NAVDBENCH: A DIAGNOSTIC BENCHMARK FOR CORE CAPABILITIES OF VISION-LANGUAGE NAVIGATION AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing vision-language navigation (VLN) benchmarks primarily measure end-task success, providing limited insight into which capabilities constrain an agent when navigation fails. We introduce NavDBench, a diagnostic benchmark that combines automatically generated navigation behaviors with controlled interventions to characterize four capabilities essential for reliable VLN: Visual Grounding, Trajectory Planning, Multi-Modal Reasoning, and Long-Horizon Memory. Rather than assigning individual tasks to isolated capabilities, NavDBench suppresses competing requirements through controlled variations in geometry, target distribution, language formulation, and instruction structure to construct capability-dominant embodied evaluations. Across five representative VLN agents, we find that visual grounding remains unreliable and sensitive to target distributions, geometric planning is a major short-horizon bottleneck, and the evaluated language variations have comparatively limited impact. Long-horizon failure reflects both accumulated local errors and difficulty maintaining task progress. Navigation-specific post-training substantially improves local capabilities across several diagnostic conditions, but these gains largely fail to propagate to autonomous exploration and sequential navigation. Our results highlight capability composition, rather than local competence alone, as a central challenge for reliable VLN.

1 INTRODUCTION

Vision-language navigation (VLN) has rapidly evolved from task-specific navigation systems toward generalist embodied agents built upon large language and vision-language models (Anderson et al., 2018; Krantz et al., 2020; Zheng et al., 2024; Zhou et al., 2024). Recent approaches increasingly integrate visual perception, language reasoning, spatial planning, memory, and action generation within unified navigation policies (Zhang et al., 2024; Cheng et al., 2025; Zhang et al., 2025; Wei et al., 2026b;a). As these models move toward general-purpose embodied agents, reliable navigation requires multiple capabilities to operate jointly: agents must ground diverse targets, interpret different task specifications, plan feasible trajectories, and maintain task progress over extended interactions.

Existing VLN benchmarks primarily characterize agents using aggregate end-task metrics such as Success Rate (SR) and Success weighted by Path Length (SPL) (Anderson et al., 2018; Krantz et al., 2020). While recent benchmarks have expanded evaluation toward continuous and physically realistic settings (Krantz et al., 2020; Wang et al., 2025), aggregate success reveals whether an agent completes a task but provides limited insight into which capability constrains it when navigation fails. This distinction is important because navigation is inherently compositional: even simple target reaching jointly depends on perception and execution, while more demanding behaviors additionally require spatial reasoning, exploration, obstacle avoidance, and memory. Consequently, neither a single navigation task nor a single ablation can be directly equated with an individual capability.

To address this limitation, we introduce NavDBench, a high-fidelity simulation benchmark for diagnosing four capabilities that we consider essential for reliable VLN: Visual Grounding, Trajectory

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

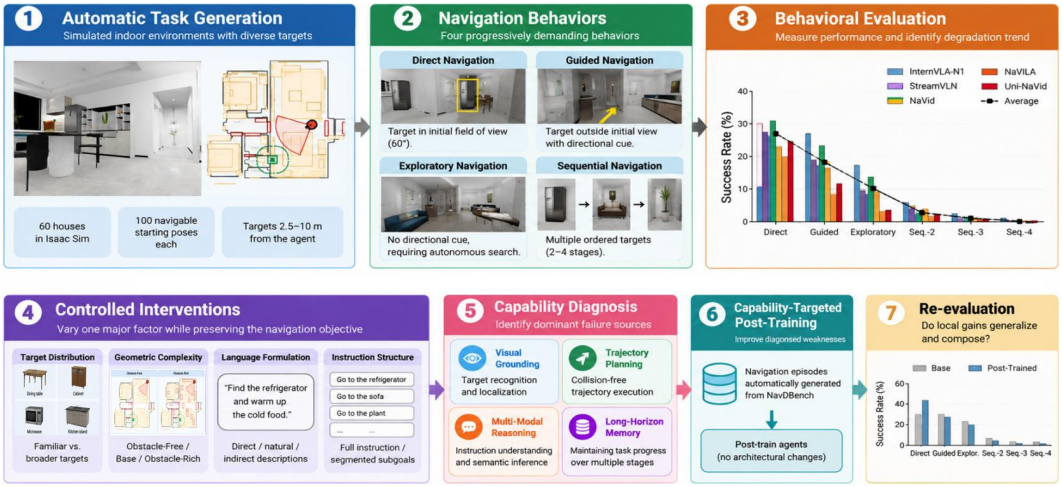


Figure 1: **Overview of the NavDBench evaluation framework.** NavDBench automatically generates navigation tasks with progressively demanding behaviors, measures behavioral degradation, and applies controlled interventions to diagnose capability-specific limitations. Capability-targeted post-training then tests whether diagnosed weaknesses can be improved and whether local gains generalize and compose into more demanding navigation behavior.

Planning, Multi-Modal Reasoning, and Long-Horizon Memory. As illustrated in Fig. 1, NavDBench provides an end-to-end diagnostic framework spanning automatic task generation, behavioral evaluation, controlled capability diagnosis, and capability-targeted post-training. It first constructs progressively demanding navigation behaviors to expose where performance degrades, and then applies controlled interventions to vary major sources of difficulty while preserving the underlying navigation objective. Rather than assigning each behavior to a single capability, NavDBench combines evidence across behaviors and interventions to construct capability-dominant diagnostic settings. Finally, post-training and re-evaluation test whether diagnosed weaknesses can be improved and whether local capability gains generalize and compose into more demanding navigation behavior.

