0	20.0	27.5	29.6	13.3	32.5	16.0	12.5	42.0	22.0	43.3	2.0	8.0	6.0
Qwen3.5-2B	6.0	28.6	10.0	22.5	15.0	13.6	43.3	22.5	12.0	5.0	42.0	50.0	30.0	4.0	12.0	18.0

Figure 4: Fine-grained performance of 8 representative VLMs across the 16 subtasks in Plan-Bench. Each cell reports accuracy (%), with color intensity indicating performance.

and Qwen3.5-4B, achieving 49.5%, 49.2%, and 43.2% overall accuracy, respectively. Beyond these leading models, most of the remaining models in these groups cluster around 20% accuracy, with only a few reaching the high-20% to mid-30% range. These results indicate that reliable reasoning and planning over physical environments remains far from solved for current VLMs.

Spatial cognition remains challenging for current VLMs. As shown in Table 1, many models stay close to the 10.4% random baseline. For example, Qwen3-VL-2B-Instruct and InternVL3.5-2B achieve only 14.3% and 13.9%, while Qwen3.5-2B scores 15.1%. This weakness is not confined to the lightweight models, as Qwen3-VL-8B-Instruct and NEO-ov reach 17.8% and 18.5%. Across the full set of evaluated models, only nine models exceed 30% accuracy, while ten remain around or below 20%. Even among stronger frontier models, the highest accuracy is only 69.5%, achieved by Kimi-K3, followed by Qwen3.8-Max at 66.4%. These results show that current VLMs still struggle to reliably integrate spatial evidence across multiple viewpoints and complex scene configurations.

Model scaling generally benefits embodied reasoning and planning. Within the same model family and architecture, performance improves with scale. Qwen3.5, for example, increases from 20.2% at 2B to 49.2% at 9B and further to 55.9% at 27B. Similar trends are observed for Gemma-4 and InternVL3.5. Gemma-4 improves from 28.5% at 12B to 43.5% at 31B, while InternVL3.5 rises from 12.6% at 2B to 19.6% at 8B. However, this relationship is not consistent across model families. Qwen3.8-27B reaches 62.2%, higher than Gemini-3.1-Pro-Preview at 46.4%. Similarly, Qwen3.5-9B achieves 49.2%, outperforming the much larger Gemma-4-31B at 43.5%. Overall, larger models tend to perform better, while architecture and training strategy also remain important factors.

4.3 Ablation Study

Strong overall performance does not imply consistent fine-grained capabilities. Figure 4 reveals variation across the 16 subtasks. Even the strongest models exhibit weaknesses on individual tasks. Qwen3.8-Max, for example, exceeds 70% accuracy on several subtasks but reaches only 38.0% on Persistent Counting, while Kimi-K3 drops to 10.0% on Rearrangement Planning despite strong performance elsewhere. Models with similar overall accuracy may still differ across tasks. Qwen3.5-2B and Ovis2.5-9B achieve comparable overall scores, yet differ considerably across Progress Tracking, Spatial Relations, Constraint Sequencing, and Capacity Planning. Overall, strong or similar aggregate performance can still coexist with task-specific weaknesses and capability differences.

More views do not necessarily lead to better spatial cognition. Current VLMs remain susceptible to interference from misleading viewpoints. We reorganize *Spatial Relations* under the *Spatial Cognition* dimension into three settings: 1 right, 1 right + 2 mislead, and 1 right + 4 mislead, where additional views contain misleading cues. The results in Figure 5 show that all seven models perform well when only the informative view is provided, but their accuracy consistently declines as more misleading views are introduced. Larger models generally show greater robustness to such distractions. Qwen3.8-27B drops from 90.9% to 68.2%, while Qwen3.5-2B falls from 93.2% to 22.7%. Yet scale alone does not explain this robustness, as Gemma-4-31B-it is less robust than the smaller Qwen3.5-9B, suggesting that model architecture and training strategy may also play important roles.

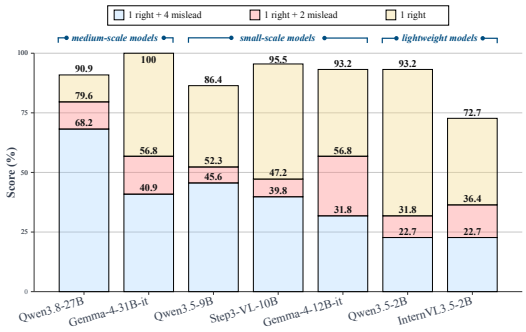

Figure 5: Robustness to Increasing Misleading Views across Model Scales.

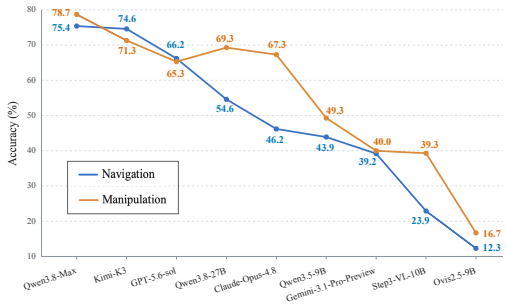

Figure 6: Navigation and Manipulation Performance across Different Models.

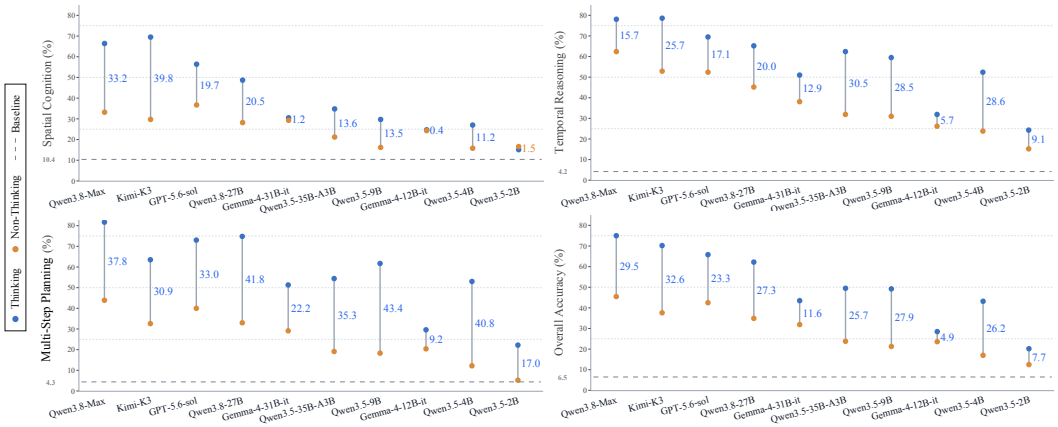

Figure 7: Comparison between Thinking and Non-Thinking modes on Plan-Bench across Spatial Cognition, Temporal Reasoning, Multi-Step Planning, and overall accuracy. Blue and orange markers denote Thinking and Non-Thinking results, while the annotated values indicate their gaps.

