

VLN-MME : Diagnosing MLLMs as Language-guided Visual Navigation Agents

Xunyi Zhao*, Gengze Zhou*†, Qi Wu

Australian Institute for Machine Learning, Adelaide University

{xunyi.zhao, gengze.zhou, qi.wu01}@adelaide.edu.au

<https://github.com/billzhao1030/VLN-MME>

Abstract

Multimodal Large Language Models (MLLMs) have demonstrated remarkable capabilities across a wide range of vision-language tasks. However, their performance as embodied agents, which requires multi-round interaction with spatial reasoning and sequential action prediction, needs further exploration. Our work investigates this potential in the context of Vision-and-Language Navigation (VLN) by introducing a unified and extensible simulation-free evaluation framework to probe MLLMs as zero-shot agents, named VLN-MME. Simplifying the evaluation with a highly modular and accessible design streamlines experiments, enabling structured comparisons and component-level ablations across diverse MLLM architectures, agent designs, and navigation tasks. Crucially, enabled by VLN-MME, we observe that enhancing prevalent agents with Chain-of-Thought (CoT) reasoning and self-reflection leads to an unexpected performance decrease. This suggests MLLMs exhibit poor context awareness in embodied navigation tasks; although they can follow instructions and structure their output, their 3D spatial reasoning fidelity is low. Furthermore, we demonstrate that agent performance could be largely improved with simple failure cases in context learning. VLN-MME lays the groundwork for systematic evaluation of general-purpose MLLMs in embodied navigation settings and reveals limitations in their sequential decision-making capabilities. We believe these findings offer crucial guidance for MLLM post-training as embodied agents.

1 Introduction

The rapid advancement of Multimodal Large Language Models (MLLMs) has catalyzed a shift from static vision-language tasks toward the development of embodied agents capable of dynamic, interactive decision-making. Central to this transition

is the evaluation of 3D spatial reasoning. However, prior benchmarks have largely focused on static spatial Question Answering or simple navigation or manipulation tasks within low-fidelity, synthetic environments (Yang et al., 2025b; Daxberger et al., 2025; Yang et al., 2025c; Cheng et al., 2025). Crucially, these settings lack substantial changes in 3D environmental states and do not require the maintenance of long-term spatial memory. In contrast, Vision-and-Language Navigation (VLN) (Anderson et al., 2018) emerges as a far more challenging paradigm for evaluating these abilities. Successfully navigating a 3D environment requires more than simple pattern recognition; it fundamentally tests an agent’s spatial understanding, its ability to maintain awareness, foresee the consequences of its actions across dynamic viewpoint changes, and the use of long-term memory to ground extended plans. Furthermore, when navigation involves multi-round interaction, it probes the model’s capacity for contextual reasoning. However, despite VLN’s potential as a comprehensive benchmark for these core agentic skills, progress in systematically evaluating MLLMs is constrained by the limitations of existing evaluation pipelines.

First, embodied navigation typically relies on high-fidelity simulators such as Matterport3D (Chang et al., 2017) or Habitat (Savva et al., 2019). When deploying large MLLMs in these multi-round, interactive settings, the computational cost grows sharply. This issue is compounded by the high volume of trajectories across diverse benchmarks (Anderson et al., 2018; Qi et al., 2020; Ku et al., 2020), making comprehensive testing prohibitively time-consuming. As a result, the computational burden of exhaustive evaluation has pushed prior studies toward largely metric-driven approaches, prioritizing end-to-end success rates over diagnostic clarity. This lack of principled error analysis obscures the underlying model behavior, making it difficult to assess critical capabilities

*Equal Contribution, † Project lead

such as generalization, robustness, or the specific alignment between visual perception and instruction following.

More critically, recent evaluation suites like NavBench (Qiao et al., 2025) have attempted to standardize this process and successfully unify the evaluation of different tasks and models. However, it is restricted to a single agent design and not systematically varying agent designs; it becomes impossible to decouple the intrinsic capabilities of the MLLM from the efficacy of specific prompting or planning strategies. As a result, the community lacks a fine-grained understanding of whether failures stem from the model’s reasoning limitations or suboptimal agent engineering.

In response to these challenges, we propose the **Vision Language Navigation Multi-Model Evaluation (VLN-MME)**, a novel framework designed to address these limitations. Our approach features a modular and simulator-free architecture that preserves navigational semantics while eliminating the setup complexity and computational overhead of traditional pipelines. Crucially, we move beyond high-level success metrics to provide a detailed error analysis, dissecting agent performance to diagnose core proficiencies in instruction following, spatial understanding, and historical reasoning.

Our contributions are summarized as follows:

- We introduce VLN-MME, a unified and modular framework that systematically evaluates the interplay between Model, Agent, and Task, addressing the limitation of fixed-agent designs in prior benchmarks.
- We design a simulator-free evaluation pipeline that preserves essential navigational semantics while drastically reducing computational overhead and setup complexity to enhance accessibility.
- We provide a comprehensive diagnostic analysis that moves beyond success metrics, categorizing fine-grained failure modes to uncover specific deficiencies in MLLM spatial reasoning and instruction following.
- We release a standardized suite of processed datasets and environmental artifacts to facilitate reproducible research and streamline the benchmarking process.

2 Related Works

Evaluating MLLMs in Spatial and Embodied Contexts Comprehensive benchmarks have emerged to test a wide spectrum of MLLM abilities (Chaoyou et al., 2023; Liu et al., 2024; Li et al., 2024c; Yue et al., 2024; Yu et al., 2024; Lu et al., 2023; Fei et al., 2025), ranging from basic perception to complex cognition. Within this landscape, specific efforts have focused on assessing 3D spatial reasoning (Yang et al., 2025b; Daxberger et al., 2025; Xu et al., 2025; Liao et al., 2024; Li et al., 2024b). However, the majority of these benchmarks rely on static QA formats, where the model provides a single response to a fixed visual input, but not for continuous state tracking. To assess sequential reasoning, several benchmarks focus on long-horizon tasks; however, these are largely restricted to digital domains like web browsing and application usage (Deng et al., 2023; Trivedi et al., 2024; Tao et al., 2025; Wang et al., 2025). More recent embodied benchmarks attempt to bridge this gap but face distinct limitations. Some works, such as 3DMEM-Bench (Hu et al., 2025), focus primarily on high-level planning. While effective for evaluating abstract reasoning, these approaches often overlook the fine-grained environmental interactions required for realistic agent execution. Conversely, benchmarks like (Yang et al., 2025c; Cheng et al., 2025) operate within low-fidelity, synthetic environments. Although they incorporate manipulation tasks, these settings generally lack photorealistic visual complexity and do not involve the massive 3D environmental state changes inherent to large-scale navigation. Consequently, these tasks rarely demand the long-term spatial memory or the rigorous sequential reasoning required to operate in dynamic, photorealistic spaces.

MLLMs for Vision-and-Language Navigation The integration of MLLMs into robotics has inspired new paradigms for VLN. Early efforts leveraged LLMs as copilots to guide specialist agents (Qiao et al., 2023), while recent works employ off-the-shelf MLLMs as zero-shot agents through elaborate prompting (Zhou et al., 2024; Long et al., 2023; Chen et al., 2024; Dong et al., 2025). Concurrently, other studies have explored fine-tuning MLLMs directly on navigation data (Zhou et al., 2025a; Lin et al., 2024; Pan et al., 2023; Zheng et al., 2023) or adapting pre-trained video models (Zhang et al., 2024b,a; Cheng et al., 2024). Despite this progress, evaluation remains

fragmented and costly due to reliance on simulators. While recent efforts like SAME (Zhou et al., 2025b) and NavBench (Qiao et al., 2025) attempt to standardize task evaluation, they face distinct limitations: the former does not specifically target MLLM agents, while the latter is restricted to fixed agent design. Consequently, existing frameworks often focus on aggregate metrics, lacking the fine-grained diagnostics required to understand specific failure modes. We bridge this gap with a unified, simulator-free framework that jointly evaluates MLLMs, agent designs, and diverse tasks.

3 Method

3.1 Modular VLN Evaluation Framework

To enable systematic assessment of MLLMs’ spatial reasoning, long-horizon planning, and sequential decision-making capabilities in embodied settings, we design a modular software stack that cleanly separates three core components: Model, Agent, and Task (Figure 1 & 2). This modularity empowers us to seamlessly interchange each component independently, enabling structured comparisons and component-level ablations to isolate the source of success or failure. Our framework is built upon three primary abstractions:

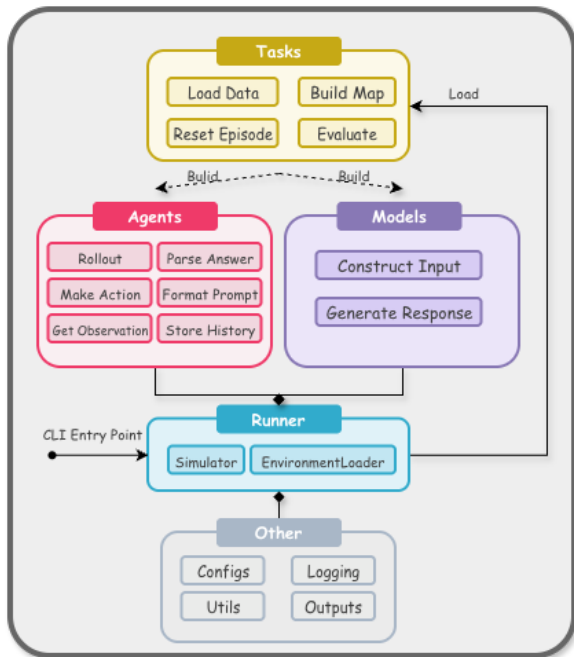


Figure 1: A high-level structure for the benchmark, centered on the interplay between **Tasks**, **Agents**, and **Models**.

Model This component serves as a unified interface for various MLLMs, handling model-specific API calls. It allows users to integrate from open-source models to proprietary APIs without altering the agent logic or evaluation protocol.

Task This component encapsulates the specific navigation challenge and manage dataset splits. By treating the task as a modular input, our framework supports diverse navigation tasks within a unified evaluation protocol.

Agent The Agent acts as the decision-making module that mediates the interaction between the MLLM and the Task. Its role is to embed environmental observations into structured prompts and parse the model’s textual output into executable actions. To rigorously test different cognitive capabilities, we implement modular agent designs varying in memory and reasoning. We compare agents using text summarized action histories (Zhou et al., 2024) against those building topological text maps (Chen et al., 2024) to test long-term spatial grounding. Similarly, to probe planning depth, we evaluate baselines ranging from direct prediction to those employing Chain-of-Thought (CoT) (Wei et al., 2022) and self-reflection (Yao et al., 2022). Detailed agent workflow and prompt templates are provided in the Appendix D.

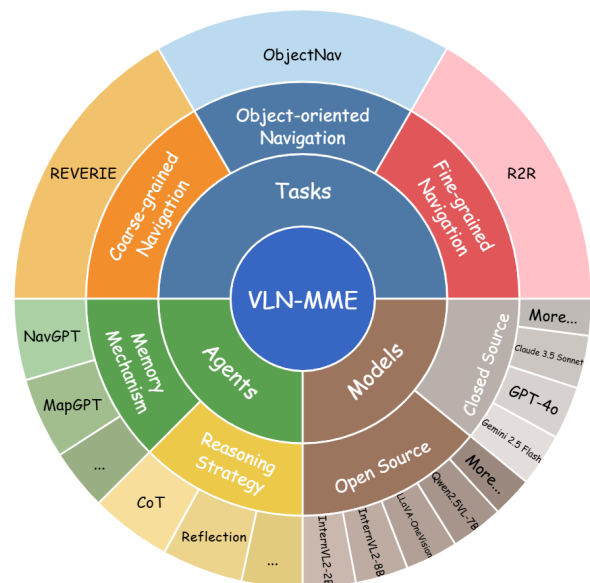


Figure 2: The composition of the VLN-MME.

Extensibility and Orchestration To ensure modularity, we adopt a unified **factory pattern** for instantiating all three component types. By reg-

istering each component with a unique identifier resolved at runtime, this design enables true “plug-and-play” extensibility: integrating a new agent or model requires only a simple registration, preserving the integrity of the core evaluation logic.