Using NavDBench, we evaluate five representative VLN agents—InternVLA-N1, StreamVLN, Uni-NaVid, NaVid, and NaVILA. Our evaluation reveals several distinct capability limitations. Visual grounding remains unreliable and sensitive to target distributions, while the evaluated language variations have comparatively limited impact. Geometric interventions expose trajectory planning and obstacle avoidance as major short-horizon bottlenecks. At longer horizons, sequential navigation reveals that performance degradation arises from both accumulated local navigation errors and difficulty maintaining task progress. Finally, navigation-specific post-training substantially improves short-horizon capabilities across multiple diagnostic conditions, yet these gains largely disappear under autonomous exploration and sequential navigation. Together, these results suggest that reliable VLN requires not only improving individual navigation capabilities, but enabling them to compose robustly as task complexity and interaction horizon increase.

2 Related Work

Vision-Language Navigation. Vision-language navigation was introduced with R2R, where agents follow natural-language instructions over navigation graphs in previously unseen indoor environments (Anderson et al., 2018). VLN-CE extended this setting to continuous environments requiring low-level navigation actions (Krantz et al., 2020). Later methods improved history modeling, global planning, and large-scale navigation supervision (Chen et al., 2021; 2022; Wang et al., 2023). More recently, LLM- and VLM-based approaches have moved toward generalist embodied agents integrating language-conditioned reasoning, visual perception, temporal context, and action generation (Zheng et al., 2024; Zhou et al., 2024; Zhang et al., 2024; Cheng et al., 2025; Zhang et al., 2025; Wei et al., 2026b;a). Despite these advances, evaluation remains dominated by aggregate navigation metrics, which provide limited insight into the capability responsible for a failure.

Embodied Evaluation and Capability Diagnosis. VLN benchmarks have progressively expanded evaluation realism and coverage. VLN-CE evaluates continuous embodied navigation (Krantz et al., 2020), while VLN-PE studies realistic robot embodiment, sensing, locomotion, and environmental variations (Wang et al., 2025). Beyond VLN, EmbodiedBench evaluates multimodal foundation models across embodied capabilities including perception, spatial awareness, instruction understanding, and long-term planning (Yang et al., 2025). NavDBench complements these efforts by connecting progressive VLN behaviors with controlled interventions to diagnose capability bottlenecks and further testing whether diagnosed weaknesses respond to targeted post-training.

3 NAVDBENCH

3.1 OVERVIEW

NavDBench characterizes Visual Grounding, Trajectory Planning, Multi-Modal Reasoning, and Long-Horizon Memory through combinations of navigation behaviors and controlled interventions. Since these capabilities operate jointly in closed-loop navigation, we do not assign individual tasks to individual capabilities. Instead, controlled settings suppress major competing requirements so that a capability of interest becomes increasingly dominant.

As illustrated in Fig. 2, NavDBench contains four base behaviors—Direct, Guided, Exploratory, and Sequential Navigation—together with controlled variations in geometric complexity, language formulation, target distribution, and instruction structure.

3.2 AUTOMATIC TASK GENERATION

We construct NavDBench in 60 synthetic household environments in Isaac Sim, sampling 100 valid navigable starting poses per environment and candidate targets 2.5–10 m from the agent. From these poses, we automatically generate four behaviors with progressively increasing requirements. Direct Navigation places the target within the initial 60° field of view. Guided Navigation places it outside the initial view but provides an egocentric directional cue that brings the target into view. Exploratory Navigation removes this cue and requires autonomous search. Sequential Navigation composes 2–4 ordered stages, with the first following the Direct setting and subsequent stages following the Guided setting. Detailed generation rules and examples are provided in the Appendix.

3.3 CONTROLLED DIAGNOSTIC INTERVENTIONS

We construct matched task variants that modify one major source of difficulty while preserving the navigation objective whenever possible. Geometric interventions vary obstacle density across Obstacle-Free, Base, and Obstacle-Rich conditions; language interventions reformulate the same objective using direct, natural, or indirect descriptions; target-distribution interventions compare the six categories commonly used in Habitat ObjectNav with broader household objects; and instruction fragmentation provides only the active subgoal during Sequential Navigation, reducing internal task-state requirements. We refer to the six ObjectNav categories as *familiar navigation targets* because of their prevalence in standard navigation benchmarks; this terminology does not assume that every evaluated model was explicitly trained on these categories. Additional implementation details and examples are provided in the Appendix.

3.4 CAPABILITY DIAGNOSIS

We diagnose capabilities from combinations of behaviors and interventions rather than from any single task. Visual Grounding is approximated using Obstacle-Free Direct Navigation and target-distribution variation; Trajectory Planning is diagnosed by varying geometric complexity for matched objectives; Multi-Modal Reasoning is examined through semantic language reformulation and visually grounded spatial-instruction following; and Long-Horizon Memory is examined through increasing sequential length and instruction fragmentation. These settings do not perfectly isolate individual capabilities, but suppress major competing requirements to provide capability-dominant embodied proxies.