Thinking improves performance, with the largest gains in multi-step planning. We compare ten models that support both Thinking and Non-Thinking modes on Plan-Bench. As shown in Figure 7, Thinking improves nearly all dimension-level and overall results, with the only exception of Qwen3.5-2B, which drops by 1.5 percentage points on Spatial Cognition. Among the three dimensions, Thinking brings the largest gains on Multi-Step Planning, with an average improvement of 31.1 percentage points, compared with 19.4 on Temporal Reasoning and 15.2 on Spatial Cognition. At the overall level, all models benefit from Thinking, with improvements ranging from 4.9 percentage points on Gemma-4-12B-it to 32.6 points on Kimi-K3. Despite longer inference time, Thinking enables models to better handle reasoning and planning tasks in embodied environments.

Navigation and Manipulation are broadly aligned, with model-specific mismatches. We group Plan-Bench into Navigation-related (*Route Planning, Instruction Alignment, and Target Discovery*) and Manipulation-related (*Dependency Planning, Constraint Sequencing, and Occlusion Order*). As shown in Figure 6, most models exhibit similar trends. For example, Qwen3.8-Max, Kimi-K3, and GPT-5.6-sol show balanced performance on the two task groups. However, some models exhibit biases, with Qwen3.8-27B achieving about 69.3% on Manipulation task but 54.6% on Navigation task, while Step3-VL-10B shows a gap of roughly 15 percentage points. These results further indicate that larger models have greater potential to serve as high-level decision makers in physical environments, with more stable and consistent performance across diverse embodied tasks.

Hallucination is a source of reasoning and planning errors in the embodied world. We analyze Qwen3.8-Max, the best-performing model on Plan-Bench, focusing on failures in *Mirror Distinction, Persistent Counting, and Target Discovery*. As shown in Figure 8, the model fails to distinguish real objects from their reflections in *Mirror Distinction*, hallucinating reflections as physical entities.

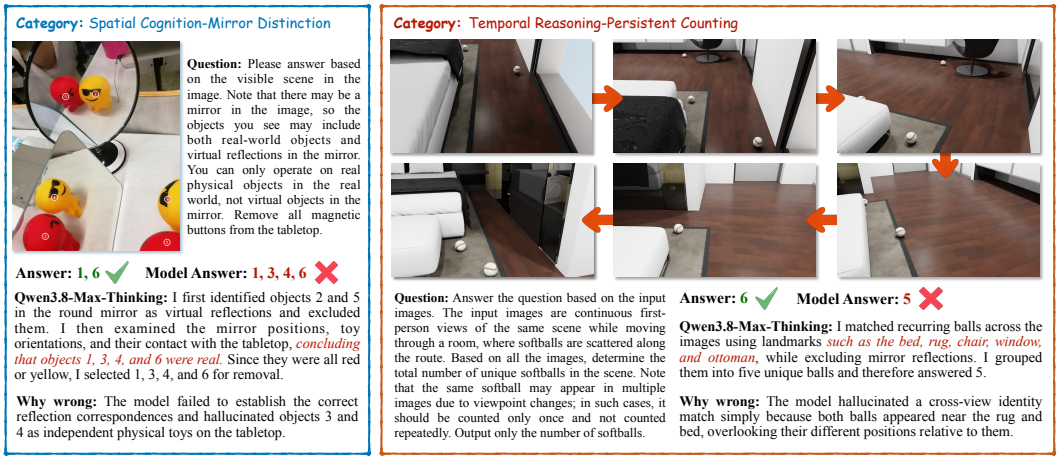

Figure 8: Representative Examples of Hallucination Errors by Qwen3.8-Max

In *Persistent Counting*, it incorrectly associates two distinct objects across views due to similar surrounding landmarks, resulting in an incorrect match. A similar phenomenon also appears in *Target Discovery*. This suggests that failures arise not only from visual recognition errors, but also from hallucinations during high-level reasoning and multi-step planning. Reducing such hallucinations may be important for improving the reliability of VLMs in embodied reasoning and planning.

5 Limitations and Discussion

We construct Plan-Bench to evaluate the reasoning and planning capabilities of VLMs in embodied tasks across Spatial Cognition, Temporal Reasoning, and Multi-step Planning, while several limitations remain to be further explored and addressed. *(i) Continually evolving frontier models.* Although we have evaluated diverse models, frontier systems evolve rapidly and performance may change across versions. Future work will track newly released models to assess progress in these capabilities. *(ii) Limited data scale and task coverage.* Plan-Bench includes 699 samples across 16 subcategories, while there is still room to expand both the dataset scale and task diversity. Further expansion will enable a more comprehensive evaluation of reasoning and planning. *(iii) Potential errors in LLM-as-judge.* Our LLM-as-judge achieves 100% agreement with human evaluation on 200 sampled cases, but some errors may still exist in the remaining samples. We will further explore more robust and programmatic evaluation procedures to improve evaluation accuracy and reliability.

6 Conclusion

We introduce Plan-Bench, a lightweight and systematic benchmark for evaluating the high-level reasoning and multi-step planning capabilities of VLMs in embodied scenarios independently of action execution. Plan-Bench covers Spatial Cognition, Temporal Reasoning, and Multi-Step Planning, spanning 16 subclasses and 699 curated samples. Evaluation of 28 VLMs shows that current models still have room for improvement, with the best-performing model achieving only 75.0% accuracy. Ablation studies further reveal that strong overall performance does not guarantee consistent capabilities across subtasks, and that models remain sensitive to misleading evidence and prone to hallucinations during reasoning. We also find that scaling generally improves results, while architecture and training may also influence performance, with Thinking bringing additional gains, particularly for Multi-Step Planning. Our results suggest that larger VLMs show greater potential as high-level planners, but reliable reasoning and planning across diverse embodied tasks remain challenging. Overall, Plan-Bench provides a principled and easy-to-deploy testbed for tracking progress in these capabilities and assessing VLMs as high-level decision makers for embodied AI.

AI USE STATEMENT

We used generative AI tools to assist with language polishing and the organization of research-related information during the preparation of this work. All AI-assisted content was reviewed, verified, and revised by the authors. Generative AI models used as evaluated models or as part of the evaluation pipeline are explicitly described in the main paper and Appendix. We take full responsibility for the final content of this work, including all text, code, and analyses.

ETHICS STATEMENT

Plan-Bench is constructed from publicly available datasets, controlled synthetic environments, and publicly accessible web resources. The benchmark does not involve human subjects or sensitive personal information and is intended solely for research on vision-language reasoning and planning.