The evaluation lifecycle is coordinated by a central Runner module, which ensures reproducibility through a lightweight configuration system. The Runner manages the initialization of the simulator-free environment (Section 3.3) and the dynamic loading of datasets (Section 3.2). During execution, it acts as the intermediary, facilitating the exchange of observations and actions between the agent and environment while logging interactions for granular post-hoc analysis. This architecture strictly separates high-level reasoning capabilities from low-level execution details, reinforcing the framework’s modularity.

3.2 Dataset Construction for Efficient Evaluation

To address the computational challenges of evaluating large models on existing, large-scale VLN datasets and to facilitate rapid experimentation, we constructed a curated benchmark for efficient yet representative evaluation. Following the broader definition of VLN (Zheng et al., 2023; Zhou et al., 2025b), our benchmark is composed of samples carefully drawn from the validation unseen splits of three main datasets: R2R (Anderson et al., 2018), REVERIE (Qi et al., 2020), and ObjectNav (Batra et al., 2020). Detailed descriptions of these datasets and the rationale behind their selection are provided in the Appendix A.

Our construction employs stratified sampling to preserve diversity across three key axes: **scene complexity**, **path difficulty**, and **linguistic richness**. For instance, in R2R and REVERIE, we stratify episodes by Matterport3D scan ID to capture environmental variance, then sample proportionally from path length bins to maintain the difficulty distribution. Linguistic diversity is ensured by randomly selecting one of the three available natural language instructions for each trajectory. For ObjectNav, we additionally enforce a balanced distribution of target object categories. This process yields a compact benchmark that statistically mirrors the characteristics of the original datasets. To further validate the fidelity of our constructed benchmark, we evaluated several previous methods from VLN specialist to Finetuned MLLM, on both the full val_unseen splits and our curated bench-

mark for R2R and REVERIE (all the details are provided in Appendix C). The results reveal a strong correlation in performance. Key metrics such as Success Rate (SR) and Success weighted by Path Length (SPL) on our benchmark closely track the performance on the full splits, with deviations typically within a 2-3 percentage point margin. This close alignment confirms that our stratified sampling approach successfully captures the intrinsic difficulty and diversity of the original datasets, establishing our benchmark as a reliable and efficient proxy for full-scale MLLM evaluation.

3.3 Simulator-Free Environment Design

While high-fidelity simulators are essential for rendering, their computational overhead creates a bottleneck for real-time evaluation. To address this, we introduce a **simulator-free** mode that implements a strategic “space-for-time” trade-off, decoupling evaluation from the rendering engine by substituting expensive GPU operations with efficient disk I/O. As detailed in Table 1, this approach drastically lowers the hardware barrier: by loading pre-cached images instead of full 3D scene geometry and textures, we reduce RAM usage from $\sim 10\text{GB}$ to $\sim 1.7\text{GB}$ ($6\times$ lower), enabling research on consumer-grade hardware. Furthermore, replacing on-the-fly rendering with direct image retrieval accelerates observation access by nearly $9\times$ (0.14s to 0.016s), shortening total episode duration by 20–30 seconds. Here, t denotes the time cost of a single navigation step within an episode, while T denotes the total time cost of a complete episode. Both quantities depend on the total number of steps as well as the specific model and agent configuration used. This efficiency is further reinforced by our discrete navigation graph, which removes the need for continuous waypoint prediction typical of VLN-CE, thereby streamlining the evaluation loop without sacrificing 3D spatial assessment.

Table 1: **Efficiency Comparison.** Resource usage and runtime of our simulator-free, pre-rendered approach versus Habitat, showing substantially lower memory consumption and faster evaluation.

Metric	Ours	Habitat	Delta
RAM Usage	$\sim 1.7\text{ GB}$	$\sim 10\text{ GB}$	5.9× Lower
Obs. Access	$\sim 0.016\text{s}$	$\sim 0.14\text{s}$	8.8× Faster
Time / Step	t	$t + 1.5\text{s}$	1.5s Faster
Time / Episode	T	$T + 25\text{s}$	$\sim 25\text{s}$ Faster

The core of this design is a standardized visual

representation optimized for MLLM perception. Rather than using distorted equirectangular projections, we capture the panoramic context as a set of four non-overlapping perspective images, each with a 90° Field of View. This format preserves visual fidelity and aligns better with the pre-training distributions of standard vision encoders. The rationale for this specific projection strategy is detailed further in Appendix C. Within these views, navigable directions are annotated with distinct numerical markers derived from the underlying connectivity graph, sorting candidates by their global heading to provide a consistent spatial reference.

To bridge the gap between visual features and symbolic reasoning, we augment the environment with semantic metadata. We employ GPT-5 to generate descriptive captions for both the general scene and specific navigation candidates. These descriptions serve as auxiliary inputs for agents that rely on semantic memory; the generation prompts and validation of these captions are documented in the Appendix B. Finally, to ensure immediate accessibility, we host all pre-rendered assets on open-source platforms. Our framework manages the automatic retrieval of these artifacts, enabling users to evaluate without complex simulator installation.

4 Experiments

4.1 Settings

Evaluation Metrics. In this work, we focus exclusively on the navigation component of all tasks, without considering object grounding in REVERIE. We adopt a standard set of navigation metrics to evaluate agent performance: (1) *Trajectory Length* (TL), which measures the average path length in meters; (2) *Navigation Error* (NE), the average distance between the agent’s final position and the goal location; (3) *Success Rate* (SR), the percentage of episodes where the final location is within 3 meters of the target; (4) *Oracle Success Rate* (OSR), the success rate assuming an optimal stopping policy; (5) *Success weighted by Path Length* (SPL) (Jain et al., 2019), which combines success with path efficiency; (6) *Normalized Dynamic Time Warping* (nDTW) (Ilharco et al., 2019), which measures the trajectory similarity to the ground truth path; (7) *Success weighted by normalized DTW* (SDTW), a combined metric capturing both goal-reaching and trajectory fidelity; and (8) *Coverage weighted by Length Score* (CLS) (Jain et al., 2019), which evaluates how well the predicted trajectory

covers the reference path while also accounting for path length efficiency.

Implementation Details We evaluate two proprietary and four open-source MLLMs in a zero-shot setting: GPT5, Gemini2.5 Pro, Qwen2.5-VL-7B (Bai et al., 2025), InternVL3-2/8B (Zhu et al., 2025), LLaVA-One-Vision-7B (Li et al., 2024a). These models are integrated into eight distinct agent configurations, categorized into two primary classes: agents using text summarization and text map as memory. Each class includes four variants: a baseline, one with CoT prompting, one with reflection-based reasoning, and one featuring both CoT and reflection. All open-source MLLMs are served using the vLLM backend (Kwon et al., 2023) to ensure efficient inference and memory management. We assess their performance on all the tasks in our benchmark. Additionally, we compare these zero-shot agents against previously finetuned VLN specialist and finetuned MLLM agents on the R2R and REVERIE tasks, evaluating performance across both the full dataset from prior evaluation methods and our benchmark. All experiments are conducted on a single NVIDIA A100 GPU with 40GB VRAM.

4.2 Performance Analysis

We evaluate our zero-shot MLLM-based agents against prior state-of-the-art methods, analyzing performance across different model architectures, reasoning strategies, and navigation tasks.

Model Capabilities. Our results, detailed in Table 2, indicate that proprietary models like GPT-5 and Gemini-2.5 Pro currently establish the upper bound for zero-shot navigation. However, among open-weights models, Qwen2.5-VL-7B demonstrates remarkable robustness, consistently outperforming peers such as LLaVA-OneVision and InternVL3-2B. For instance, in the baseline NavGPT configuration on the Fine-Grained task, Qwen2.5-VL-7B achieves a Success Rate (SR) of 27.5%, substantially surpassing LLaVA-OneVision (11.5%) and InternVL3-2B (13.5%). This positions Qwen2.5-VL as a strong and capable baseline for the open-source embodied community.

Agent Architecture and Reasoning Strategies. Counterintuitively, the integration of advanced prompting strategies like CoT or reflection does not consistently yield performance improvements and often proves detrimental. As shown in Table 2