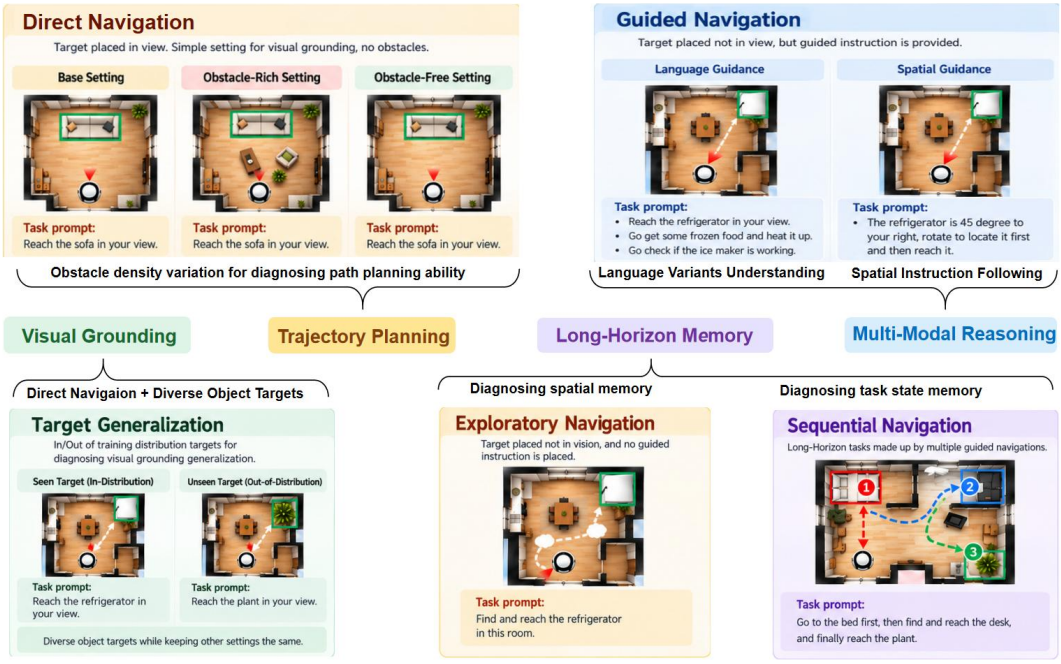


Figure 2: **Diagnostic design of NavDBench.** NavDBench combines progressively demanding navigation behaviors with controlled task and environmental interventions to diagnose four essential VLN capabilities. Rather than assigning each behavior to a single capability, evidence is combined across behaviors and controlled settings to construct capability-dominant diagnostic evaluations.

3.5 Capability-Targeted Post-Training

Finally, we post-train representative agents using automatically generated navigation episodes without introducing new architectural components. Re-evaluation under the same behaviors and diagnostic conditions tests which capabilities respond to additional supervision and whether local improvements propagate to exploration and long-horizon navigation.

4 Experiments

4.1 Experimental Setup

We evaluate five representative VLN agents: InternVLA-N1, StreamVLN, NaVid, NaVILA, and Uni-NaVid. All agents are evaluated in the same Isaac Sim household environments using Success Rate (SR) as the primary metric. Unless otherwise specified, each diagnostic comparison uses matched task subsets in which the target and navigation objective are preserved while the intervention of interest is varied. Absolute SR can therefore differ across diagnostic subsets, and we primarily interpret the within-subset performance change induced by each intervention. Additional model configurations, episode statistics, success criteria, and implementation details are provided in the Appendix.

4.2 Capability Diagnosis

Table 1 and Fig. 3 show a clear degradation as navigation requirements increase. Average SR decreases from 25.1% in Direct Navigation to 17.6% in Guided and 9.6% in Exploratory Navigation, before falling to 3.3%, 0.6%, and 0.2% for two-, three-, and four-stage Sequential Navigation. These behavioral results reveal where agents become unreliable, but not which capability causes the failure. We therefore use the controlled interventions in Sec. 3.3 to diagnose each capability.

Table 1: Overall success rate (%) across navigation behaviors. Sequential Navigation contains 2–4 ordered stages.

Model	Direct	Guided	Explor.	Seq.-2	Seq.-3	Seq.-4
InternVLA-N1	29.7	29.1	16.8	4.7	1.0	0.3
StreamVLN	21.6	20.2	11.3	4.4	1.0	0.5
NaVid	28.9	17.0	12.7	2.8	0.4	0.1
NaVILA	17.5	17.2	5.5	3.1	0.5	0.2
Uni-NaVid	27.8	4.7	1.8	1.4	0.0	0.0
Average	25.1	17.6	9.6	3.3	0.6	0.2

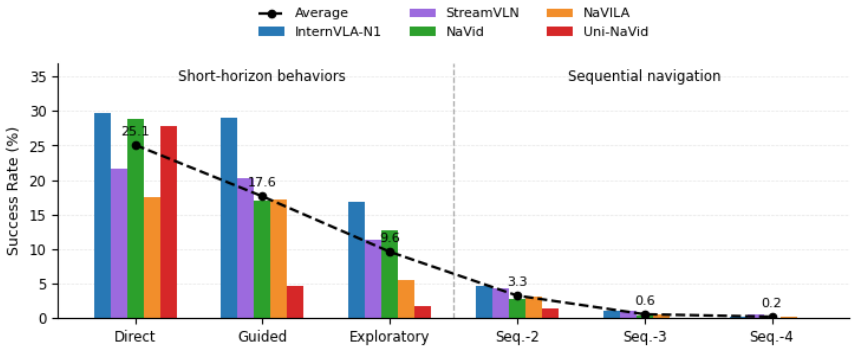


Figure 3: **Overall navigation performance.** Success rates of five VLN agents across Direct, Guided, Exploratory, and Sequential Navigation. The dashed line denotes the cross-model average.