REPRODUCIBILITY STATEMENT

We provide detailed task definitions, dataset statistics, data sources, prompt templates, LLM-as-judge settings, and model-specific evaluation configurations in the Appendix. Additional fine-grained results are also included to facilitate reproduction and further analysis of Plan-Bench.

REFERENCES

Anthropic. Introducing claude opus 4.8, 2026. URL `https://www.anthropic.com/news/claude-opus-4-8`.

Yinpei Dai, Hongze Fu, Jayjun Lee, Yuejiang Liu, Haoran Zhang, Jianing Yang, Chelsea Finn, Nima Fazeli, and Joyce Chai. Robomme: Benchmarking and understanding memory for robotic generalist policies. *arXiv preprint arXiv:2603.04639*, 2026.

Brian Danchilla. Three. js framework. In *Beginning WebGL for HTML5*, pp. 173–203. Springer, 2012.

Haiwen Diao, Mingxuan Li, Silei Wu, Linjun Dai, Xiaohua Wang, Hanming Deng, Lewei Lu, Dahua Lin, and Ziwei Liu. From pixels to words–towards native vision-language primitives at scale. In *International Conference on Learning Representations*, volume 2026, pp. 109909–109929, 2026a.

Haiwen Diao, Jiahao Wang, Penghao Wu, Yuhao Dong, Yuwei Niu, Yue Zhu, Zhongang Cai, Weichen Fan, Linjun Dai, Silei Wu, et al. From pixels to words–towards native one-vision models at scale. *arXiv preprint arXiv:2605.28820*, 2026b.

Haiwen Diao, Penghao Wu, Hanming Deng, Jiahao Wang, Shihao Bai, Silei Wu, Weichen Fan, Wenjie Ye, Wenwen Tong, Xiangyu Fan, et al. Sensenova-u1: Unifying multimodal understanding and generation with neo-unify architecture. *arXiv preprint arXiv:2605.12500*, 2026c.

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2694–2703. IEEE, 2023.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 11198–11201, 2024.

Zhenxuan Fan, Bo Zhang, Yutong Lin, Yuqian Yuan, Juekai Lin, Liang Liang, Zhuoyi Huang, Wenqiao Zhang, Juncheng Li, Siliang Tang, et al. Robospa: Can vla models go beyond simple scenes and short-horizon tasks? *arXiv preprint arXiv:2609.05324*, 2026.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Google. Gemini 3.1 pro model card, 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/. Model Card.

Songhao Han, Boxiang Qiu, Yue Liao, Siyuan Huang, Chen Gao, Shuicheng Yan, and Si Liu. Robocerebra: A large-scale benchmark for long-horizon robotic manipulation evaluation. *Advances in Neural Information Processing Systems*, 38, 2026.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Ailin Huang, Chengyuan Yao, Chunrui Han, Fanqi Wan, Hangyu Guo, Haoran Lv, Hongyu Zhou, Jia Wang, Jian Zhou, Jianjian Sun, et al. Step3-vl-10b technical report. *arXiv preprint arXiv:2601.09668*, 2026.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4392–4412, 2020.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *Conference on Robot Learning*, pp. 80–93. PMLR, 2023.

Hengzhi Li, Justin Zhang, Brendon Jiang, Alexander Naehu, Regan Song, Megan Tjandrasuwita, Chanakya Ekbote, Steven-Shine Chen, Adithya Balachandran, Wei Dai, et al. Puzzleworld: A benchmark for multimodal, open-ended reasoning in puzzlehunts. In *International Conference on Learning Representations*, volume 2026, pp. 11819–11843, 2026a.

Jiachun Li, Shaoping Huang, Zhuoran Jin, Chenlong Zhang, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Mmr-life: Piecing together real-life scenes for multimodal multi-image reasoning. *arXiv preprint arXiv:2603.02024*, 2026b.

Yi Li, Yuquan Deng, Jesse Zhang, Joel Jang, Marius Memmel, Caelan Garrett, Fabio Ramos, Dieter Fox, Anqi Li, Abhishek Gupta, et al. Hamster: Hierarchical action models for open-world robot manipulation. In *International Conference on Learning Representations*, volume 2025, pp. 24040–24068, 2025.

Jiaying Lin, Guodong Wang, and Rynson WH Lau. Progressive mirror detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3694–3702. IEEE, 2020.

Zhen Liu, Xinyu Ning, Zhe Hu, Xinxin Xie, Weize Li, Zhipeng Tang, Chongyu Wang, Zejun Yang, Hanlin Wang, Yitong Liu, et al. Goal2skill: Long-horizon manipulation with adaptive planning and reflection. *arXiv preprint arXiv:2604.13942*, 2026.

Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, et al. Ovis2. 5 technical report. *arXiv preprint arXiv:2508.11737*, 2025.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Tianshuo Yang, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, et al. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 38405–38453, 2025.

OpenAI. Introducing gpt-5.6, 2026. URL https://openai.com/zh-Hans-CN/index/gpt-5-6.

Rohith Peddi, Shivvrat Arya, Bharath Challa, Likhitha Pallapothula, Akshay Vyas, Bhavya Gouripeddi, Qifan Zhang, Jikai Wang, Vasundhara Komaragiri, Eric Ragan, et al. Captaincook4d: A dataset for understanding errors in procedural activities. *Advances in Neural Information Processing Systems*, 37:135626–135679, 2024.

Lu Qiu, Yi Chen, Yuying Ge, Yixiao Ge, Ying Shan, and Xihui Liu. Egoplan-bench2: A benchmark for multimodal large language model planning in real-world scenarios. *International Journal of Computer Vision*, 134(5):222, 2026.

Qwen Team. Introducing qwen3.8-max, 2026a. URL https://qwen.ai/blog?id=qwen3.8.

Qwen Team. Qwen3.5, 2026b. URL https://huggingface.co/collections/Qwen/qwen35. Official Model Card.

Qwen Team. Qwen3.8-27B, 2026c. URL https://huggingface.co/Qwen/Qwen3.8-27B. Official Model Card.

Lucy Xiaoyang Shi, Brian Ichter, Michael Equi, Liyiming Ke, Karl Pertsch, Quan Vuong, James Tanner, Anna Walling, Haohuan Wang, Niccolo Fusai, et al. Hi robot: Open-ended instruction following with hierarchical vision-language-action models. *arXiv preprint arXiv:2502.19417*, 2025.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026a.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026b.

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. In *International Conference on Learning Representations*, volume 2025, pp. 62624–62650, 2025a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Yuechen Xie, Xiaoyan Zhang, Yicheng Shan, Zhu Hao, Rui Tang, Rong Wei, Mingli Song, Yuanyu Wan, and Jie Song. Spatialqa: A benchmark for evaluating spatial logical reasoning in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2645–2657, 2026.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026a.