Agent / MLLM	Fine-Grained Navigation								Coarse-grained Navigation								Object-Oriented Navigation							
	TL↓	NE↓	nDTW↑	SDTW↑	CLS↑	OSR↑	SR↑	SPL↑	TL↓	NE↓	nDTW↑	SDTW↑	CLS↑	OSR↑	SR↑	SPL↑	TL↓	NE↓	nDTW↑	SDTW↑	CLS↑	OSR↑	SR↑	SPL↑
Text Summarization Memory Agents																								
NavGPT																								
GPT-5	9.12	6.28	37.96	22.02	34.11	57.50	38.50	29.23	9.37	8.03	27.11	14.88	21.05	49.50	30.00	20.76	9.63	4.41	26.52	16.33	20.13	79.00	48.00	23.84
Gemini-2.5 Pro	8.94	6.17	38.07	22.43	34.62	60.50	41.00	32.67	9.21	7.89	27.47	15.13	21.38	52.00	33.50	24.38	9.48	4.34	26.78	16.54	20.43	79.00	51.50	27.19
InternVL3-2B	9.89	8.56	21.25	5.75	21.59	27.00	13.50	5.46	10.13	10.18	15.30	2.97	17.17	16.33	7.33	2.50	10.35	6.27	13.67	2.55	15.57	40.50	21.50	3.57
InternVL3-8B	11.74	7.55	25.28	13.38	26.22	50.50	28.00	12.61	11.90	9.25	17.32	8.18	18.84	30.67	20.00	7.18	11.55	4.63	14.51	5.09	17.36	56.00	39.00	7.69
LLaVA-OV-7B	8.04	8.40	26.34	5.70	25.53	20.50	11.50	4.94	9.85	9.35	19.27	6.09	18.39	20.00	14.67	5.19	9.52	5.93	16.54	5.11	14.71	41.00	27.50	4.51
Qwen2.5-VL-7B	8.54	6.99	35.97	18.88	34.85	44.00	27.50	17.11	8.94	8.55	23.79	9.88	23.55	27.33	18.67	9.00	9.07	4.65	21.83	10.63	23.52	56.50	37.50	13.18
NavGPT w/ CoT																								
GPT-5	9.47	6.42	39.17	18.89	32.54	54.50	36.00	27.31	9.71	8.24	32.53	12.04	26.82	46.00	28.50	20.42	9.92	4.58	31.06	14.03	25.07	70.00	46.50	23.67
Gemini-2.5 Pro	9.31	6.29	39.64	19.12	32.86	50.00	32.50	23.88	9.56	8.11	32.83	12.22	27.14	41.50	24.00	16.93	9.76	4.49	31.38	14.24	25.31	65.50	42.00	19.27
InternVL3-2B	6.21	8.84	26.15	4.28	27.45	14.50	8.00	4.47	4.98	9.87	23.39	3.11	25.09	9.00	5.33	3.24	6.78	5.56	25.51	5.34	23.98	41.00	25.00	6.66
InternVL3-8B	9.07	7.56	29.22	10.90	29.59	35.50	19.00	10.95	7.96	9.39	24.93	9.14	26.40	22.00	15.33	9.31	6.43	5.31	27.22	10.30	27.35	43.50	34.50	12.67
LLaVA-OV-7B	7.97	8.85	23.20	5.60	24.43	22.50	12.50	5.41	8.43	9.47	21.09	7.13	21.38	20.33	14.00	5.90	9.69	5.66	15.90	6.24	17.92	47.50	33.50	7.31
Qwen2.5-VL-7B	9.04	7.97	30.23	11.56	31.36	37.50	21.00	11.41	8.29	9.85	24.02	8.82	25.59	24.33	15.67	8.10	6.86	5.13	28.63	12.54	28.32	44.50	33.00	13.25
NavGPT w/ Reflection																								
GPT-5	9.61	6.37	40.04	19.53	33.14	50.50	33.00	24.18	9.82	8.13	33.03	12.51	27.08	43.00	25.00	17.34	10.03	4.47	30.02	15.06	24.09	68.00	43.00	20.62
Gemini-2.5 Pro	9.44	6.23	40.39	19.71	33.47	55.00	37.50	28.73	9.67	7.98	33.34	12.72	27.33	47.50	29.00	21.68	9.88	4.38	30.27	15.21	24.34	72.50	47.50	25.14
InternVL3-2B	6.50	8.94	29.91	5.00	30.13	16.00	8.00	5.20	7.20	9.41	25.49	4.60	24.93	15.00	8.50	4.80	6.43	5.56	26.53	6.14	25.28	43.00	28.00	7.83
InternVL3-8B	4.54	7.99	34.25	8.89	34.44	19.00	12.00	9.53	6.82	10.20	23.58	6.04	27.79	17.33	11.00	5.97	8.35	5.02	18.79	6.68	22.14	51.00	32.50	9.12
LLaVA-OV-7B	2.81	8.01	38.17	8.43	38.39	11.00	10.50	9.44	5.15	9.34	28.04	5.91	30.22	14.67	9.33	5.82	9.35	5.58	15.48	6.11	17.05	47.50	34.00	7.90
Qwen2.5-VL-7B	6.93	7.17	36.51	16.44	33.59	32.50	24.00	14.95	6.96	8.76	26.88	7.97	25.78	18.00	12.00	7.97	7.55	5.06	23.33	11.04	25.23	50.50	35.50	14.67
NavGPT w/ CoT & Reflection																								
GPT-5	9.52	6.31	39.56	20.08	33.02	56.50	38.50	29.81	9.77	8.18	32.04	13.07	26.03	48.50	30.00	22.17	9.97	4.36	29.08	15.51	23.06	73.50	48.50	25.92
Gemini-2.5 Pro	9.36	6.19	39.83	20.24	33.27	52.00	34.00	25.74	9.61	8.04	32.32	13.21	26.24	43.50	25.50	18.36	9.81	4.27	29.33	15.66	23.22	68.00	43.50	21.28
InternVL3-2B	7.15	9.24	22.45	2.05	23.47	15.00	4.50	1.70	7.30	9.78	22.33	5.30	24.07	15.00	9.33	4.63	6.94	6.25	21.29	6.11	21.47	37.50	24.50	7.26
InternVL3-8B	7.22	7.47	36.62	16.18	35.98	32.50	22.00	15.33	8.95	9.07	24.11	10.24	25.78	28.67	17.33	10.07	9.18	5.30	18.59	5.99	21.45	51.50	32.50	8.14
LLaVA-OV-7B	7.61	8.48	28.01	6.31	26.32	22.00	10.00	5.83	8.44	8.68	24.73	8.31	22.60	22.00	14.00	6.78	8.55	5.66	21.26	6.74	19.69	44.00	28.50	7.25
Qwen2.5-VL-7B	7.82	7.53	34.86	17.65	34.80	38.50	25.50	17.68	7.19	9.48	27.07	7.87	28.22	18.00	11.67	7.89	7.60	5.39	26.52	11.49	26.98	47.00	36.00	13.67
Text Map Memory Agents																								
MapGPT																								
GPT-5	9.32	6.21	39.03	21.08	33.05	52.50	34.00	25.83	9.53	7.94	29.07	13.02	23.04	45.00	26.00	18.29	9.73	4.33	28.06	16.08	22.01	74.00	44.00	20.91
Gemini-2.5 Pro	9.17	6.09	39.34	21.22	33.23	57.50	39.50	30.72	9.38	7.81	29.26	13.17	23.21	50.00	31.50	23.16	9.58	4.26	28.24	16.19	22.13	79.50	49.50	26.24
InternVL3-2B	9.84	8.61	20.18	3.85	21.89	22.50	11.00	3.71	10.19	9.59	18.10	5.24	19.97	19.00	12.00	4.41	10.35	5.93	13.59	3.69	15.57	46.50	27.50	4.41
InternVL3-8B	6.78	7.70	34.34	13.06	33.78	32.00	18.00	12.46	7.26	9.16	26.63	8.04	27.62	22.33	13.67	7.87	5.95	5.26	28.03	8.98	27.39	44.50	31.50	11.61
LLaVA-OV-7B	4.97	8.44	31.70	5.29	32.62	15.50	8.50	5.59	8.58	8.96	23.33	7.33	22.39	22.00	14.67	6.48	7.92	5.86	20.66	4.63	18.34	35.50	22.50	4.28
Qwen2.5-VL-7B	8.16	7.13	34.31	17.39	33.37	38.00	26.00	17.31	10.52	8.53	21.27	10.85	20.78	32.33	21.67	8.96	9.77	4.82	20.06	9.06	22.13	52.50	36.50	11.05
MapGPT w/ CoT																								
GPT-5	9.62	6.51	40.09	17.04	32.08	50.00	32.50	27.14	9.83	8.31	34.05	10.06	28.03	42.50	24.50	19.68	9.93	4.67	32.08	13.01	26.02	68.50	44.50	23.21
Gemini-2.5 Pro	9.46	6.38	40.36	17.17	32.24	48.50	31.00	23.37	9.68	8.19	34.23	10.14	28.22	40.00	23.00	15.84	9.78	4.59	32.27	13.16	26.21	62.50	39.50	18.76
InternVL3-2B	5.09	8.99	26.43	3.94	27.62	14.00	6.50	4.12	5.22	9.88	23.01	1.77	24.48	8.00	4.00	1.58	6.66	6.31	21.14	4.69	21.05	35.50	20.00	4.76
InternVL3-8B	5.96	8.79	31.15	8.64	33.11	22.50	12.00	9.66	5.82	9.07	30.01	9.09	30.77	17.33	13.33	8.77	6.55	5.34	26.74	9.38	27.22	46.50	34.00	12.33
LLaVA-OV-7B	8.34	8.96	19.61	2.84	19.10	13.50	8.50	2.83	7.82	9.63	20.22	4.10	21.21	13.00	8.67	3.38	7.64	6.43	19.11	3.21	18.63	31.00	16.00	4.14
Qwen2.5-VL-7B	7.45	7.78	30.38	10.86	29.67	28.50	17.00	10.47	6.48	8.49	30.81	11.11	31.34	21.00	16.33	10.61	9.51	4.78	22.14	8.65	23.29	54.00	32.00	9.95
MapGPT w/ Reflection																								
GPT-5	9.72	6.41	41.02	18.09	33.03	54.00	36.50	28.19	9.91	8.26	35.06	11.04	29.07	42.50	25.50	20.73	10.02	4.62	31.09	14.05	25.08	69.50	45.50	23.65
Gemini-2.5 Pro	9.57	6.28	41.28	18.16	33.21	49.50	32.00	24.31	9.77	8.14	35.27	11.13	29.25	41.50	24.00	16.88	9.87	4.54	31.24	14.17	25.23	64.00	41.00	19.94
InternVL3-2B	2.37	8.55	32.83	2.97	33.54	6.00	4.00	3.26	2.45	9.58	27.97	2.81	31.14	5.00	3.67	3.31	9.72	5.85	15.52	3.04	17.39	44.00	25.00	4.50
InternVL3-8B	5.85	7.80	36.05	11.48	35.18	28.50	16.50	10.89	6.49	8.80	27.19	6.91	28.24	19.33	12.67	6.64	6.23	5.50	26.92	8.12	25.74	40.50	30.00	10.36
LLaVA-OV-7B	8.35	8.50	26.62	5.50	23.33	20.00	10.00	5.50	8.20	9.47	25.84	6.00	25.16	19										

this conclusion.

Task Difficulty Hierarchy. Finally, the results reveal a clear difficulty hierarchy across navigation categories. Object-Oriented Navigation proves the most tractable, with agents consistently achieving higher success rates. Fine-Grained Navigation presents moderate difficulty, while Coarse-Grained Navigation emerges as the most challenging task for MLLMs. This suggests that interpreting high-level instructions to navigate towards less precisely defined locations remains a substantial hurdle for current zero-shot reasoning.

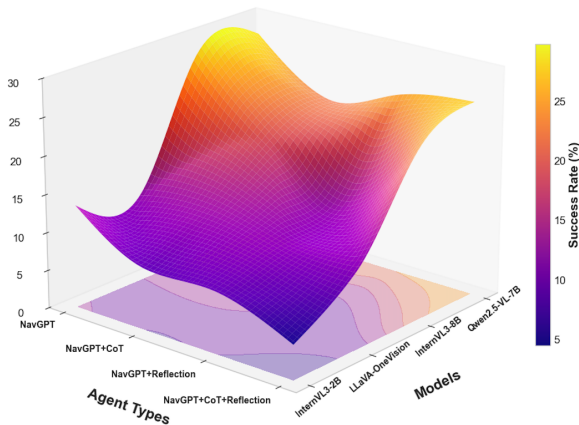


Figure 3: Performance comparison of agents using text summarization memory under different reasoning strategies across multiple backbone MLLMs.

4.3 Discussion

As discussed in section 4.2, we reveal some counterfactual behavior when MLLMs perform embodied navigation. We further conduct an error analysis to understand their error pattern and find that they are hindered by fundamental limitations across several cognitive dimensions (see more details in Appendix E). Interestingly, we find that the high navigation failure rate is overwhelmingly dominated by looping behaviors, shown in Figure 4. It is not a superficial issue but symptomatic of deeper challenges in instruction fidelity, spatial reasoning, historical context utilization, and the grounding of multi-modal perception into action. We discuss these three interconnected aspects below.

Instruction Following and Reasoning Fidelity.

A primary challenge is the limited fidelity with which MLLMs adhere to complex instructions, particularly those governing their reasoning process. While the models can follow basic output formatting prompts, they struggle with more abstract

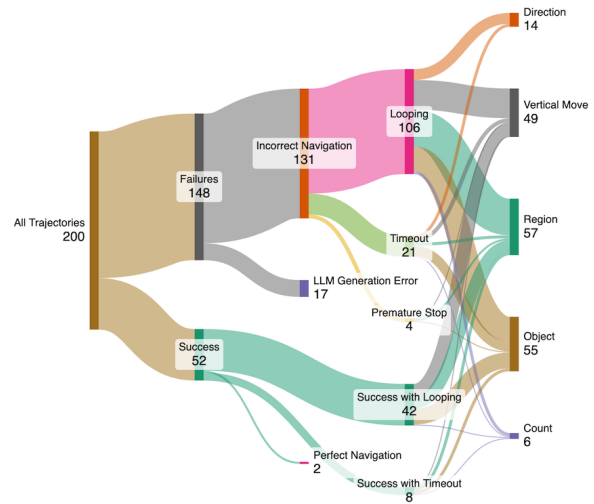


Figure 4: A high-level analysis of success and failure modes for Qwen2.5-VL-7B model using an agent with text map memory.

meta-instructions. For instance, when prompted with CoT or reflection mechanisms to explicitly “reason based on history and the map,” the agents often diverge, reverting to a reactive, myopic reasoning pattern that ignores the very context they were instructed to use. This disconnect helps explain why adding CoT and reflection did not consistently improve performance (Table 2); the models did not faithfully execute the intended reasoning strategy. This suggests a significant gap between simply conditioning a model on a prompt and instilling a robust, procedural reasoning capability.

Spatial and Environmental Understanding.

Our fine-grained error analysis reveals that profound weaknesses in spatial understanding are the root cause of most navigational failures. Of 131 errors analyzed, a staggering 106 were due to persistent looping, a direct consequence of the model’s inability to ground instructions in the 3D environment. This manifests in specific, recurring issues like poor region recognition, failure to reason about verticality on stairs, and basic directional confusion. The fact that providing an explicit topological map failed to yield significant gains highlights a deeper problem: the agent cannot connect abstract spatial knowledge to its visual perception and actions. Furthermore, the agent critically fails at sequential decision-making, which is essential for navigation. The rampant looping behavior clearly shows that the agent does not learn from its trajectory to avoid repeating mistakes. This is not a problem of memory capacity, as the history rarely exceeds the

model’s context window, but rather one of memory utilization. The model has access to its past actions but cannot ground its current decisions in that history to self-correct. In fact, the observation that simpler history formats can outperform complex ones suggests that too much historical information creates a cognitive load, confusing the agent instead of guiding it.