4.2.1 Visual Grounding

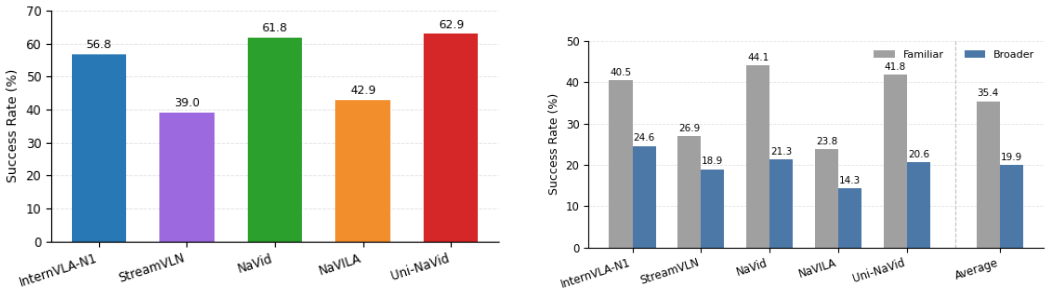
We approximate embodied visual grounding using Obstacle-Free Direct Navigation, where the target is directly specified and initially visible while obstacles between the agent and target are removed. This suppresses search, long-horizon memory, and obstacle avoidance; after grounding the target, the remaining behavior primarily requires orienting toward it and moving forward. As shown in Fig. 4a, removing geometric complexity substantially improves performance, yet no model exceeds 63% SR. Thus, although geometric difficulty explains a substantial fraction of Direct Navigation failures, reliable target grounding and target-directed execution remain unresolved even under this deliberately simplified setting.

We further test generalization across target categories in Fig. 4b. Average SR decreases from 35.4% on familiar navigation targets to 19.9% on broader household objects, a 15.5-point drop, with degradation observed across all five agents. Visual grounding is therefore limited both in basic reliability and in generalization beyond familiar target distributions.

4.2.2 Trajectory Planning

Figure 5 compares matched objectives across Obstacle-Free, Base, and Obstacle-Rich conditions. Geometric complexity strongly affects Direct and Guided Navigation: removing obstacles improves Direct SR by 17.4–35.2 points across models, while obstacle-rich conditions consistently reduce performance. The same trend extends to Guided Navigation; for example, StreamVLN increases from 20.2% to 46.9% after obstacle removal, while its SR decreases to 13.7% under the Obstacle-Rich condition.

The pattern is less uniform in Exploratory Navigation; notably, obstacle removal does not improve InternVLA-N1. This indicates that easier local geometry alone cannot resolve autonomous exploration, where search and other capabilities remain necessary. Overall, the controlled geometry results identify trajectory planning and obstacle avoidance as major short-horizon bottlenecks, while also showing that they cannot fully explain failures in more compositional behaviors.



(a) Obstacle-Free Direct Navigation.

(b) Target-category generalization.

Figure 4: **Visual grounding diagnosis.** (a) Obstacle-Free Direct Navigation provides an approximate embodied grounding proxy by minimizing competing navigation requirements. (b) Performance on familiar navigation targets and broader household categories.

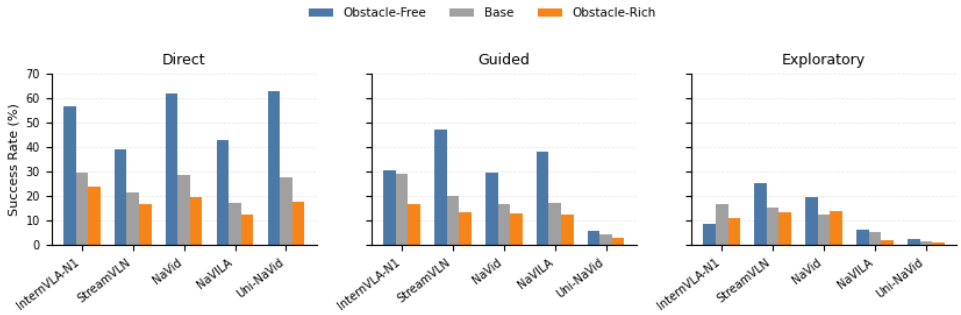


Figure 5: **Trajectory planning diagnosis.** SR under Obstacle-Free, Base, and Obstacle-Rich conditions for Direct, Guided, and Exploratory Navigation. The navigation objective is preserved while geometric complexity is varied.

4.2.3 Multi-Modal Reasoning

Figure 6 compares matched objectives under base and alternative language formulations. Changes are small and inconsistent across Direct, Guided, and Exploratory Navigation: the largest decrease is 3.0 points for NaVid in Guided Navigation, while several models instead improve under the variants. No consistent degradation is observed as the same objective is reformulated.

This pattern contrasts sharply with the geometric interventions in Fig. 5, where changing obstacle complexity frequently shifts SR by tens of percentage points. This does not imply that multi-modal reasoning is solved; rather, within the language and spatial-instruction variations evaluated by NavDBench, semantic reformulation is not a dominant source of failure. Current agents appear substantially more robust to these linguistic variations than to changes in geometric complexity.

4.2.4 Long-Horizon Memory

Sequential Navigation exhibits the sharpest degradation in Fig. 3. However, this decline cannot be attributed to memory alone. Because the first sequential stage follows the Direct setting and subsequent stages follow the Guided setting, we use $p_D p_G^{N-1}$ as a simple compounding reference, where p_D and p_G denote the corresponding single-stage success rates and N is the number of stages. This reference assumes independent stages and is not intended as a prediction of sequential performance, but provides an approximate baseline for degradation caused by accumulated local navigation errors.