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, et al. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. In *International Conference on Learning Representations*, volume 2026, pp. 25966–26003, 2026b.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, et al. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. *arXiv preprint arXiv:2502.09560*, 2025b.

Tianshuo Yang, Guanyu Chen, Yutian Chen, Zhixuan Liang, Yitian Liu, Zanxin Chen, Chunpu Xu, Haotian Liang, Jiangmiao Pang, Yao Mu, et al. Hivla: A visual-grounded-centric hierarchical embodied manipulation system. *arXiv preprint arXiv:2604.14125*, 2026.

Xin Yang, Haiyang Mei, Ke Xu, Xiaopeng Wei, Baocai Yin, and Rynson WH Lau. Where is my mirror? In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8809–8818, 2019.

Zhutian Yang, Caelan Garrett, Dieter Fox, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. Guiding long-horizon task and motion planning with vision language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 16847–16853. IEEE, 2025c.

Ruilin Yao, Bo Zhang, Jirui Huang, Xinwei Long, Yifang Zhang, Tianyu Zou, Shili Xiong, Yi Rong, Yufei Wu, Shichao Su, et al. Lens: Multi-level evaluation of multimodal reasoning with large language models. In *International Conference on Learning Representations*, volume 2026, pp. 2627–2654, 2026.

Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, et al. Mimo-vl technical report. *arXiv preprint arXiv:2506.03569*, 2025.

Yixian Zhang, Huanming Zhang, Feng Gao, Xiao Li, Zhihao Liu, Chunyang Zhu, Jiaxing Qiu, Yuchen Yan, Jiyuan Liu, Wenhao Tang, et al. Harness vla: Steering frozen vlas into reliable manipulation primitives via memory-guided agents. *arXiv preprint arXiv:2607.08448*, 2026.

Yiyang Zhou, Haoqin Tu, Zijun Wang, Zeyu Wang, Niklas Muennighoff, Fan Nie, Chaorui Deng, Shen Yan, Haoqi Fan, Yejin Choi, et al. When visualizing is the first step to reasoning: Mira, a benchmark for visual chain-of-thought. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26154–26164, 2026.

Zilin Zhu, Longteng Guo, Yanghong Mei, Bowen Pang, Zongxun Zhang, Xingjian He, Ruyi Ji, and Jing Liu. When robots do the chores: A benchmark and agent for long-horizon household task execution. *arXiv preprint arXiv:2605.14504*, 2026.

A BENCHMARK DETAILS

A.1 TASK DEFINITIONS

Plan-Bench consists of 16 subtasks that instantiate different reasoning and planning challenges in embodied scenarios. Below, we provide detailed definitions of each subtask and clarify the specific capability evaluated by each task.

Occlusion Order (OO). Given a scene containing partially occluded objects, the model is required to infer their relative ordering from visual occlusion cues. This task evaluates whether the model can reason about object ordering when complete object boundaries are not directly observable.

Mirror Distinction (MD). The model must distinguish real objects from their reflections and correctly reason about the scene content. The task is designed to evaluate whether the model can distinguish real objects from reflections and correctly handle reflection-induced visual ambiguity.

Action Recovery (AR). Given partial information from one viewpoint, the model is asked to recover the corresponding information that is missing or unclear from another viewpoint. This task examines whether the model can compare observations across multiple views and integrate complementary visual cues to reconstruct the complete scene information.

Target Discovery (TD). The model must reason about the target's spatial relationship under mirror reflections and plan a feasible path toward it. This task evaluates spatial reasoning and path planning in reflection-ambiguous scenes.

State Recognition (SR). Given multiple visual observations of a scene, the model must integrate information while identifying redundantly observed content. This task evaluates multi-view scene understanding and the ability to distinguish new evidence from repeated information.

Spatial Relations (SpR). The model must infer spatial relationships between specified entities from one or multiple views. The task evaluates spatial reasoning and the ability to integrate relevant evidence while ignoring misleading observations.

Progress Tracking (PT). Given a sequence depicting an ongoing activity, the model is required to determine the current progress or stage of the activity. This task evaluates whether the model can track how an embodied procedure evolves over time.

Temporal Ordering (TO). The model must determine the correct temporal order of events, actions, or scene states from sequential visual evidence. The task evaluates the ability to recover temporal dependencies and distinguish earlier events from later ones.

Persistent Counting (PC). The model is asked to count target entities across a visual sequence while maintaining their identity over time. This task evaluates whether the model can accumulate information across frames without repeatedly counting the same persistent entity.

Instruction Alignment (IA). Given an instruction together with a sequence of visual observations, the model must identify the observation that corresponds to the specified instruction. The task evaluates the alignment between language-described actions and their visual realization over time.

Motion Retrieval (MR). The model is required to identify the motion or action segment that matches a specified query from a temporal sequence. This task evaluates whether the model can recognize and retrieve task-relevant motion patterns among multiple observed actions.

Constraint Sequencing (CS). Given a set of actions together with ordering constraints, the model must construct a valid action sequence satisfying all specified constraints. The task evaluates multi-step planning under explicit requirements on the relative order of actions.

Capacity Planning (CP). The model is required to produce a feasible plan under capacity limitations on containers or other task-relevant entities. This task evaluates whether the model can jointly reason about action ordering and limited capacity when constructing a multi-step solution.

Rearrangement Planning (RP). Given an initial and a target configuration, the model must determine a sequence of actions that transforms the scene into the desired arrangement. The task evaluates planning capabilities and the consequences of successive rearrangement operations.

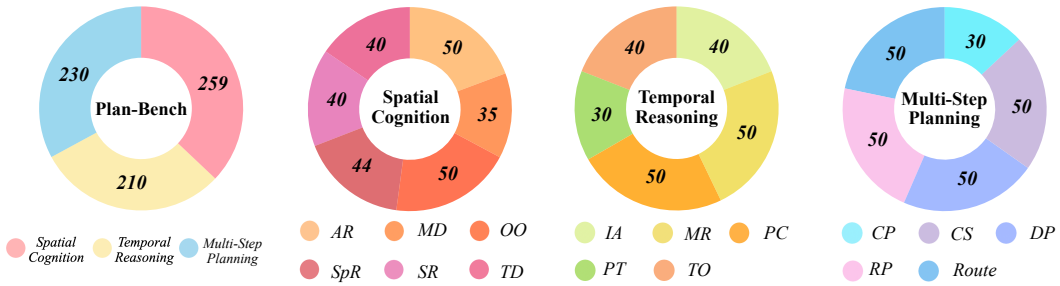

Figure 9: Dataset composition of Plan-Bench across three dimensions and 16 subtasks.