Perception-Action Grounding. Finally, we observe a critical gap between multimodal perception and embodied action. The MLLM’s visual grounding is functional at a recognition level; for example, it can often correctly identify a “staircase” or a target “chair” in its textual reasoning trace. This indicates that the visual and language modalities are connected. However, this recognition consistently fails to translate into correct action. The agent sees the stairs but walks past them in a loop. It may even get very close to the goal, demonstrating it has successfully grounded the target object visually, yet fails to execute the final ‘STOP’ action. This is powerfully illustrated by our success-case analysis, where most successful episodes involved inefficient looping near the target before stopping. This “perception-action gap” shows that the greatest challenge for MLLMs in VLN is not just seeing and describing the world, but effectively acting within it.

4.4 Diagnostic Case Study

To demonstrate the utility of our benchmark as a scientific tool for uncovering MLLM capabilities, we conducted an in-depth investigation on a specific subset of “Hard Negatives.” This subset consists of 25 trajectories from the fine-grained task where all open-source models consistently failed using the standard text-summary memory agent. We revisited these failure cases using Qwen2.5VL-7B as the primary navigator and designed two experimental settings to isolate the source of the failure:

(1) Oracle-Guided Navigation. To test if the failure was due to a lack of visual understanding or a lack of high-level reasoning, we introduced a stronger model (Qwen3VL-4B (Yang et al., 2025a)) acting as an Oracle Assistant. When the navigator struggled (*e.g.*, looping, moving in a critically wrong direction, or entering the wrong region), it was allowed to query the Oracle. The Oracle provided high-level reasoning guidance and suggested potential plans and actions. The navigator achieved a significant success rate with this reasoning sup-

port as shown in Table 3, suggesting that the base model possesses the fundamental navigational capability but lacks the high-level planning or error-correction logic required for these hard cases.

Table 3: Effect of Oracle Assistance on Hard Negatives.

Method	SR ↑	OSR ↑	SPL ↑
Baseline (Qwen2.5VL-7B)	0.00	0.00	0.00
+ Oracle Assistant (Qwen3VL)	52.00	68.00	41.28

(2) Failure-Aware In-Context Learning. To test if the model could self-correct given awareness of common pitfalls, we replaced the zero-shot evaluation with a few-shot setup. We provided the model with N examples of potential failure cases identified in our error analysis ($N = \{1, 2, 3\}$). As shown in Table 4, providing these “negative” examples in the prompt yielded a performance increase. However, the gains were modest compared to the Oracle intervention, indicating that while awareness of failure modes helps, active reasoning support is more critical for solving complex navigation challenges.

Table 4: Effect of Failure-Aware Few-Shot Prompting on Hard Negatives.

Method	SR ↑	OSR ↑	SPL ↑
Zero-shot (Baseline)	0.00	0.00	0.00
1-shot Failure Example	12.00	24.00	9.42
2-shot Failure Examples	16.00	28.00	11.25
3-shot Failure Examples	16.00	32.00	14.29

5 Conclusion

In this work, we introduce VLN-MME, a modular, simulator-free framework for diagnosing MLLMs as zero-shot embodied agents. Our evaluation reveals a critical “perception-action gap”: models possess robust visual grounding yet fail at sequential decision-making, with Chain-of-Thought strategies often proving detrimental due to limited context awareness. Crucially, our diagnostic study on hard negatives confirms that these failures stem primarily from deficient strategic planning rather than perceptual limitations, as evidenced by substantial performance recovery under oracle guidance. Ultimately, VLN-MME highlights that future MLLM development must prioritize bridging the gap between static visual understanding and dynamic, history-aware strategic reasoning.

Limitation

We present and discuss the limitations of our work to outline the scope of our analysis. First, due to computational and temporal constraints, our evaluation covers a representative selection of MLLMs and agent designs rather than an exhaustive spectrum. However, we mitigate this by releasing a highly modular interface that enables the community to seamlessly integrate and benchmark emerging models. Second, while our framework supports broader capabilities such as dialog-based navigation and multi-language instructions, we restricted our current analysis to standard instruction following to establish a rigorous baseline. We designed the dataset construction pipeline to be extensible, allowing future researchers to easily incorporate these diverse tasks. Finally, this work is primarily diagnostic: we identify critical deficiencies in MLLM 3D spatial reasoning and strategic planning but do not propose specific algorithmic remedies. We posit these findings as a strategic guide for future work in embodied MLLM post-training and agent design.

References

- Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. 2020. ObjectNav Revisited: On Evaluation of Embodied Agents Navigating to Objects. In *arXiv:2006.13171*.
- Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3d: Learning from rgb-d data in indoor environments. In *2017 International Conference on 3D Vision (3DV)*, pages 667–676. IEEE.
- Fu Chaoyou, Chen Peixian, Shen Yunhang, Qin Yulei, Zhang Mengdan, Lin Xu, Yang Jinrui, Zheng Xiawu, Li Ke, Sun Xing, and 1 others. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 3.
- Jiaqi Chen, Bingqian Lin, Ran Xu, Zhenhua Chai, Xiaodan Liang, and Kwan-Yee K Wong. 2024. Mapgpt: Map-guided prompting for unified vision-and-language navigation. *arXiv preprint arXiv:2401.07314*.
- Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. 2021. History aware multimodal transformer for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 34:5834–5847.
- Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. 2022. 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*, pages 16537–16547.
- An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Zaitian Gongye, Xueyan Zou, Jan Kautz, Erdem Biyik, Hongxu Yin, Sifei Liu, and Xiaolong Wang. 2024. Navila: Legged robot vision-language-action model for navigation. *arXiv preprint arXiv:2412.04453*.
- Zhili Cheng, Yuge Tu, Ran Li, Shiqi Dai, Jinyi Hu, Shengding Hu, Jiahao Li, Yang Shi, Tianyu Yu, Weize Chen, and 1 others. 2025. Embodiedeval: Evaluate multimodal llms as embodied agents. *arXiv preprint arXiv:2501.11858*.
- Erik Daxberger, Nina Wenzel, David Griffiths, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfei Yang, Afshin Dehghan, and 1 others. 2025. Mm-spatial: Exploring 3d spatial understanding in multimodal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7395–7408.
- Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36:28091–28114.
- Xiangyu Dong, Haoran Zhao, Jiang Gao, Haozhou Li, Xiaoguang Ma, Yaoming Zhou, Fuhai Chen, and Juan Liu. 2025. Se-vln: A self-evolving vision-language navigation framework based on multimodal large language models. *arXiv preprint arXiv:2507.13152*.
- Hao Fei, Yuan Zhou, Juncheng Li, Xiangtai Li, Qingshan Xu, Bobo Li, Shengqiong Wu, Yaoting Wang, Junbao Zhou, Jiahao Meng, and 1 others. 2025. On path to multimodal generalist: General-level and general-bench. In *Forty-second International Conference on Machine Learning*.
- Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. 2021. A recurrent vision-and-language bert for navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1643–1653.
- Wenbo Hu, Yining Hong, Yanjun Wang, Leison Gao, Zibu Wei, Xingcheng Yao, Nanyun Peng, Yonatan

- Bitton, Idan Szpektor, and Kai-Wei Chang. 2025. 3dllm-mem: Long-term spatial-temporal memory for embodied 3d large language model. *arXiv preprint arXiv:2505.22657*.
- Gabriel Ilharco, Vihan Jain, Alexander Ku, Eugene Ie, and Jason Baldridge. 2019. General evaluation for instruction conditioned navigation using dynamic time warping. *arXiv preprint arXiv:1907.05446*.
- Vihan Jain, Gabriel Magalhaes, Alexander Ku, Ashish Vaswani, Eugene Ie, and Jason Baldridge. 2019. Stay on the path: Instruction fidelity in vision-and-language navigation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1862–1872.
- Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. 2020. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *European Conference on Computer Vision*, pages 104–120. Springer.
- Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. 2020. 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)*, pages 4392–4412.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.
- Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Wensi Ai, Benjamin Martinez, and 1 others. 2024b. Behavior-1k: A human-centered, embodied ai benchmark with 1,000 everyday activities and realistic simulation. *arXiv preprint arXiv:2403.09227*.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.
- Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024c. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.
- Yuan-Hong Liao, Rafid Mahmood, Sanja Fidler, and David Acuna. 2024. Reasoning paths with reference objects elicit quantitative spatial reasoning in large vision-language models. *arXiv preprint arXiv:2409.09788*.
- Bingqian Lin, Yunshuang Nie, Ziming Wei, Jiaqi Chen, Shikui Ma, Jianhua Han, Hang Xu, Xiaojun Chang, and Xiaodan Liang. 2024. Navcot: Boosting llm-based vision-and-language navigation via learning disentangled reasoning. *arXiv preprint arXiv:2403.07376*.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.
- Yuxing Long, Xiaoqi Li, Wenzhe Cai, and Hao Dong. 2023. Discuss before moving: Visual language navigation via multi-expert discussions. *arXiv preprint arXiv:2309.11382*.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.
- Bowen Pan, Rameswar Panda, SouYoung Jin, Rogerio Feris, Aude Oliva, Phillip Isola, and Yoon Kim. 2023. Langnav: Language as a perceptual representation for navigation. *arXiv preprint arXiv:2310.07889*.
- Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. 2020. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9982–9991.
- Yanyuan Qiao, Haodong Hong, Wenqi Lyu, Dong An, Siqi Zhang, Yutong Xie, Xinyu Wang, and Qi Wu. 2025. Navbench: Probing multimodal large language models for embodied navigation. *arXiv preprint arXiv:2506.01031*.
- Yanyuan Qiao, Yuankai Qi, Zheng Yu, Jing Liu, and Qi Wu. 2023. March in chat: Interactive prompting for remote embodied referring expression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15758–15767.
- Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, and 1 others. 2019. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9339–9347.

- Xijia Tao, Yihua Teng, Xinxing Su, Xinyu Fu, Jihao Wu, Chaofan Tao, Ziru Liu, Haoli Bai, Rui Liu, and Lingpeng Kong. 2025. Mmsearch-plus: A simple yet challenging benchmark for multimodal browsing agents. *arXiv e-prints*, pages arXiv-2508.
- Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2020. Vision-and-dialog navigation. In *Conference on Robot Learning*, pages 394–406.
- Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. 2024. Appworld: A controllable world of apps and people for benchmarking interactive coding agents. *arXiv preprint arXiv:2407.18901*.
- Weixuan Wang, Dongge Han, Daniel Madrigal Diaz, Jin Xu, Victor Rühle, and Saravan Rajmohan. 2025. Odysseybench: Evaluating llm agents on long-horizon complex office application workflows. *arXiv preprint arXiv:2508.09124*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.
- Peiran Xu, Sudong Wang, Yao Zhu, Jianing Li, and Yunjian Zhang. 2025. Spatialbench: Benchmarking multimodal large language models for spatial cognition. *arXiv preprint arXiv:2511.21471*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2025b. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643.
- Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, and 1 others. 2025c. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. *arXiv preprint arXiv:2502.09560*.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.
- Weihao Yu, Zhengyuan Yang, Lingfeng Ren, Linjie Li, Jianfeng Wang, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Lijuan Wang, and Xinchao Wang. 2024. Mm-vet v2: A challenging benchmark to evaluate large multimodal models for integrated capabilities. *arXiv preprint arXiv:2408.00765*.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.
- Jiazhao Zhang, Kunyu Wang, Shaoan Wang, Minghan Li, Haoran Liu, Songlin Wei, Zhongyuan Wang, Zhizheng Zhang, and He Wang. 2024a. Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks. *arXiv preprint arXiv:2412.06224*.
- Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and Wang He. 2024b. Navid: Video-based vlm plans the next step for vision-and-language navigation. *arXiv preprint arXiv:2402.15852*.
- Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. 2023. Towards learning a generalist model for embodied navigation. *arXiv preprint arXiv:2312.02010*.
- Gengze Zhou, Yicong Hong, Zun Wang, Xin Eric Wang, and Qi Wu. 2025a. Navgpt-2: Unleashing navigational reasoning capability for large vision-language models. In *European Conference on Computer Vision*, pages 260–278. Springer.
- Gengze Zhou, Yicong Hong, Zun Wang, Chongyang Zhao, Mohit Bansal, and Qi Wu. 2025b. Same: Learning generic language-guided visual navigation with state-adaptive mixture of experts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7794–7807.
- Gengze Zhou, Yicong Hong, and Qi Wu. 2024. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7641–7649.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internv13: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Appendix

In this supplementary material, we aim to provide additional details to support the main content of our paper: Section A provides the rationale for selecting datasets across different levels of instruction granularity and defines the experimental scope and context. Section B details the process for generating semantic annotations for the environment. Section C presents the construction and rationale of agent-centric visual observations, supported by ablation studies and empirical evidence demonstrating the representativeness of the resulting benchmark. Section D explains the detailed workflow, action parsing, and complete prompt structure used for all agent variants. Section E introduces the custom tool for trajectory visualization and analysis, provides a detailed quantitative analysis of agent failures and successes, and illustrates common agent behaviors through several qualitative case studies. Additionally, Section F includes our declaration on the use of large language models to aid in polishing the manuscript.