As shown in Fig. 7, segmentation consistently improves sequential performance. InternVLA-N1 increases from 4.7% to 6.2%, 1.0% to 2.2%, and 0.33% to 0.70% for two, three, and four stages, respectively; StreamVLN improves from 4.4% to 5.9%, 1.0% to 1.2%, and 0.5% to 0.7%. Since

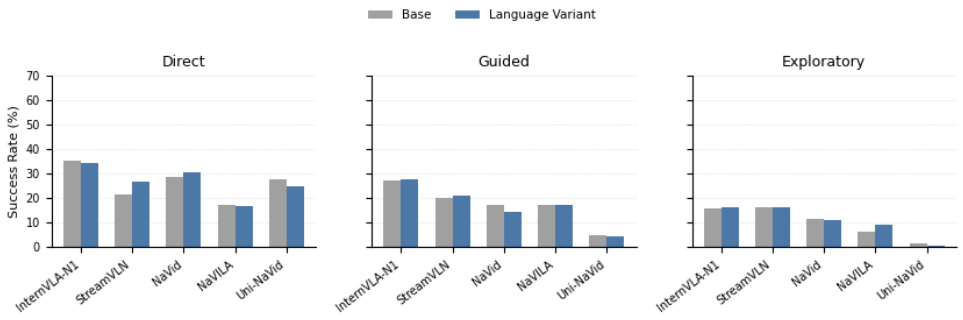


Figure 6: **Multi-modal reasoning diagnosis.** SR under base and language-variant instructions for Direct, Guided, and Exploratory Navigation, with the underlying navigation objective preserved.

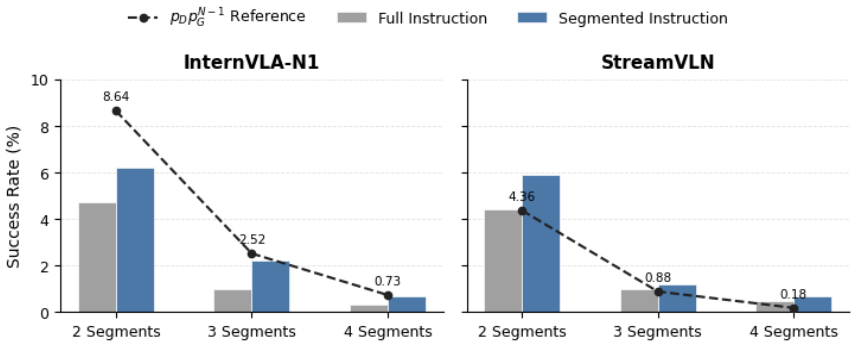


Figure 7: **Long-horizon memory diagnosis.** Sequential performance under full and segmented instructions for InternVLA-N1 and StreamVLN. Segmentation externally provides the active subgoal, reducing task-state requirements; the dashed $p_D p_G^{N-1}$ curve provides a reference for accumulated independent stage failures.

the physical target sequence is unchanged, this recovery provides evidence that maintaining task progress contributes to long-horizon failure.

Segmentation nevertheless does not prevent rapid degradation with sequence length. The compounding reference shows that local errors also accumulate as more stages must succeed. For InternVLA-N1, the four-stage reference $p_D p_G^3$ is approximately 0.73%, close to its 0.70% segmented SR. We therefore interpret the reference as a diagnostic baseline rather than a prediction. Long-horizon failure reflects both accumulated local navigation errors and difficulty maintaining task state, suggesting that improving memory alone cannot compensate for weak per-stage reliability.

4.3 Case Study: Typical Failure Modes

Beyond aggregate performance, qualitative inspection of navigation episodes reveals recurring failure patterns that are consistent with the capability-level diagnostics above. First, agents sometimes fail to ground a target even when the requested object is clearly visible directly in front of the robot, consistent with the remaining failures under Obstacle-Free Direct Navigation. Second, agents frequently become stuck after repeatedly colliding with or attempting to move through obstacles, illustrating the difficulty of translating a navigation objective into feasible collision-free execution. Third, in long-horizon Sequential Navigation, agents may revisit previously completed targets instead of progressing to the remaining subgoals, indicating difficulty maintaining task progress across stages. These cases provide qualitative examples of how the diagnosed capability limitations manifest during closed-loop navigation. Representative failure videos will be released on the project website upon publication.

Table 2: **Effect of capability-targeted post-training on StreamVLN.** Success rate (SR, %) before and after post-training, with absolute change in percentage points (pp).

Evaluation	Condition / Behavior	Base SR	Post-Trained SR	Δ pp
Target Distribution	Familiar Targets	26.70	49.30	+22.60
	Broader Targets	14.10	29.90	+15.80
Diagnostic Condition	Standard	18.20	37.90	+19.70
	Obstacle-Free	40.91	62.50	+21.59
	Obstacle-Rich	16.21	32.14	+15.93
	Language Variant	25.91	35.65	+9.74
Navigation Behavior	Direct	18.20	37.90	+19.70
	Guided	19.57	22.58	+3.01
	Exploratory	14.41	12.47	-1.94
	Sequential-2	2.87	2.68	-0.19
	Sequential-3	0.77	0.19	-0.58
	Sequential-4	0.96	0.00	-0.96

4.4 Capability-Targeted Post-Training

We finally ask whether diagnosed weaknesses respond to additional navigation supervision and whether these improvements compose into more demanding behavior. We post-train StreamVLN on automatically generated navigation episodes without architectural modifications and re-evaluate it under the same diagnostic conditions.