Table 2: Data sources and question-type distribution across the 16 subtasks in Plan-Bench.

Task	Subtask	Single	Open	Data Source
Spatial Cognition	Action Recovery	30	20	Robomme (Dai et al., 2026)
	Mirror Distinction	0	35	MSD (Yang et al., 2019) PMD (Lin et al., 2020)
	Occlusion Order	18	32	three.js (Danchilla, 2012)
	Spatial Relations	0	44	three.js (Danchilla, 2012)
	State Recognition	20	20	three.js (Danchilla, 2012)
	Target Discovery	40	0	PMD (Lin et al., 2020)
Temporal Reasoning	Instruction Alignment	15	25	RxR (Ku et al., 2020)
	Motion Retrieval	0	50	mevis (Ding et al., 2023)
	Persistent Counting	0	50	OmniGibson (Li et al., 2023)
	Progress Tracking	0	30	captaincook4d (Peddi et al., 2024)
	Temporal Ordering	20	20	captaincook4d (Peddi et al., 2024)
Multi-Step Planning	Capacity Planning	30	0	three.js (Danchilla, 2012)
	Constraint Sequencing	0	50	CLEVR (Johnson et al., 2017)
	Dependency Planning	0	50	Website SpatialQA (Xie et al., 2026)
	Rearrangement Planning	0	50	OmniGibson (Li et al., 2023)
	Route Planning	10	40	OmniGibson (Li et al., 2023)

Dependency Planning (DP). The model must construct an action plan while respecting prerequisite relations among actions or subgoals. This task evaluates whether the model can identify dependencies and ensure that required preconditions are satisfied before executing dependent actions.

Route Planning (Route). Given spatial information about an environment, the model must determine a feasible route from the current location to a specified destination. The task evaluates multi-step navigation planning under spatial connectivity, obstacle, and path-feasibility constraints.

A.2 Dataset Statistics

Plan-Bench maintains a balanced distribution across dimensions and subtasks. As shown in Figure 9, Plan-Bench contains 699 samples distributed across three dimensions, with 259 samples for Spatial Cognition, 210 for Temporal Reasoning, and 230 for Multi-Step Planning. The samples are further distributed across 16 subtasks, with most subtasks containing between 30 and 50 instances. Overall, the benchmark maintains a relatively balanced distribution across both dimensions and subtasks, avoiding excessive concentration on any single capability.

Plan-Bench exhibits diversity in visual inputs. As shown in Tables 3 and 4, the benchmark covers samples containing from 1 to 16 images, with 64.9% of the samples involving multiple images and 47.2% containing six or more images. The 3,007 images also span a wide range of resolutions, from below 0.5 MP to above 4 MP, with 35.8% exceeding 1 MP. These statistics highlight the diversity of Plan-Bench in both visual context length and image resolution, rather than relying on a fixed input configuration.

Table 3: Distribution of the number of images per sample.

Images	Samples	Ratio
1	245	35.1%
2–5	124	17.7%
6–7	181	25.9%
8–16	149	21.3%
Total	**699**	**100%**

Table 4: Distribution of image resolutions.

Resolution	Images	Ratio
$\leq$0.5 MP	1,430	47.6%
0.5–1 MP	502	16.7%
1–2 MP	363	12.1%
2–4 MP	643	21.4%
$\geq$4 MP	69	2.3%
Total	**3,007**	**100%**

Table 5: Distribution of planning lengths in the Multi-Step Planning subset of Plan-Bench.

Steps	Samples	Percentage
1–3 steps	70	30.43%
4–6 steps	131	56.96%
7–10 steps	29	12.61%
Total	**230**	**100%**

To further characterize the planning complexity of the Multi-Step Planning subset, we analyze the number of steps required by the reference plans. As shown in Table 5, the 230 samples span planning horizons ranging from 1 to 10 steps. Among them, 70 samples (30.43%) require 1–3 steps, while the majority of the samples, 131 (56.96%), involve 4–6 planning steps. In addition, 29 samples (12.61%) require relatively longer plans of 7–10 steps. This distribution indicates that Plan-Bench covers planning problems with varying horizon lengths, while placing substantial emphasis on multi-step tasks that require models to reason over several interdependent actions rather than making only short-horizon decisions.

A.3 Dataset Filtering Procedure

We perform a multi-stage filtering procedure to improve the quality and reliability of Plan-Bench. Starting from 2,580 candidate QA pairs, we remove samples that fail one or more quality requirements and finally obtain 699 validated samples. The entire filtering process is conducted before model evaluation and is independent of model performance.

Each candidate sample is examined according to five quality considerations: (1) visual sufficiency, where the answer must be determinable from the provided visual evidence; (2) annotation correctness, including the validity of reference answers and answer choices; (3) capability alignment, ensuring that the sample measures the intended reasoning or planning ability; (4) redundancy control, which removes highly correlated samples originating from similar scenes, videos, or viewpoints; and (5) other quality issues identified during manual review. The detailed statistics of removed samples are reported in Table 7.

For samples collected from videos or multi-view environments, redundancy filtering is performed at the scene level rather than only at the QA level. Samples sharing the same underlying scene, highly similar viewpoints, or overlapping temporal segments are grouped together. When multiple samples evaluate equivalent reasoning patterns, only the highest-quality instance is retained to avoid over-representing specific environments.

To examine whether filtering introduces unintended changes in benchmark composition, we compare the category distribution before and after filtering. As shown in Table 6, the retention ratios remain similar across Spatial Cognition, Temporal Reasoning, and Multi-Step Planning. The final benchmark maintains coverage across all three dimensions and 16 subtasks while removing low-quality or redundant samples.

Table 6: Category distribution before and after sample filtering.

Category	Before Filtering	Removed	Retained
Spatial Cognition	952	693	259
Temporal Reasoning	768	558	210
Multi-Step Planning	860	630	230
Total	**2580**	**1881**	**699**

Table 7: Statistics of sample filtering during Plan-Bench construction.

Filtering Reason	Removed Samples	Percentage
Insufficient visual evidence or ambiguous answers	637	33.9%
Incorrect annotation or reference answer	364	19.4%
Capability mismatch with target task	297	15.8%
Duplicate or highly correlated scenes	512	27.2%
Other quality issues	71	3.8%
Total removed	**1881**	**72.91%**
Final retained	**699**	**27.09%**

A.4 Data Sources

Plan-Bench combines diverse data sources and question formats across subtasks. As shown in Table 2, the 16 subtasks are constructed from a mixture of existing datasets, controlled generation tools, simulation environments, and web-sourced data. Different sources are selected according to the capability targeted by each subtask, ranging from temporal activities to spatial reasoning and multi-step planning. The benchmark also includes single-choice, open-ended, and mixed-format subtasks, providing diverse evaluation settings rather than relying on a uniform question format.