A Dataset Selection and Experimental Scope

A.1 Selection Rationale: The Granularity Spectrum

Our primary criterion for dataset selection is to cover the full spectrum of linguistic granularity. To evaluate MLLM spatial reasoning comprehensively, we categorize the instruction \mathcal{W} into three distinct levels of granularity. We select one representative dataset for each level to ensure broad coverage:

1. **Fine-grained VLN:** \mathcal{W} describes the complete sequence of actions and observations $\{s_0, a_0, s_1, a_1, \dots, s_T, a_T\}$ step-by-step.
 - *Selected Dataset:* **R2R** (Anderson et al., 2018). This represents the standard fine-grained task with 22k human-annotated instructions (avg. 32 words) guiding agents along ground-truth paths of approx. 7 steps (10 meters).
2. **Coarse-grained VLN:** \mathcal{W} refers to a remote, out-of-sight target at v_T using high-level linguistic descriptions (e.g., “the cold tap in the first bedroom on level two”).
 - *Selected Dataset:* **REVERIE** (Qi et al., 2020). This represents the coarse-

grained task. It inherits the R2R topology but uses high-level instructions (avg. 21 words) for paths ranging from 4 to 7 steps.

3. **Zero-grained VLN:** \mathcal{W} consists of a single term indicating the target (e.g., an object category), requiring the agent to infer the path without linguistic guidance.

- *Selected Dataset:* **objnav-MP3D** (Batra et al., 2020). We utilize the standard validation split of 11 scenes from the Habitat objnav dataset (Savva et al., 2019) in MP3D (Chang et al., 2017) to represent zero-grained navigation across 21 goal categories.

A.2 Scope and Extensibility

To isolate the reasoning capabilities of MLLMs, we strictly define the scope of this work while highlighting the extensibility of our approach:

Focus on Spatial Reasoning: We operate within discrete action spaces (connectivity graphs) and exclude continuous control (VLN-CE (Krantz et al., 2020)) or physical robot deployment. Furthermore, while REVERIE typically involves object localization, we restrict our evaluation to the navigation success only. This allows us to focus purely on sequential decision-making, spatial reasoning, and understanding action consequences, avoiding the confounding factors of sim-to-real gaps or object detection failures.

Extensibility to Other Tasks: In this work, we prioritize static instructions and exclude interactive datasets (e.g., CVDN (Thomason et al., 2020)) or multilingual variants (e.g., RxR (Ku et al., 2020)). However, our framework is designed to be task-agnostic. Future extensions to interactive or multilingual tasks can be seamlessly embedded into our framework without requiring significant architectural changes.

B Generation of Semantic Annotations

To enrich the agent’s environmental understanding in our simulator-free setup, we generated two types of semantic annotations: descriptive captions for navigable markers and concise summaries for each viewpoint.

B.1 Marker Caption Generation

The visual markers indicating navigable viewpoints in the panoramic images were annotated with short,

descriptive captions. This process provides the agent with crucial semantic cues about the direction of potential paths. We used GPT-5 for this task. For each viewpoint, the model was provided with the panoramic image containing numbered visual markers and prompted to generate a JSON object mapping each marker index to a descriptive sentence. The prompt used for this captioning process is shown in Figure 5.

```
Observe the panoramic images provided, each labeled with distinct markers of green circles (indexed in number). For each marker, briefly specify the area or room it leads toward and describe what visually distinguishes it from the others in a short sentence. Present your response in JSON format, where each marker's index is a key and the corresponding short, descriptive sentence is the value.

For example, if the image contains three markers, the response should look like:
{
  "1": "Decorative partition and dining area; leads toward an interior space or adjacent room.",
  "2": "Chair and decorative cabinet area; leads to a wall-mounted decoration and seating.",
  "3": "Hallway with external view and door; leads towards the entrance or exterior patio."
}
```

Figure 5: GPT5 prompt for generating marker caption

B.2 Viewpoint Summary Generation

In addition to marker captions, a single, holistic summary of the scene was generated for each viewpoint to give map-based agents a global understanding of their current location. For this process, we adopt the same methodology presented in NavGPT (Zhou et al., 2024).

The generation follows a two-stage process. First, initial descriptions are generated for images from a viewpoint using the BLIP-2 model. To elicit descriptions that are rich in object details and relevant to indoor scenes, each image is fed to BLIP-2 with the prompt: “*This is a scene of*”.

As this initial step often produces redundant information across different images of the same viewpoint, a second summarization step is employed. The descriptions generated by BLIP-2 (Li et al., 2023) are consolidated into a single, coherent sentence using a GPT-5 summarizer. The model is prompted with the following template in Figure 6, where “{description}” is replaced by the text from BLIP-2.

```
Here is a single scene view from top, down
and middle: {description}
Summarize the scene in one sentence:
```

Figure 6: GPT-5 prompt for generating viewpoint summarization

C Construction of Agent-Centric Visual Observations

C.1 Panoramic Construction Details

To provide the MLLM with a full 360-degree visual context from the agent’s perspective, we construct a single panoramic image at each step. This process leverages the four pre-rendered, world-oriented images associated with each viewpoint and reorients them based on the agent’s current heading. This method serves as a lightweight, simulator-free proxy for real-time rendering.

Pre-rendered Image Set As described in the main paper, each viewpoint in the environment is associated with four high-resolution images, each with a 90-degree vertical Field of View (vFOV). These images are centered on the four cardinal directions relative to the global coordinate system: 0° (North), 90° (East), 180° (South), and 270° (West).

Heading Correction and Image Selection Since an agent’s heading is continuous (e.g., 60°), it will not always align perfectly with one of the four pre-rendered directions. To resolve this, we implement a heading correction mechanism. The agent’s current continuous heading is first mapped to the closest cardinal direction. This is achieved by quantizing the heading angle to the center of the 90-degree quadrant it falls within. For instance, any agent heading $h \in [45, 135)$ is mapped to the 90° image, which then serves as the agent’s **Front** view.

Panoramic Image Composition Once the **Front** image is determined through heading correction, the remaining three images are assigned to the agent’s relative directions: **Left**, **Right**, and **Back**. These four images are then concatenated horizontally in the following order to form a single panoramic strip: [**Left**, **Front**, **Right**, **Back**].

To ensure the MLLM can correctly interpret this composite view, we explicitly annotate the image by overlaying the corresponding directional labels above each of the four segments, as illustrated in

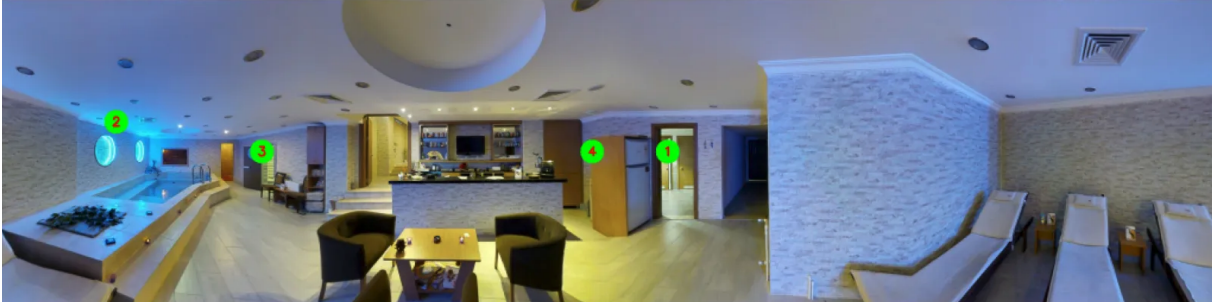


Figure 7: An example of the composite visual observation provided to the MLLM. The four pre-rendered images are stitched together in an agent-centric order (Left, Front, Right, Back) based on the agent’s corrected heading.

Figure 7. This provides a clear, agent-centric visual input that grounds the model in its current orientation.

C.2 Justification of Observation Format

The choice of using a 4-image panoramic observation was not an arbitrary simplification, but a design decision informed by systematic experimentation across four observation formats. To determine the optimal visual input for the MLLM, we evaluated:

1. The standard 36-image panoramic sweep (dense sampling).
2. A 24-image variant tested in prior work (Zhou et al., 2024).
3. A single-image stitched panorama (warped projection).
4. Our proposed 4-image 360° view (90° vFOV each).

To ensure the comparison was meaningful, we conducted these tests on two representative MLLMs (InternVL3-8B and Qwen2.5-VL-7B) on the fine-grained navigation task. As shown in Table 5, our proposed 4-image format consistently outperforms significantly denser representations (24 or 36 images) and the single stitched panorama.

We hypothesize that while 24 or 36 images provide more granular visual information, they significantly increase the token count and complexity, potentially overwhelming the MLLM’s context window or introducing irrelevant visual noise. Conversely, the single stitched panorama likely suffers from projection distortion or loss of fine-grained detail. The 4-image format strikes an optimal balance, providing high-resolution, undistorted views of the cardinal directions without excessive token overhead.

C.3 Benchmark Representativeness and Baseline Comparison

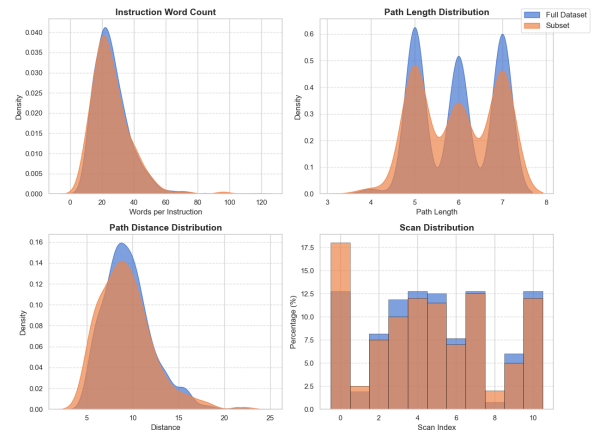


Figure 8: A statistical comparison of our benchmark’s fine-grained task subset data against the original R2R val_unseen split.

To verify the representativeness of our dataset, we first examine the instruction and path characteristics. As illustrated in Figure 8, the data distribution of our sampled benchmark aligns closely with the original R2R val_unseen split, indicating that our subset successfully preserves the intrinsic diversity and complexity of the full dataset.

Complementing this statistical alignment, we evaluated a diverse set of prior methods, ranging from traditional VLN specialists (RecVLNBert (Hong et al., 2021), HAMT (Chen et al., 2021), DUET (Chen et al., 2022)) to finetuned MLLMs (NaviLLM (Zheng et al., 2023), NavGPT2 (Zhou et al., 2025a)), on both the full val_unseen splits and our curated benchmark for R2R and REVERIE. The results reveal a strong performance correlation: key metrics such as Success Rate (SR) and Success weighted by Path Length (SPL) on our benchmark closely track the performance on the full splits, with

Table 5: **Ablation Study on Visual Observation Formats.** We compare our proposed 4-image view against standard dense sweeps (36 and 24 images) and a single stitched panorama. The 4-image format yields the highest Success Rate (SR) and Oracle Success Rate (OSR) across both MLLMs. **Bold** indicates the best performance.