As shown in Table 2, post-training substantially improves short-horizon target-directed navigation. Gains transfer across familiar (+22.6 pp) and broader (+15.8 pp) target categories, indicating that the improvement is not restricted to the most familiar target set. Consistent gains also appear across geometric conditions: +21.6 pp in Obstacle-Free, +19.7 pp in Standard, and +15.9 pp in Obstacle-Rich environments. Performance under language variation improves by 9.7 pp. Together, these results show that additional navigation supervision can improve the local capabilities exposed by our diagnostics and that these improvements transfer across multiple evaluation conditions.

However, capability-level improvements do not automatically compose into stronger behavior. Direct Navigation improves by 19.7 points, whereas the gain falls sharply to 3.0 points in Guided Navigation. Exploratory Navigation instead decreases by 1.9 points, and no improvement is observed for any Sequential setting. The contrast becomes particularly informative when considered together with the diagnostic results: post-training substantially strengthens target-directed execution under diverse local conditions, yet provides little benefit once successful navigation additionally requires autonomous search or sustained multi-stage task progress.

These results expose a distinction between improving individual navigation capabilities and composing them over increasingly demanding interactions. Additional supervision can substantially strengthen local target-directed navigation, but these gains alone are insufficient for reliable exploration and long-horizon behavior. This motivates future approaches that improve not only local navigation competence, but also exploration, subgoal progression, and persistent task-state management.

5 Conclusion

We present NavDBench, a diagnostic benchmark that combines progressive navigation behaviors with controlled interventions to characterize Visual Grounding, Trajectory Planning, Multi-Modal Reasoning, and Long-Horizon Memory. Across five representative agents, we find that grounding remains unreliable and distribution-sensitive, geometric planning constitutes a major short-horizon bottleneck, evaluated language variations have comparatively limited impact, and long-horizon failure reflects both accumulated local errors and difficulty maintaining task progress.

Post-training shows that these capability limitations respond differently to additional supervision: local target-directed navigation improves substantially across several diagnostic conditions, yet the gains largely disappear under autonomous exploration and sequential navigation. Reliable VLN

therefore requires not only stronger individual capabilities, but also their robust composition as task complexity and interaction horizon increase. NavDBench provides a framework for diagnosing these capability boundaries and testing whether future improvements translate from local navigation competence to reliable long-horizon behavior.

REFERENCES

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History aware multimodal transformer for vision-and-language navigation. In *Advances in Neural Information Processing Systems*, 2021.

Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16537–16547, 2022.

An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Zaitian Gongye, Xueyan Zou, Jan Kautz, Erdem Biyik, Hongxu Yin, Sifei Liu, and Xiaolong Wang. Navila: Legged robot vision-language-action model for navigation. In *Proceedings of Robotics: Science and Systems*, 2025. doi: 10.15607/RSS.2025.XXI.018.

Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *European Conference on Computer Vision*, 2020.

Liuyi Wang, Xinyuan Xia, Hui Zhao, Hanqing Wang, Tai Wang, Yilun Chen, Chengju Liu, Qijun Chen, and Jiangmiao Pang. Rethinking the embodied gap in vision-and-language navigation: A holistic study of physical and visual disparities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9455–9465, 2025. doi: 10.1109/ICCV51701.2025.00882.

Zun Wang, Jialu Li, Yicong Hong, Yi Wang, Qi Wu, Mohit Bansal, Stephen Gould, Hao Tan, and Yu Qiao. Scaling data generation in vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12009–12020, 2023.

Meng Wei, Chenyang Wan, Jiaqi Peng, Xiqian Yu, Yuqiang Yang, Delin Feng, Wenzhe Cai, Chenming Zhu, Tai Wang, Jiangmiao Pang, and Xihui Liu. Ground slow, move fast: A dual-system foundation model for generalizable vision-language navigation. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Meng Wei, Chenyang Wan, Xiqian Yu, Tai Wang, Yuqiang Yang, Xiaohan Mao, Chenming Zhu, Wenzhe Cai, Hanqing Wang, Yilun Chen, Xihui Liu, and Jiangmiao Pang. Streamvln: Streaming vision-and-language navigation via slowfast context modeling. In *IEEE International Conference on Robotics and Automation*, 2026b.

Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, Heng Ji, Huan Zhang, and Tong Zhang. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 70576–70631. PMLR, 2025.

Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and He Wang. Navid: Video-based vlm plans the next step for vision-and-language navigation, 2024.

Jiazhao Zhang, Kunyu Wang, Shaoan Wang, Minghan Li, Haoran Liu, Songlin Wei, Zhongyuan Wang, Zhizheng Zhang, and He Wang. Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks. In *Proceedings of Robotics: Science and Systems*, 2025. doi: 10.15607/RSS.2025.XXI.013.

Table 3: **Summary of NavDBench navigation behaviors.** The four behaviors progressively introduce autonomous search and longer interaction horizons.

Behavior	Initial Target	Direction Cue	Stages
Direct	Visible (60°)	No	1
Guided	Outside view	Yes	1
Exploratory	Outside view	No	1
Sequential	Direct $\rightarrow$ Guided	Later stages	2–4

Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. Towards learning a generalist model for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13624–13634, 2024.

Gengze Zhou, Yicong Hong, Zun Wang, Xin Eric Wang, and Qi Wu. Navgpt-2: Unleashing navigational reasoning capability for large vision-language models. In *European Conference on Computer Vision*, 2024.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their valuable feedback.