B Evaluation Protocol

B.1 Model Prompt Templates

Since Plan-Bench contains both single-choice and open-ended questions, we implement a unified prompting function to support both question types within the same evaluation pipeline. As shown below, all samples share a common *Base Prompt*, which specifies the general response behavior and output requirements. We then append a question-type-specific instruction according to the sample format: the *Single-Choice Prompt* requires the model to return only one option letter, while the *Open-Ended Prompt* requires only the requested structured answer. This design provides a consistent prompting framework while preserving the output requirements of different question types.

Base Prompt

Answer directly. Do not explain your reasoning. For single-choice questions, answer with one option letter only, such as A. For open questions, follow the required answer format exactly.

Single-Choice Prompt

This is a single-choice question. Return one option letter only.

Open-Ended Prompt

This is an open question. Return only the requested structured answer.

B.2 LLM-as-Judge Prompt

We use DeepSeek-V4-Flash as an LLM-based judge to evaluate model responses. For each sample, the judge receives the model output, ground-truth answer, and an optional answer hint specifying additional valid alternatives or equivalence conditions. A response is labeled `True` when it is semantically equivalent to the ground-truth answer or satisfies an equivalence condition explicitly defined by the answer hint; otherwise, it is labeled `False`. The judge is instructed to evaluate only answer correctness and to return no additional explanation.

System Prompt (Answer Equivalence Judge)

You are a precise answer-equivalence judge for evaluating model responses.
Your task is to determine whether a model output should be considered correct with respect to the ground-truth answer and the provided answer hint.
The **ground-truth answer** specifies the reference answer, while the **answer hint** provides additional valid alternatives or equivalence conditions that should be accepted.
Judge the model output according to the following rules:

1. **Semantic Equivalence:** Return `True` if the model output expresses the same answer as the ground-truth answer, even if the wording or formatting differs.
2. **Answer-Hint Equivalence:** Return `True` if the model output satisfies any valid alternative or equivalence condition explicitly specified in the answer hint, including acceptable order swaps, position exchanges, alternative object names, or equivalent action descriptions.
3. **Partial Correctness:** Return `False` if the response is only partially correct, omits required information, or contains an incorrect component.
4. **Contradictory or Additional Content:** Return `False` if the response contains information that contradicts the ground-truth answer or violates the conditions specified by the answer hint.

5. **Formatting Differences:** Ignore superficial differences in capitalization, punctuation, spacing, or other formatting when they do not change the meaning of the answer.

Do not introduce alternative interpretations beyond those supported by the ground-truth answer and answer hint. Do not infer missing information. Do not evaluate the reasoning process; evaluate only the final answer.
Return only one of the following:
`True`
or
`False`
Do not provide any explanation or additional text.

User Prompt (Answer Verification)

Determine whether the model output is equivalent to the ground-truth answer under the conditions specified by the answer hint.
Ground-truth answer:
{`ground_truth`}
Answer hint:
{`answer_hint`}
Model output:
{`model_output`}
Return only `True` or `False`.

B.3 Answer Hint and Equivalence Rules

Answer hints handle multiple valid answers in open-ended questions. Among the open-ended questions in Plan-Bench, 486 samples require answer hints across 13 subcategories, including *Occlusion Order*, *Mirror Distinction*, *State Recognition*, *Rearrangement Planning*, and *Dependency Planning*. These hints specify valid equivalence conditions beyond the reference answer, such as interchangeable object references, flexible orderings, position exchanges, and equivalent action sequences. As illustrated in Figure 10, the answer hint is provided only to the LLM judge, allowing semantically equivalent predictions to be accepted without exposing additional information to the evaluated model.

B.4 Human Verification of the Judge

To further assess the reliability of our LLM-based evaluation, we conduct a human verification study on 200 responses generated by Qwen3.8-Max, Gemma-4-31B-it, and Step3-VL-10B. We adopt stratified sampling to ensure diverse and balanced coverage across the three evaluation dimensions and their corresponding subclasses, selecting 67 samples from *Spatial Cognition*, 67 from *Temporal Reasoning*, and 66 from *Multi-Step Planning*. Among the 200 sampled instances, 133 contain answer hints specifying acceptable alternatives or equivalence conditions, such as alternative object names, interchangeable orderings, or equivalent action sequences, while the remaining 67 have no answer hints. All 200 responses are independently evaluated by human annotators and DeepSeek-V4-Flash based on the same ground-truth answers and answer hints. The two evaluations agree on all 200 samples, achieving 100% agreement, including the 133 cases involving answer hints. These results demonstrate that our LLM-based judge can reliably account for equivalent valid answers and provide consistent evaluation for Plan-Bench.

B.5 Model Configurations

We evaluate open-weight VLMs through VLMEvalKit using their corresponding model wrappers and checkpoints. Unless specified below, we retain the decoding and image-processing settings defined by each wrapper. Open-weight models are evaluated on NVIDIA A800 GPUs in BF16 precision. All models receive the prompt templates described in Appendix B.1, while proprietary models are evaluated through API interfaces under the same question and answer protocol.



Figure 10: Examples of answer hints for multiple valid answers.

Qwen series. Qwen3-VL-2B-Instruct, Qwen3-VL-8B-Instruct, and Qwen3-VL-30B-A3B-Instruct use the Qwen3VLChat wrapper with custom prompts disabled. Their configurations set temperature to 0.7, top-p to 0.8, top-k to 20, presence penalty to 1.5, and the maximum number of new tokens to 16,384. We use the same wrapper for Qwen3.5-2B, 4B, 9B, 27B, and 35B-A3B, as well as Qwen3.8-27B, with temperature 1.0, top-p 0.95, top-k 20, and presence penalty 1.5. For these Qwen3.5 and Qwen3.8 models, the initial generation limit is 16,384 tokens. After the initial generation, we instruct each model to summarize its generated response and present the final answer within an additional 512-token budget. This additional budget is used only for answer condensation, extraction, and formatting based on the already generated content, rather than for further substantive reasoning or revising the underlying answer. Where supported, Thinking and Non-Thinking evaluations use the same checkpoint with the corresponding mode selected in the wrapper.

Gemma-4 series. Gemma-4-12B-it and Gemma-4-31B-it use the Gemma4 wrapper with the corresponding model processor and chat template. The wrapper uses greedy decoding with a maximum of 16,384 new tokens. For Thinking and Non-Thinking evaluations, we set the chat template's `enable_thinking` option accordingly and extract the final answer from the generated response.