Model	Obs. Format	TL ↓	NE ↓	SR ↑	OSR ↑	SPL ↑	nDTW ↑	SDTW ↑	CLS ↑
InternVL3-8B	36 images	12.40	8.20	22.80	44.00	9.10	22.00	11.10	22.50
	24 images	12.10	7.95	24.50	46.80	10.40	23.40	12.00	24.30
	Panoramic (single)	13.20	8.80	20.40	41.50	8.20	20.00	10.10	20.70
	4-image (Ours)	11.74	7.55	28.00	50.50	12.61	25.28	13.38	26.22
Qwen2.5-VL-7B	36 images	9.40	7.60	24.00	39.50	14.80	33.10	17.20	31.80
	24 images	9.05	7.30	25.80	41.70	15.90	34.40	17.90	33.20
	Panoramic (single)	10.20	8.55	21.40	37.90	13.00	30.00	15.10	28.10
	4-image (Ours)	8.54	6.99	27.50	44.00	17.11	35.97	18.88	34.85

Methods	R2R										REVERIE										
	Val Unseen					Subset					Val Unseen					Subset					
	TL	NE	OSR	SR	SPL	TL	NE	OSR	SR	SPL	TL	NE	OSR	SR	SPL	TL	NE	OSR	SR	SPL	
<i>VLN Specialist:</i>																					
VLNBERT (2021)	12.01	3.93	70	63	57	12.06	3.76	70	63	56	13.98	4.18	35	31	25	14.07	4.22	34	31	24	
HAMT (2021)	12.14	3.92	71	63	58	12.10	3.51	73	65	59	14.08	4.12	37	33	30	14.62	4.11	37	32	29	
DUET (2022)	13.94	3.31	81	72	60	13.25	3.54	79	71	60	22.11	3.93	51	47	34	22.28	3.96	49	43	32	
<i>Finetuned MLLM:</i>																					
NavILLM (2023)	12.81	3.38	81	66	54	15.80	3.32	86	66	55	16.04	5.76	54	45	37	19.02	5.80	55	34	27	
NavGPT-2 (2025a)	12.79	3.35	78	70	67	12.39	3.04	82	74	62	-	-	-	-	-	-	-	-	-	-	

Table 6: Performance of baseline agents on the R2R and REVERIE tasks, with results compared across the previous and our benchmark.

deviations typically within a 2 to 3 percentage point margin. This close alignment confirms that our stratified sampling successfully captures the intrinsic difficulty and diversity of the original datasets, establishing our benchmark as a reliable proxy for full-scale evaluation.

When comparing these baselines to our proposed framework, however, a significant performance gap remains. As referenced in Table 6, VLN specialists and finetuned agents achieve substantially higher success rates (e.g., 72% SR on R2R) compared to the zero-shot average. Despite this gap, our zero-shot agents demonstrate non-trivial navigation capabilities, establishing a crucial baseline for generalization without the cost of task-specific training.

D Agent Workflow and Prompt Design

A central component of our framework lies in the design of agent and prompts that guide multimodal large language models (MLLMs) to behave as navigation agents. Since the main paper provides only a brief overview, we expand here with a full account of the workflow, structure, components, and variations used across all eight agent types. Our agents are divided into two families: *Text Summarization as Memory (NavGPT)*, which relies purely on local observations and history, and *Text*

Map as Memory (MapGPT), which augments navigation with dynamically constructed topological maps. Within each family, we instantiate four variants: a baseline version, a chain-of-thought (CoT) agent, a reflection-enabled agent, and a combined CoT+Reflection agent. This section explains the design philosophy of each family and the detailed structure of their prompts.

D.1 Overall Agent Workflow

Although our agents are implemented as executable modules, their behavior is fundamentally driven by prompt design. Conceptually, each agent operates in a closed interaction loop with a self-defined simulator wrapper, where perception, reasoning, and action selection are mediated entirely through structured prompting.

At each navigation step, the agent first interacts with our self-designed simulator via the runner module to retrieve the full set of environmental information and visual observations required for decision making. This includes the panoramic visual input, the navigation instruction, the agent’s current heading and elevation, the navigation history, and the set of navigable action options. For text map memory agents, this additionally includes the dynamically constructed topological map and node connectivity information described in Sec-

tions D.3 and D.5.

These elements are then formatted into a prompt following the unified structure introduced in Section D.2. The system component provides persistent role definitions, navigation rules, and output constraints, while the task component injects step-specific state information. For multimodal models, the prompt is paired with the corresponding visual observation and passed to the MLLM as a single inference call.

Upon receiving the prompt, the MLLM generates a structured textual output that encodes its navigation decision. Depending on the agent variant, this output may include explicit reasoning traces (CoT), reflective self-evaluation, or both, as detailed in Sections D.3–D.7. Crucially, regardless of internal reasoning style, the output always conforms to a strict, predefined format that exposes a single actionable decision.

The agent then parses this output to extract the intended action and feeds it back to the simulator to execute the corresponding movement. The environment transitions to a new state, and the interaction loop repeats until the agent selects the stop action or reaches the maximum step limit. This design ensures that all agent behavior—state interpretation, reasoning, and control—is governed by prompt construction and output parsing, rather than task-specific procedural logic.

The detailed mechanics of prompt composition are described in Sections D.2–D.7, while the rules and implementation of action extraction from model outputs are specified in Section D.8.

D.2 Overall Prompt Structure

All prompts are composed of two distinct parts: the *system* and the *task* component. The system portion defines the global context of the agent, introducing the VLN setting, enumerating the input elements, and stating the rules the model must follow when reasoning about navigation. It also enforces the strict output format required for downstream evaluation. The task portion is dynamic, providing the specific input to the agent at each time step: the instruction, navigation history, agent orientation, and the set of navigable options. Together, these two components establish both the constraints and the situational awareness necessary for coherent decision making. Figure 9 illustrates an example of the full text summarization as memory baseline prompt.

```
[System Prompt]
You are a Vision-Language Navigation (VLN) agent in an indoor environment.
Your task is to select ONE next action that follows the navigation instruction,
based on:
1. Observation - A panoramic image at the current location (4 views: left,
front, right, back) with numbered green-circle markers for possible moves.
2. Instruction - Step-by-step route. Some steps may already be completed;
focus only on the remaining.
3. History - Past moves, including mistakes and loops.
4. Action Options - A dictionary of possible movement choices grouped by
direction. Each option has a unique number as its ID and a text description.
The direction names (Left, Front, Right, Back) are for grouping only — they are
NOT part of the output.

Key Rules:
- Avoid loops: Do not repeat recent viewpoints or oscillate unless required.
- Do not revisit: If a viewpoint was visited multiple times, stop instead.
- Detect arrival: If the scene matches the destination in the instruction,
choose 'Stop'.
- Prefer progress: Select new paths that clearly advance toward the goal.
- Follow heading: Use current heading/elevation when interpreting the
instruction.

Output format:
Action: <Option_ID>. <Option_description>
- <Option_ID> must be ONLY the numeric ID from the action options (e.g.,
'3', '4').
- DO NOT include the direction name in the output.
If stopping:
Action: Stop. Have reached the destination and stop here.

[Navigation Prompt]
Instruction: {instruction}
History: {history}
Current heading: {heading}°, elevation: {elevation}°
Action options: {action_options}

Choose the best next action ID and its description.
Return the result in the exact format specified above.
```

Figure 9: Text summarization memory baseline agent prompt structure

D.3 Text Summarization as Memory Agents

The NavGPT-style agents are designed to operate with information that would be available in a simulator-based VLN setup but translated into our simulator-free representation. The system prompt explicitly instructs the model to select a single next action, referencing only the option identifiers, and to obey a series of rules that reduce common navigation errors such as looping, oscillation, and premature stopping.

The task inputs are carefully structured. The navigation **instruction** is passed in verbatim, ensuring the model has access to the original language guidance. The **history** describes prior movements in natural language, with each step recorded as a turning angle, forward displacement, and the semantic description of the destination viewpoint. This representation provides both spatial reasoning cues and semantic grounding. The agent’s **current heading and elevation** are provided as numerical values, anchoring the model’s interpretation of orientation. Finally, the **action options** are represented as a dictionary keyed by relative directions, where each entry contains a marker ID and a semantic description of the corresponding navigable viewpoint, along with an explicit “Stop” action. This structured but naturalistic representation ensures the model can ground its decisions in both geometry and seman-

tics.

D.4 Reasoning-Enhanced Text Summarization Memory Agent Variants

To probe the role of explicit reasoning, we introduce three reasoning-augmented variants of text summarization memory agent. In the CoT version, the system prompt is modified so that the agent first produces a reasoning trace encapsulated in `<Reasoning>` tags before committing to its final action choice. This design encourages more transparent step-by-step deliberation. The Reflection variant modifies the output format further: after producing an action, the agent generates a reflective evaluation wrapped in `<Reflection>` tags, followed by a `<Final Decision>` statement declaring whether to keep or revise its action. If the reflection deems the decision unsound, the agent replans rather than moving. The CoT+Reflection version combines both mechanisms, first reasoning explicitly and then reflecting on the proposed choice, providing the richest form of introspective navigation. These modifications shift the model from direct action prediction toward a more deliberative, self-monitoring behavior.

```
Instruction: Walk down the stairs. Walk forward and stop next to the door that is next to the recycling bin.

History:
Navigation starts.
Step 1: Turning heading direction -37.78 degrees from right 36.90 to left 0.88, and forward 3.42 meters towards Staircase with glass panels and wall-mounted artwork; leads toward an upper interior area or room.

Current heading: -0.88°, elevation: 27.21°

Action options:
{
  "Left": {
    "3": "Glass-enclosed gym area with exercise equipment; leads toward the fitness room.",
    "4": "Open gym area with visible workout machines; leads toward the main exercise space.",
    "5": "Brightly lit gym area with additional equipment; leads further into the fitness zone."
  },
  "Front": {
    "1": "Hallway with a brown wall and utility area; leads toward a corridor or storage space."
  },
  "Right": {},
  "Back": {
    "2": "Wall with artwork and a staircase; leads toward a lower level or exit."
  },
  "Stop": "Have reach the destination and stop here."
}

Choose the best next action ID and its description.
Return the result in the exact format specified above.
```

Figure 10: Text map memory agent prompt example (Step 1)

D.5 Text Map as Memory Agents

While NavGPT-style focuses on local decision making, the MapGPT-style agents introduces a form of spatial memory through a dynamically constructed topological graph. At each step, the agent augments its prompt with a **map connectivity** field, expressed in natural language, that lists adjacency relationships between viewpoints (e.g., “node_0 is connected to node_1, node_2”). This evolving graph representation enables the MLLM to reason not only about immediate action choices but also about the broader connectivity of the explored environment.

The navigation history for text map memory agents is likewise enriched. Instead of recording only motion trajectories, it includes the current node identifier, a semantic description of the viewpoint, and the sets of visited and unvisited nodes. This structure gives the agent both a local semantic grounding and a global perspective on the exploration state. A complete prompt example after one navigation step is illustrated in Figure 10.

D.6 Reasoning-Enhanced Text Map Memory Agent Variants

The CoT, Reflection, and CoT+Reflection augmentations are applied to text map memory agents in the same manner as for text summarization agents, modifying only the output structure while retaining the additional map input. Thus, the text map memory agents explores how explicit reasoning interacts not just with semantic cues, but also with global topological memory.

D.7 Example CoT Reasoning

Figure 11 presents two representative CoT outputs. In both cases, the model correctly decomposes the navigation instruction into progress estimation, further analysis, and prediction. However, despite being provided with full interaction history, the reasoning predominantly relies on local observations and the most recent instruction, while largely neglecting prior context.

This illustrates a broader issue: even when the input tokens are well within the context window, the model exhibits poor reasoning fidelity across multiple rounds of context. Instead of leveraging accumulated history for richer reasoning, the MLLM tends to perform single-turn grounding of the immediate observation. This behavior highlights a critical limitation for VLN tasks, where successful

```

<Reasoning>:
1. Progress estimate: The instruction is to walk up stairs and wait at the top. The current observation shows a staircase leading upward, which aligns with the instruction.
2. Further analysis: The current heading and elevation suggest we are at the top of the stairs, as there are no more stairs visible ahead. The back view confirms we are at the top, as the glass-walled corridor leads to an exit or open area, not further stairs.
3. Prediction: Since the instruction specifies waiting at the top, and the observation confirms we are at the top, the next logical action is to stop.
</Reasoning>

-----

<Reasoning>:
1. Progress estimate: The instruction requires going up the stairs and turning left at the top. The current observation shows the staircase leading upward, which aligns with the first part of the instruction.
2. Further analysis: The current heading is 93.59°, which suggests we are facing the staircase. The left option indicates a staircase with wall-mounted artwork, which is consistent with the instruction to turn left at the top.
3. Prediction: The next logical step is to move towards the left staircase to continue following the instruction.
</Reasoning>

```

Figure 11: Two examples of chain-of-thought (CoT) reasoning generated by Qwen2.5-VL-7B. Both cases demonstrate structured step-wise reasoning but limited integration of historical context.

navigation often depends on integrating long-term history with dynamic, stepwise decision-making.