A ADDITIONAL BENCHMARK DETAILS

A.1 AUTOMATIC TASK GENERATION

NavDBench is constructed in 60 synthetic household environments in Isaac Sim. For each environment, we sample 100 valid navigable starting poses and obtain object-level information directly from the simulator. Candidate navigation targets are selected according to their visibility, relative direction, and distance from the agent. We restrict target distance to 2.5–10 m: very short distances can produce trivial successes with little navigation, whereas substantially longer distances introduce additional geometric complexity that can confound capability diagnosis.

For each sampled starting pose, we automatically construct four navigation behaviors:

Direct Navigation. A target object is selected within the agent's initial 60° field of view and 2.5–10 m from the starting position. The instruction directly specifies the target object. This setting requires the agent to visually ground the specified target and execute a trajectory toward it, while minimizing explicit search requirements.

Guided Navigation. The target is selected outside the initial 60° field of view and within the same distance range. The instruction additionally provides an egocentric directional cue specifying the target's relative direction. Following this cue brings the target into view, after which the agent must locate and reach it.

Exploratory Navigation. The target is initially outside the field of view and no directional cue is provided. The agent must autonomously search the environment, discover the target, and navigate toward it.

Sequential Navigation. We compose 2–4 navigation stages into a single ordered objective. The first target follows the Direct Navigation setting, while subsequent targets follow the Guided Navigation setting with directional cues. Success requires completing all targets in the specified order.

Table 3 summarizes the major differences among the four navigation behaviors.

Table 4: **Controlled interventions used for capability diagnosis.** Each intervention varies a major source of difficulty while preserving the navigation objective whenever possible.

Intervention	Controlled Variation	Preserved Components	Primary Diagnostic Use
Geometric Complexity	Obstacle-Free / Base / Obstacle-Rich	Target identity and instruction	Trajectory Planning
Language Formulation	Direct / natural / indirect descriptions	Physical navigation objective	Multi-Modal Reasoning
Target Distribution	Familiar / broader household targets	Navigation-generation constraints	Visual Grounding generalization
Instruction Structure	Full / segmented instruction	Ordered physical target sequence	Long-Horizon Memory

A.2 Controlled Diagnostic Interventions

A central design principle of NavDBench is to preserve the underlying navigation objective while modifying a major source of difficulty. These interventions do not perfectly isolate individual capabilities, which is generally infeasible in closed-loop embodied navigation. Instead, they suppress major competing requirements and create approximately capability-dominant diagnostic settings.

Geometric Complexity. For matched navigation objectives, we construct three geometric conditions. *Base* uses the original environment configuration. *Obstacle-Free* removes obstacles that introduce additional geometric difficulty between the agent and target, while *Obstacle-Rich* introduces additional obstacles. The target and instruction are preserved across matched variants. Comparing SR across these conditions measures sensitivity to trajectory planning and obstacle avoidance.

For Visual Grounding, we additionally combine Direct Navigation with the Obstacle-Free condition. Because the target is initially visible and major intervening obstacles are removed, search, memory, and obstacle avoidance requirements are minimized. The remaining task primarily requires grounding the specified object, orienting toward it, and moving toward it. We therefore treat this setting as an approximate embodied grounding proxy rather than a pure perception test.

Language Formulation. We generate alternative formulations of the same navigation objective while preserving the physical target. These include direct target descriptions, more natural language formulations, and indirect descriptions that may require semantic or functional inference. Guided Navigation additionally evaluates the interpretation of egocentric spatial cues. These variants probe the aspects of Multi-Modal Reasoning associated with semantic language understanding and visually grounded spatial-instruction following.

Target Distribution. We divide navigation targets into the six object categories commonly used in Habitat ObjectNav and a broader collection of household objects available in our Isaac Sim environments. We refer to the former as *familiar navigation targets* because of their prevalence in standard navigation benchmarks. This terminology does not imply that every evaluated model was explicitly trained on these categories.

Instruction Fragmentation. In standard Sequential Navigation, the complete ordered instruction is provided at the beginning of the episode. In the segmented setting, only the currently active subgoal is provided; after a target is successfully completed, the instruction is updated to the next subgoal. The physical target sequence is unchanged. This intervention externally maintains the active task stage and therefore reduces, but does not eliminate, the requirement for internal task-state maintenance.

Table 4 summarizes the controlled interventions and their primary diagnostic roles.

Table 5: **Overall navigation performance.** Success rate (%) across Direct, Guided, Exploratory, and Sequential Navigation.

Model	Direct	Guided	Exploratory	Seq.-2	Seq.-3	Seq.-4
InternVLA-N1	29.70	29.10	16.80	4.70	1.00	0.33
StreamVLN	21.60	20.21	11.30	4.40	1.00	0.50
NaVid	28.88	16.98	12.75	2.80	0.40	0.10
NaVILA	17.48	17.22	5.51	3.10	0.50	0.20
Uni-NaVid	27.75	4.69	1.81	1.40	0.00	0.00
Average	25.08	17.64	9.63	3.28	0.58	0.23

Table 6: **Obstacle-Free Direct Navigation.** SR (%) under Base and Obstacle-Free conditions.

Model	Base	Obstacle-Free	Δ pp
InternVLA-N1	29.70	56.76	+27.06
StreamVLN	21.60	39.00	+17.40
NaVid	28.88	61.78	+32.90
NaVILA	17.48	42.86	+25.38
Uni-NaVid	27.75	62.93	+35.18

B Additional Quantitative Results

B.1 Overall Navigation Performance

Table 5 provides the exact success rates underlying the overall behavioral comparison in the main paper.