InternVL3.5 series. InternVL3.5-2B, InternVL3.5-8B, and InternVL3.5-30B-A3B use the InternVLChat V2.0 wrapper with the vLLM backend enabled. We retain the wrapper's default image processing and generation settings for these models.

Step3-VL series. Step3-VL-10B uses a local vLLM inference wrapper with greedy decoding, a maximum of 16,384 new tokens, and support for up to 24 images per prompt. Its DeepSeek-R1-style parser separates the reasoning trace from the final answer; only the final answer is submitted for evaluation.

GLM-4.6V series. GLM-4.6V-Flash uses a local vLLM wrapper with Thinking enabled. Its configuration sets temperature to 0.8, top-p to 0.6, top-k to 2, repetition penalty to 1.1, and the maximum number of new tokens to 16,384.

Ovis series. Ovis2-8B and Ovis2.5-9B are evaluated using their corresponding VLMEvalKit wrappers. We retain their wrapper-defined image preprocessing and generation settings.

MiMo-VL series. MiMo-VL-7B-SFT uses the Qwen2VLChat wrapper with custom prompts disabled and vLLM inference. Its image-size bounds are set to $1280 \times 28 \times 28$ and $16384 \times 28 \times 28$ pixels; the remaining generation settings follow the wrapper configuration.

NEO series. NEO-2B-SFT and NEO-9B-SFT use the NEOChat wrapper with a patch size of 16, a downsampling ratio of 0.5, image-size bounds of $1280 \times 32 \times 32$ and $4096 \times 32 \times 32$ pixels, and the maximum number of new tokens to 16,384. We retain the wrapper's default generation settings.

NEO-ov is evaluated under the same benchmark protocol, although its model-specific configuration is not included in the supplied files.

SenseNova series. SenseNova-U1 is evaluated under the common Plan-Bench prompting and answer protocol, and the maximum number of new tokens to 16,384. Its model-specific inference settings are not specified in the supplied configuration.

Qwen3.8-max. Qwen3.8-max is evaluated through an API interface with a maximum response length of 16,384 tokens. The supplied configuration does not specify a temperature override.

GPT-5.6-sol. GPT-5.6-sol is evaluated through an API interface with a maximum response length of 16,384 tokens. The supplied configuration does not specify a temperature override.

Gemini-3.1-Pro-Preview. Gemini-3.1-Pro-Preview is evaluated through an API interface with temperature set to 0 and a maximum response length of 16,384 tokens.

Claude-Opus-4.8. Claude-Opus-4.8 is evaluated through an API interface with temperature set to 0 and a maximum response length of 16,384 tokens.

Kimi-K3. Kimi-K3 is evaluated under the common Plan-Bench prompting and answer protocol. Its model-specific API settings are not specified in the supplied configuration files.

C Additional Results

C.1 Results on All 16 Subtasks

Tables 8–10 report model accuracy on all 16 subtasks of Plan-Bench, revealing differences that are hidden by the three dimension-level scores. In Spatial Cognition, GPT-5.6-sol achieves 82.9% on Mirror Distinction and 87.5% on State Recognition, but only 30.0% on Target Discovery and 31.3% on Spatial Relations. Gemini-3.1-Pro-Preview shows a contrasting profile, reaching 75.0% on Spatial Relations but only 22.0% on Occlusion Order. Qwen3.5-27B also varies sharply across the category, with 56.8% on Spatial Relations but 18.0% on Occlusion Order. These contrasts indicate that success on one form of spatial evidence does not reliably extend to the other visual reasoning demands within the same dimension, even when the tasks draw on similar visual scenes and viewpoints that expose different but complementary scene cues about object locations.

Temporal Reasoning likewise exhibits substantial variation between subtasks. Qwen3.5-27B achieves 96.7% on Progress Tracking and 80.0% on Temporal Ordering, but only 28.0% on Persistent Counting. GPT-5.6-sol follows a similar pattern, scoring 93.3% on Progress Tracking and 30.0% on Persistent Counting. Even Qwen3.5-35B-A3B, which reaches 40.0% on Persistent Counting, performs much better on Progress Tracking at 86.7%. These gaps suggest that inferring task progress and event order is easier for these models than maintaining accurate counts when objects or actions recur across visual sequences and must be distinguished from previous observations or distractors that appear between relevant activity steps in multi-frame observations.

In Multi-Step Planning, strong results on one type of constraint do not consistently transfer to other planning problems. Gemini-3.1-Pro-Preview scores 93.3% on Capacity Planning but only 10.0% on Rearrangement Planning. Similarly, Qwen3.5-27B scores 90.0% on Capacity Planning but 44.0% on Route Planning. Rearrangement Planning is particularly difficult for several smaller models: InternVL3.5-8B and NEO-9B-SFT both score 0.0% despite obtaining 30.0% and 53.3% on Capacity Planning, respectively. Thus, the detailed results expose task-specific weaknesses in spatial reasoning, temporal tracking, and multi-step planning that aggregate accuracy cannot capture.

C.2 Additional Error Analysis

Figure 11 presents two additional Target Discovery failures by Qwen3.8-Max. In both cases, the target appears in a mirror, so selecting the correct action requires the model to associate the reflection with a real object and infer that object's location. In the first example, the model notices the similarity between the reflected picture and the directly visible picture, but speculates that another identical picture exists elsewhere. It therefore selects an action that leads away from the actual target. In the second example, the model likewise treats the reflected and visible paper towel dispensers as two separate objects, incorrectly concluding that the target is behind the camera.

These examples show how hallucination can arise when the model fails to establish the correspondence between a reflection and a physical object. Although it recognizes their visual similarity, it hallucinates a second, identical object and reasons about its nonexistent location. This mistaken interpretation then leads to an incorrect navigation decision, showing how object hallucination can propagate from scene understanding to action planning in Target Discovery.

Table 8: Performance of different VLMs on the six Spatial Cognition subtasks of Plan-Bench.