D.8 Details of Action Parsing

After the MLLM generates a structured textual response, the agent must convert this free-form output into an executable navigation action. Although the prompting enforces a strict output format, in practice MLLM outputs may contain formatting noise, partial deviations, or additional explanatory text. We therefore adopt a robust, rule-based action parsing strategy that differs slightly across agent variants.

Baseline and CoT Agents. For the baseline and chain-of-thought (CoT) agents, the parsing procedure is intentionally minimal. Since these agents are instructed to output only a single action declaration, the parser searches for an Action field and extracts the associated action value, which may be expressed either as a numeric viewpoint identifier or as a relative direction token (e.g., Left, Front). No additional fields are expected, and any successfully extracted action is immediately validated against the current set of navigable viewpoints before being executed.

Reflection-Enabled Agents. Reflection-based agents require more elaborate parsing, as their outputs may contain multiple structured components, including reflective analysis and decision revision. Specifically, these agents may produce optional

Reflection and Decision (or Final Decision) fields in addition to the final Action.

To accommodate this variability, we employ a two-stage parsing strategy. First, the parser robustly extracts the action value by searching for an Action declaration with flexible tolerance to formatting noise, allowing for extra characters, mark-down symbols, or minor deviations around the action delimiter. The parser prioritizes numeric viewpoint identifiers when present, and falls back to direction-based actions otherwise. To reduce false positives, the search is constrained to a local window following the detected action marker.

Second, if present, the parser extracts the optional Reflection and Decision segments using separate pattern matching rules. These fields are treated as auxiliary metadata: they do not influence control flow directly but are preserved for logging, analysis, and qualitative inspection of the agent’s self-evaluation behavior.

Action Validation and Execution. Regardless of agent type, all parsed actions undergo a strict validation step. Numeric actions are checked against the current viewpoint-to-node mapping and verified to correspond to a navigable location in the simulator. Direction-based actions are resolved into concrete viewpoint identifiers via the provided action option dictionary and validated in the same manner. If an extracted action does not correspond to a valid navigable transition, the agent raises a parsing exception and re-prompts the model with an explicit error signal.

In addition to movement actions, the parser also detects explicit termination signals (i.e., the stop action) and converts them into a terminal agent state. Parsing failures occur only when no valid Action field can be identified; in such cases, the output is treated as malformed and returned to the MLLM for correction.

Overall, regardless of reasoning verbosity, this parsing design ensures that all agent variants produce a single, well-defined navigation action at each step, while remaining robust to minor deviations in model-generated text. Importantly, the parser enforces a clear separation between reasoning content and control decisions, preserving the prompt-centric nature of the agent framework.

VLN Result Visualizer

Agent: baseline, Model: qwen2_5_vl, Task: R2R, Split: val_unseen_subset

Select an episode to visualize the agent's step-by-step navigation process.

Select Episode ID

r2r_5079_0

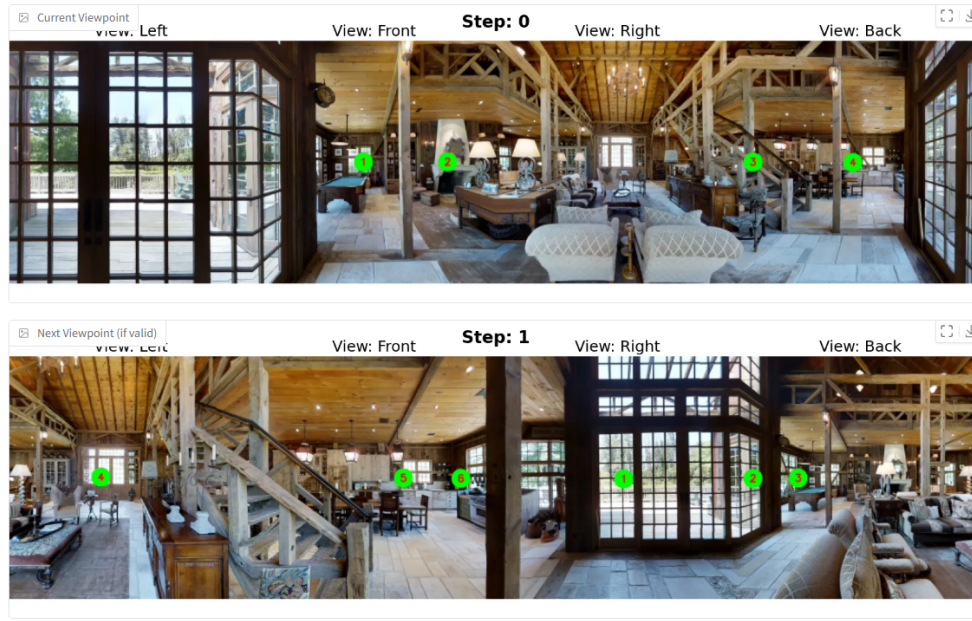
Previous Step

Next Step

 Navigation Status: Navigating...

Current Viewpoint: ['d8d0fa6bfc541889d727767910ea39e']

Next Chosen Number: 3



Step 1

Instruction: Walk past the furniture and to the left of the stairs. Stop by the post at the top of the stairs.

LLM Output:

Action: 3. Dining area with a staircase; leads toward a dining room and upper level access.

Agent Action:

- Turned Angle: 132.67679223642557
- Next Viewpoint ID: 3

Full Trajectory: d8d0fa6bfc541889d727767910ea39e -> 01cada103ecb4ad3864861c13baee57d

GT Path: d8d0fa6bfc541889d727767910ea39e -> 01cada103ecb4ad3864861c13baee57d -> 2b519d8eee9c4abb88444a397e87cd6f -> c69f595ae3e646ee9e447c87de461ab6 -> b9775867c1864f72a54d873c04d87f47

History Log:

- Navigation starts.
- Step 1: Turning heading direction 132.68 degrees from right 95.80 to left 131.52, and forward 2.16 meters towards Dining area with a staircase; leads toward a dining room and upper level access.

Figure 12: The main interface of the VLN Result Visualizer.

E Episode-level Analysis

E.1 Trajectory Visualization and Analysis Tool

This section details the custom tool developed for the qualitative analysis of agent trajectories. The tool, named the VLN Result Visualizer, developed using Gradio, provides an interactive interface for

a step-by-step inspection of any navigation episode, which is crucial for understanding the nuances of agent behavior beyond aggregate metrics.

The visualizer is built entirely using the Gradio framework. Its primary function is to parse the evaluation result files and present the information in a human-readable format. At the top of the interface, a user can specify the configuration

used during evaluation, including the **agent type**, **MLLM model**, **task**, and **data split**. Once a configuration is loaded, a dropdown menu is populated with all episode IDs from that run, allowing for the selection of any specific trajectory for analysis.

The core of the interface, shown in Figure 12, is the visual observation panel. It displays the agent’s panoramic view for the Current Viewpoint and, if a valid move is made, the panoramic view of the chosen Next Viewpoint. Each panoramic image is a composite of four individual images presented in the agent-centric order of [Left, Front, Right, Back], with the global orientation (e.g., “View: Right”) explicitly labeled above each segment. Navigable options are clearly marked with green circular markers.

Below the visual panel, detailed textual information for the current step is provided. This includes: the step number, the original navigation instruction, the raw LLM Output, the parsed agent action (turn angle and forward distance), the full trajectory path taken so far, and the ground truth path. Critically, the tool also flags the exact step at which the agent’s path first deviated from the ground truth, enabling quick identification of crucial mistakes. This is followed by the complete history log that was fed into the model at that step, allowing for an in-depth analysis of the agent’s reasoning context.

To further enhance usability, the interface provides status indicators directly on the display during navigation. The ID of the current viewpoint and the numeric ID of the next chosen marker are displayed at the top of the screen, providing immediate context without needing to consult the text logs.

At the conclusion of each episode, a final summary panel presents the quantitative evaluation metrics, such as Success Rate, SPL, Navigation Error etc., offering a direct link between the agent’s step-by-step actions and its final performance score. This tool was indispensable for conducting the detailed error analysis presented in this paper.

E.2 Detailed Error and Success Analysis

This section provides a detailed analysis of agent performance across 200 navigation trajectories. By breaking down both failures and successes, we can identify the primary challenges MLLM-based agents face in the VLN task.

The results reveal a significant performance gap, with 148 failures compared to only 52 successes. An initial breakdown of the failures, as shown in

Figure 13, indicates that the vast majority (131 out of 148) stem from **incorrect navigation** rather than technical **MLLM Generation Errors** (17 cases). This suggests that while the models are generally capable of producing valid actions, their decision-making logic is the primary point of failure.

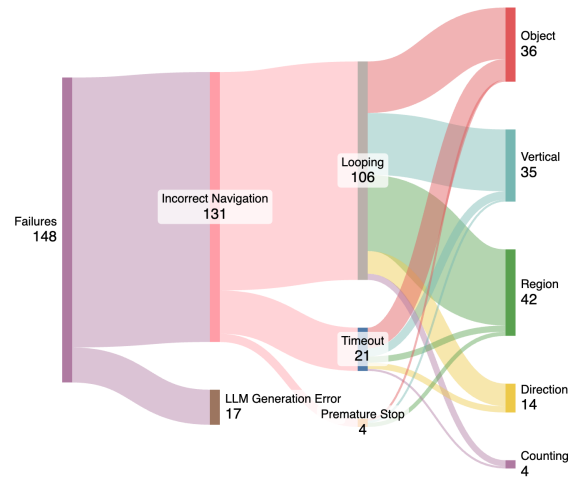


Figure 13: Analysis of error sources in 148 failure episodes.

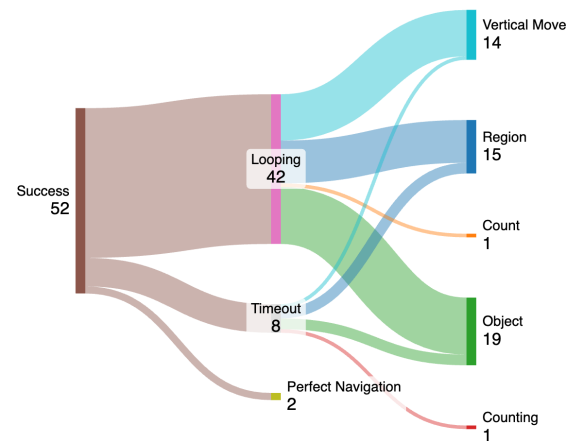


Figure 14: Analysis of navigation behavior in 52 successful episodes.

Within the incorrect navigation errors, **looping** is the most dominant failure mode, accounting for a remarkable 106 cases. This behavior, where the agent repeatedly revisits the same viewpoints, points to a fundamental difficulty in spatial aware-

ness and state tracking. The root causes for these loops, as well as for timeouts, are primarily failures in high-level scene understanding. Specifically, **region recognition** (37 cases in looping), **vertical movement understanding** (30 cases), and **object detection** (25 cases) are the most frequent triggers for getting stuck. This highlights the agent’s struggle to match abstract instructions (e.g., “go to the kitchen”, “go upstairs”) with visual evidence.

Conversely, an analysis of the 52 successful trajectories provides a more nuanced picture of the agent’s capabilities, as illustrated in Figure 14. A striking finding is that only 2 trajectories were completed perfectly. The vast majority of successes (42 cases) were achieved despite the agent exhibiting **looping behavior**, typically near the target. This suggests that while agents can eventually recover from local confusion, their navigation is often highly inefficient. The challenges in these near-success cases mirror those in the failures: difficulties with **object recognition** (16 cases), **vertical movement** (13 cases), and **region understanding** (11 cases) still persist, causing inefficiency even when the final goal is reached.