B.2 Visual Grounding

Table 6 reports Direct Navigation performance before and after removing major geometric obstacles. The Obstacle-Free condition is used as the simplified embodied grounding proxy in the main paper.

Table 7 reports performance across the two target distributions.

B.3 Trajectory Planning

Tables 8–10 provide the exact results of the geometric-complexity intervention for each short-horizon navigation behavior.

B.4 Multi-Modal Reasoning

Tables 11–13 report the exact results of the language-formulation intervention.

B.5 Long-Horizon Memory

Table 14 reports the exact Sequential Navigation results under full and segmented instructions. The physical target sequence is identical across the two conditions; segmentation only externally provides the active subgoal.

Because the first stage follows Direct Navigation and subsequent stages follow Guided Navigation, the compounding reference used in the main paper is

$$P_{\text{compound}}(N) = p_D p_G^{N-1},$$

where p_D and p_G denote the corresponding single-stage Direct and Guided success rates. This reference assumes independent stages and is intended only as a diagnostic baseline for accumulated local navigation errors.

Table 7: **Target-category generalization.** SR (%) on familiar navigation targets and broader household targets.

Model	Familiar	Broader	Δ pp
InternVLA-N1	40.5	24.6	-15.9
StreamVLN	26.9	18.9	-8.0
NaVid	44.1	21.3	-22.8
NaVILA	23.8	14.3	-9.5
Uni-NaVid	41.8	20.6	-21.2
Average	35.4	19.9	-15.5

Table 8: **Geometric intervention: Direct Navigation.** SR (%) under Obstacle-Free, Base, and Obstacle-Rich conditions.

Model	Free	Base	Rich
InternVLA-N1	56.76	29.70	23.83
StreamVLN	39.00	21.60	16.65
NaVid	61.78	28.88	19.88
NaVILA	42.86	17.48	12.34
Uni-NaVid	62.93	27.75	17.77

C Qualitative Failure Analysis

Manual inspection of navigation episodes reveals three recurring failure modes that complement the quantitative diagnostics in the main paper.

Failure to Ground Clearly Visible Targets. Agents occasionally fail to navigate toward the requested object even when it is clearly visible directly in front of the robot. Such cases are consistent with the substantial residual error observed under Obstacle-Free Direct Navigation, where search and geometric complexity are strongly reduced.

Obstacle-Induced Stagnation. Agents can repeatedly collide with or attempt to move through obstacles without recovering to a feasible trajectory. In some episodes, the agent remains locally stuck despite the target direction being apparent. These behaviors are consistent with the strong sensitivity to geometric complexity observed in the trajectory-planning diagnosis.

Revisiting Completed Targets. In Sequential Navigation, agents sometimes return to previously completed targets instead of progressing toward the remaining subgoals. This failure becomes particularly apparent over longer sequences and is consistent with difficulty maintaining task progress across stages.

Representative videos of these failure modes will be released on the project website upon publication.

Table 9: **Geometric intervention: Guided Navigation.** SR (%) under Obstacle-Free, Base, and Obstacle-Rich conditions.

Model	Free	Base	Rich
InternVLA-N1	30.61	29.10	16.71
StreamVLN	46.94	20.21	13.70
NaVid	29.80	16.98	13.20
NaVILA	38.37	17.22	12.45
Uni-NaVid	5.71	4.69	3.06

Table 10: **Geometric intervention: Exploratory Navigation.** SR (%) under Obstacle-Free, Base, and Obstacle-Rich conditions.

Model	Free	Base	Rich
InternVLA-N1	8.79	16.80	10.97
StreamVLN	25.52	15.59	13.50
NaVid	19.67	12.75	14.01
NaVILA	6.28	5.51	2.06
Uni-NaVid	2.51	1.81	1.29

Table 11: **Language intervention: Direct Navigation.** SR (%) under base and language-variant instructions.

Model	Base	Variant	Δ pp
InternVLA-N1	35.14	34.18	-0.96
StreamVLN	21.54	26.86	+5.32
NaVid	28.89	30.80	+1.91
NaVILA	17.30	16.83	-0.47
Uni-NaVid	27.77	24.96	-2.81

Table 12: **Language intervention: Guided Navigation.** SR (%) under base and language-variant instructions.

Model	Base	Variant	Δ pp
InternVLA-N1	27.02	27.81	+0.79
StreamVLN	20.00	21.27	+1.27
NaVid	17.44	14.48	-2.96
NaVILA	17.35	17.25	-0.10
Uni-NaVid	4.96	4.62	-0.34

Table 13: **Language intervention: Exploratory Navigation.** SR (%) under base and language-variant instructions.

Model	Base	Variant	Δ pp
InternVLA-N1	15.67	16.53	+0.86
StreamVLN	16.47	16.53	+0.06
NaVid	11.42	11.27	-0.15
NaVILA	6.37	9.35	+2.98
Uni-NaVid	1.46	0.75	-0.71

Table 14: **Full versus segmented Sequential Navigation.** SR (%) for InternVLA-N1 and StreamVLN.

Model	Stages	Full	Segmented	Δ pp
InternVLA-N1	2	4.70	6.20	+1.50
	3	1.00	2.20	+1.20
	4	0.33	0.70	+0.37
StreamVLN	2	4.40	5.90	+1.50
	3	1.00	1.20	+0.20
	4	0.50	0.70	+0.20

Table 15: **Compounding reference for Sequential Navigation.** Values are $p_D p_G^{N-1}$ expressed as percentages.

Model	2 Stages	3 Stages	4 Stages
InternVLA-N1	8.64	2.51	0.73
StreamVLN	4.36	0.88	0.18