Model	OO	MD	AR	TD	SR	SpR
▼ Commercial Models						
GPT-5.6-sol	58.0	82.9	54.0	30.0	87.5	31.3
Claude-Opus-4.8	64.0	74.3	28.0	27.5	42.5	22.7
Gemini-3.1-Pro-Preview	22.0	45.7	28.0	30.0	42.5	75.0
▼ Large Open-Weight Models						
Qwen3.5-27B	18.0	54.3	36.0	27.5	47.5	56.8
Qwen3.5-35B-A3B	12.0	51.4	20.0	35.0	50.0	50.0
Qwen3-VL-30B-A3B	18.0	42.9	24.0	7.5	25.0	20.5
InternVL3.5-30B-A3B	6.0	2.9	24.0	10.0	20.0	22.7
▼ Medium Open-Weight Models						
Qwen3.5-9B	6.0	45.7	18.0	30.0	35.0	52.3
Gemma-4-12B-it	10.0	37.1	24.0	25.0	22.5	34.1
InternVL3.5-8B	22.0	14.3	20.0	15.0	15.0	27.3
NEO-9B-SFT	18.0	0.0	30.0	20.0	22.5	43.2
MiMo-VL-7B-SFT	14.0	40.0	14.0	15.0	17.5	22.7
SenseNova-U1	8.0	8.6	0.0	20.0	45.5	30.0
Qwen3-VL-8B-Instruct	24.0	11.4	8.0	22.5	20.0	20.5
NEO-ov	24.0	0.0	16.0	12.5	17.5	36.4
Ovis2-8B	8.0	0.0	14.0	20.0	20.0	27.3
▼ Small Open-Weight Models						
Qwen3.5-4B	6.0	40.0	20.0	17.5	40.0	45.5
Qwen3-VL-2B-Instruct	16.0	2.9	16.0	12.5	15.0	20.5
InternVL3.5-2B	18.0	0.0	16.0	15.0	10.0	20.5
NEO-2B-SFT	10.0	2.9	10.0	10.0	12.5	15.9

Table 9: Performance of different VLMs on the five Temporal Reasoning subtasks of Plan-Bench.

Model	PT	TO	PC	IA	MR
▼ Commercial Models					
GPT-5.6-sol	93.3	75.0	30.0	80.0	82.0
Claude-Opus-4.8	93.3	50.0	34.0	40.0	44.0
Gemini-3.1-Pro-Preview	86.7	67.5	34.0	47.5	62.0
▼ Large Open-Weight Models					
Qwen3.5-27B	96.7	80.0	28.0	72.5	68.0
Qwen3.5-35B-A3B	86.7	80.0	40.0	55.0	62.0
Qwen3-VL-30B-A3B	40.0	10.0	10.0	17.5	36.0
InternVL3.5-30B-A3B	33.3	22.5	10.0	20.0	42.0
▼ Medium Open-Weight Models					
Qwen3.5-9B	93.3	72.5	28.0	62.5	58.0
Gemma-4-12B-it	80.0	7.5	14.0	22.5	48.0
InternVL3.5-8B	26.7	25.0	24.0	35.0	30.0
NEO-9B-SFT	10.0	15.0	20.0	7.5	36.0
MiMo-VL-7B-SFT	13.3	22.5	12.0	17.5	38.0
SenseNova-U1	30.0	7.5	8.0	7.5	26.0
Qwen3-VL-8B-Instruct	20.0	20.0	8.0	15.0	34.0
NEO-ov	13.3	12.5	14.0	17.5	28.0
Ovis2-8B	30.0	15.0	14.0	12.5	26.0
▼ Small Open-Weight Models					
Qwen3.5-4B	86.7	67.5	24.0	40.0	58.0
Qwen3-VL-2B-Instruct	13.3	15.0	14.0	10.0	30.0
InternVL3.5-2B	10.0	17.5	18.0	15.0	28.0
NEO-2B-SFT	10.0	15.0	2.0	2.5	8.0

Table 10: Performance of different VLMs on the five Multi-Step Planning subtasks of Plan-Bench.

Model	CS	CP	RP	DP	Route
▼ Commercial Models					
GPT-5.6-sol	52.0	86.7	62.0	86.0	84.0
Claude-Opus-4.8	70.0	90.0	34.0	68.0	66.0
Gemini-3.1-Pro-Preview	42.0	93.3	10.0	56.0	40.0
▼ Large Open-Weight Models					
Qwen3.5-27B	82.0	90.0	62.0	62.0	44.0
Qwen3.5-35B-A3B	62.0	86.7	54.0	52.0	30.0
Qwen3-VL-30B-A3B	24.0	33.3	10.0	18.0	0.0
InternVL3.5-30B-A3B	26.0	36.7	0.0	6.0	10.0
▼ Medium Open-Weight Models					
Qwen3.5-9B	72.0	86.7	50.0	70.0	40.0
Gemma-4-12B-it	38.0	26.7	6.0	54.0	22.0
InternVL3.5-8B	22.0	30.0	0.0	10.0	6.0
NEO-9B-SFT	10.0	53.3	0.0	8.0	8.0
MiMo-VL-7B-SFT	24.0	26.7	2.0	4.0	18.0
SenseNova-U1	24.0	30.0	0.0	36.0	18.0
Qwen3-VL-8B-Instruct	26.0	36.7	8.0	8.0	12.0
NEO-ov	14.0	50.0	0.0	4.0	10.0
Ovis2-8B	16.0	50.0	0.0	4.0	10.0
▼ Small Open-Weight Models					
Qwen3.5-4B	72.0	80.0	28.0	56.0	40.0
Qwen3-VL-2B-Instruct	10.0	30.0	0.0	2.0	8.0
InternVL3.5-2B	8.0	20.0	0.0	2.0	4.0
NEO-2B-SFT	2.0	30.0	0.0	8.0	4.0

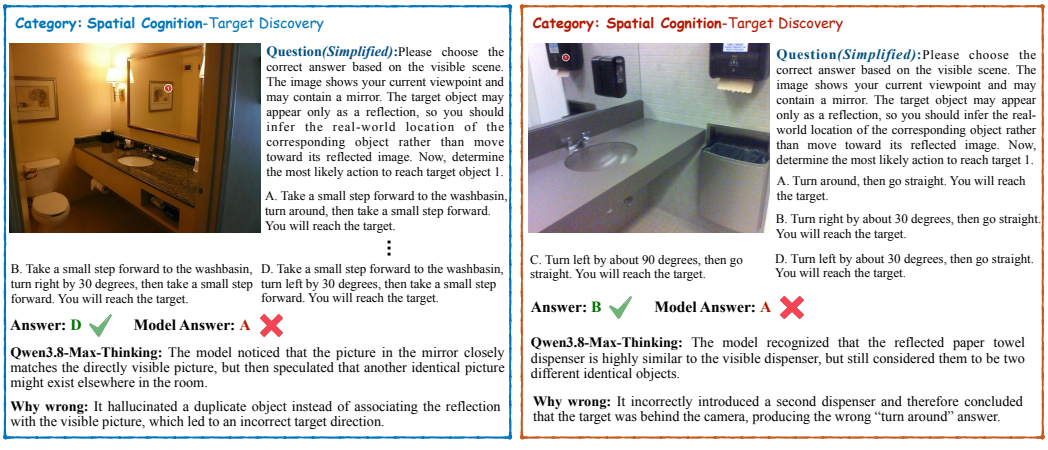

Figure 11: Additional Target Discovery failures by Qwen3.8-Max. The model treats reflected targets as separate physical objects, leading to incorrect navigation decisions.