In conclusion, the data indicates that the primary obstacle for these MLLM agents is not language generation but robust spatial and semantic reasoning. The pervasive issue of looping, both in failed and successful episodes, underscores a weakness in creating and maintaining a stable understanding of the environment. Future work should focus on enhancing these core reasoning capabilities to improve both the reliability and efficiency of navigation.

E.3 Case Studies

In this section, we provide a qualitative analysis of five navigation episodes to illustrate the agent’s common behaviors, highlighting both its capabilities and frequent failure modes.

Figures 15 and 16 showcase successful episodes that also reveal subtle inefficiencies. In Figure 15, the agent correctly identifies the target treadmill but exhibits redundant behavior by moving away and looping back before executing the final stop action. Similarly, the episode in Figure 16 demonstrates a strong recovery capability, yet the agent struggles with precise vertical positioning, causing it to loop on the stairs rather than stopping at the correct step.

Conversely, Figures 17, 18, and 19 depict common failure scenarios. The trajectory in Figure 17 represents a case of ‘oracle success,’ where the

agent navigates to the immediate vicinity of the destination downstairs but ultimately fails by getting trapped in a repetitive loop on the staircase. Figure 18 illustrates a multi-faceted failure; the agent not only fails to ground the directional instruction and identify the ‘hallway’ but also produces a malformed output, resulting in an un-parsable command and an invalid action error. Finally, Figure 19 demonstrates a failure in semantic region understanding, where the agent is unable to correctly interpret the goal of stopping ‘inside of the sauna’.

F The Use of Large Language Models (LLMs)

As disclosed, we utilized LLMs (GPT5, Google Gemini etc.) to aid in polishing the manuscript’s prose. Its role was to improve grammatical correctness and sentence clarity, with all final content being reviewed and approved by the authors, who take full responsibility for this work.



Figure 15: A successful but inefficient trajectory. After observing the target treadmill, the agent loops around the room before stopping.

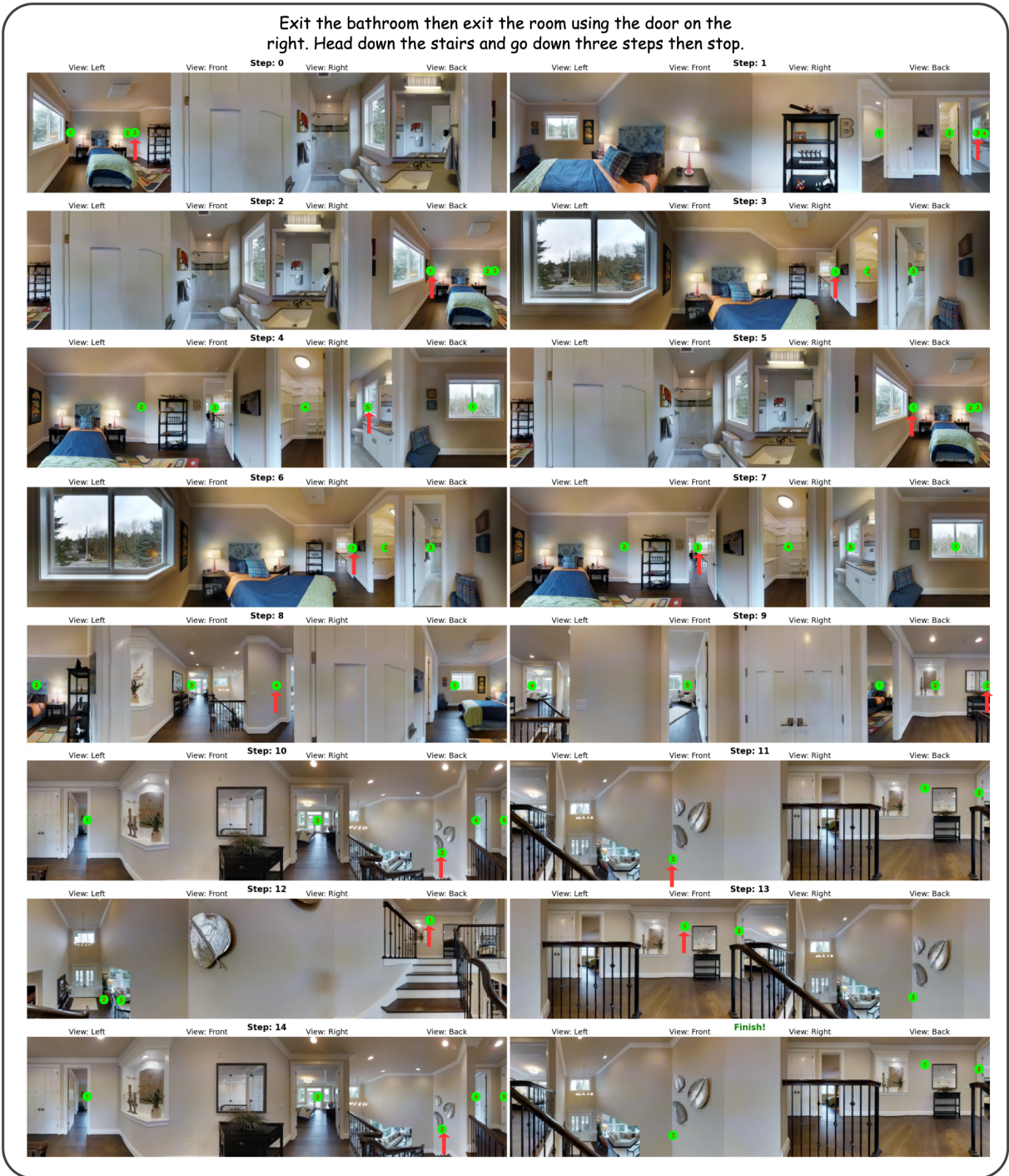


Figure 16: A successful episode showcasing recovery capabilities. However, the agent exhibits looping behavior during vertical movement on the stairs.

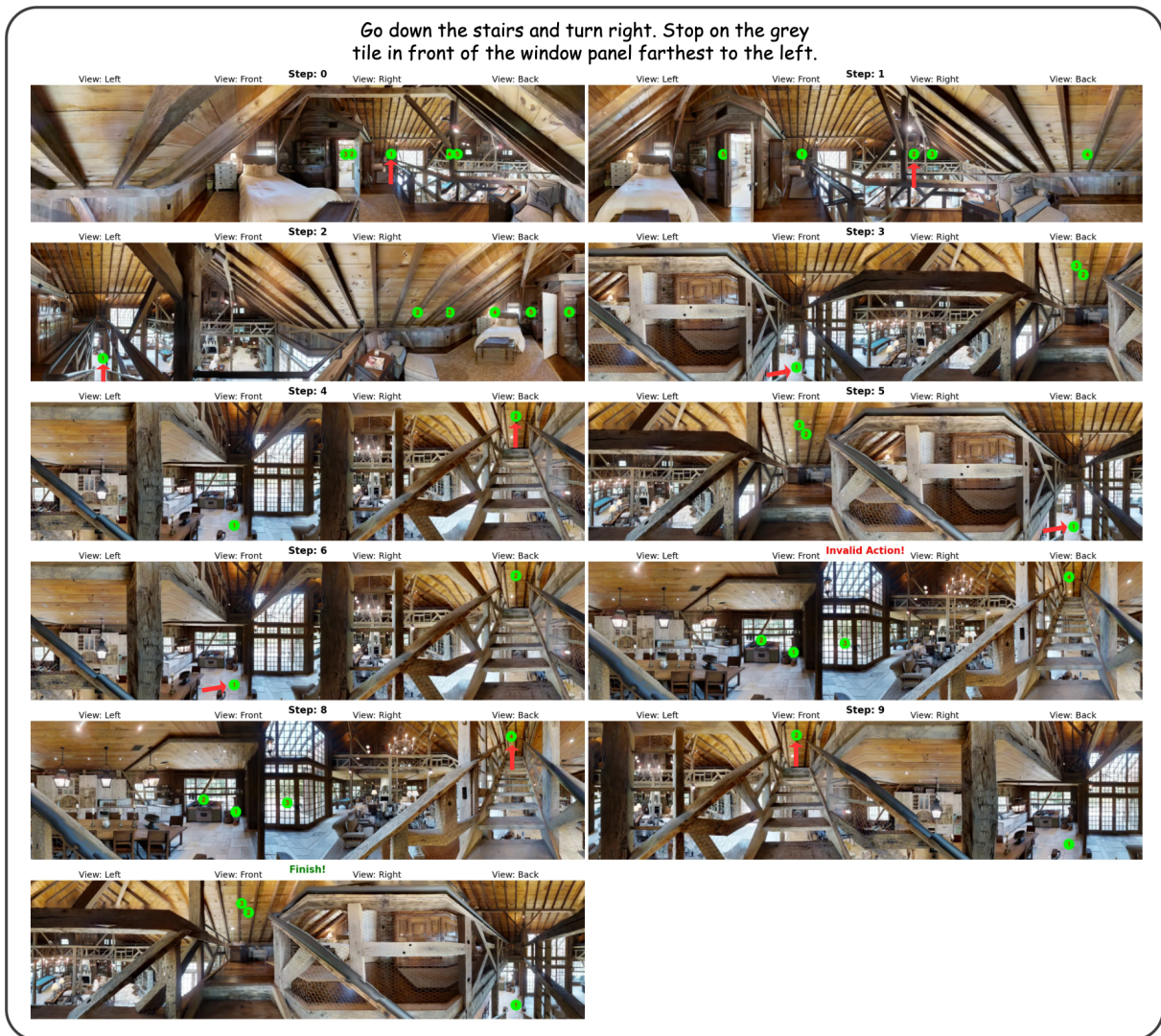


Figure 17: A failure case with oracle success. The agent reaches the correct general area but fails to stop, getting stuck in a loop on the stairs.



Figure 18: Navigation failure due to misinterpreting a directional instruction and a model generation error that produced an invalid action.

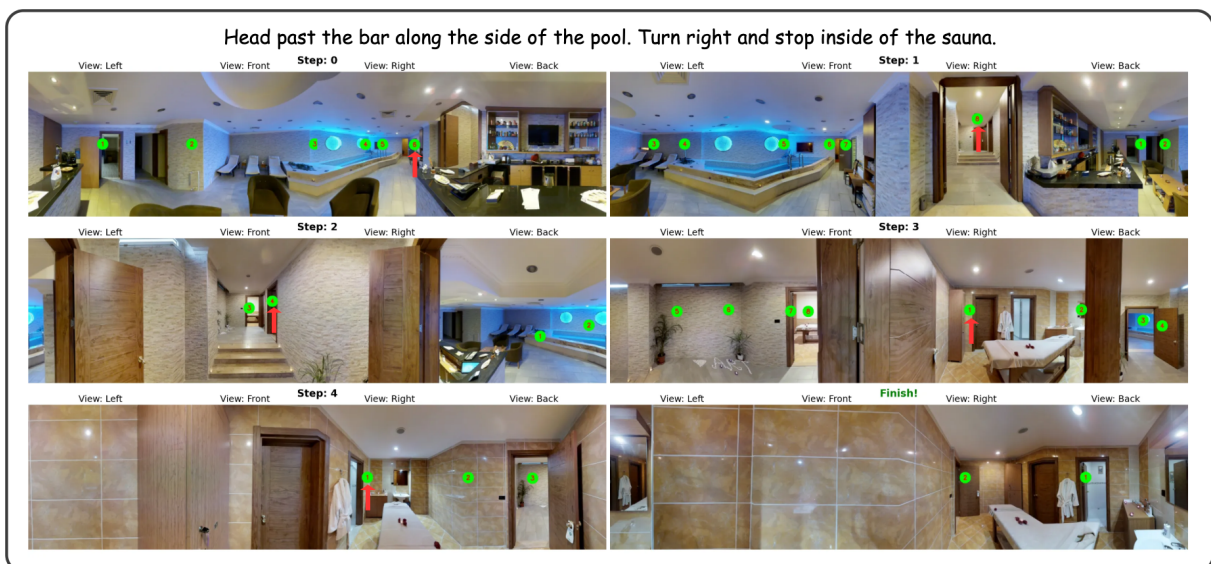


Figure 19: A failure episode caused by the agent's inability to understand and navigate into the specified target region (sauna).