Robotics: Science and Systems 2026
Sydney, Australia, July 13–17, 2026

Beyond Isolation: A Unified Benchmark for General-Purpose Navigation

Shuzheng Sun[1,2,†] Tianyi Yang[1,†] Tengyue Wang[1,†] Yikai Xue[1,†] Zhengjie Xu[1,†]
Lingming Zhang Qichen Zhang[2] Chao Liang[2] Zhipeng Zhang[1,2]✉
[1]AutoLab, SAI, Shanghai Jiao Tong University [2]Research Lab, Anyverse Dynamics

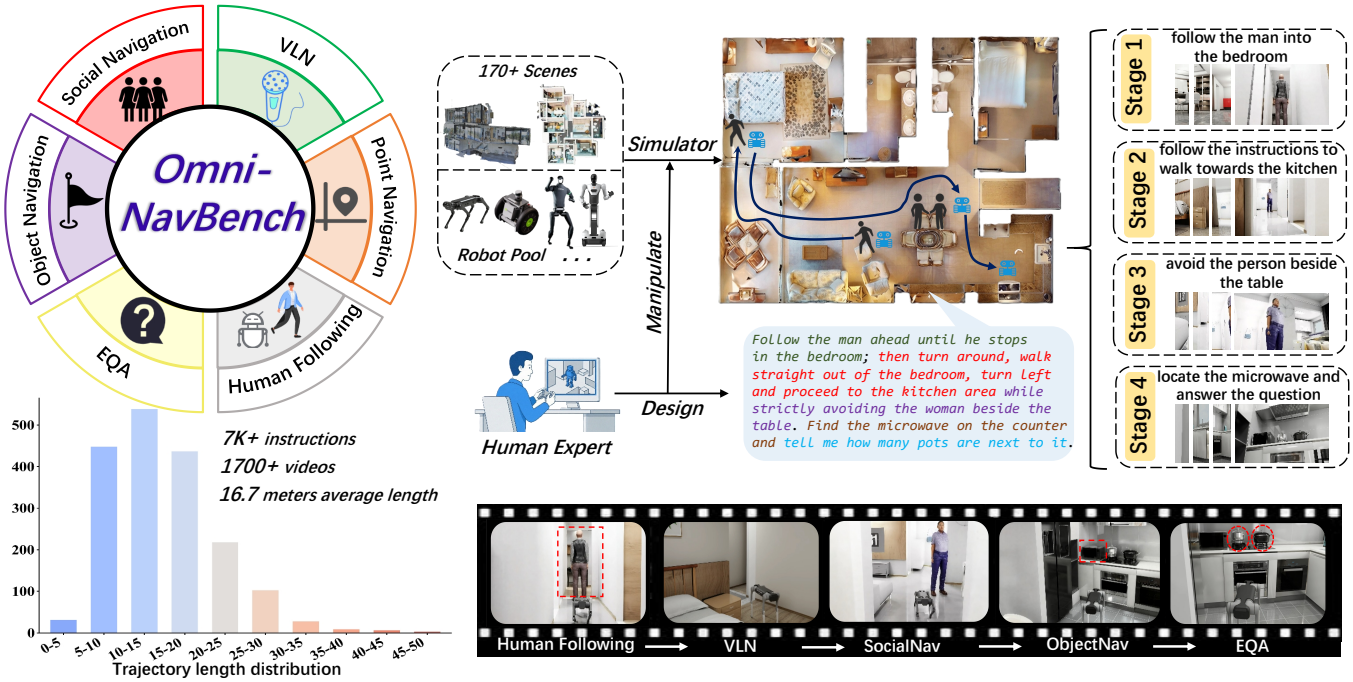

Fig. 1: **Overview of the proposed OmniNavBench.** We introduce a unified benchmark designed to evaluate cross-skill coordination and cross-embodiment generalization in embodied navigation. The benchmark constructs composite instructions by dynamically composing a sequence of primary sub-tasks (*i.e.*, VLN, PointNav, ObjectNav, Human Following), which must be executed while concurrently satisfying overarching constraints for SocialNav and EQA.

***Abstract*—The pursuit of general-purpose embodied agents is currently hindered by fragmented evaluation protocols that isolate navigation skills and fixate on specific robot morphologies. This disconnect fails to reflect real-world scenarios where agents must orchestrate diverse behaviors across varying physical embodiments. To bridge this gap, we introduce OmniNavBench, a holistic benchmark designed to rigorously assess cross-skill coordination and cross-embodiment generalization. Distinguished from existing datasets, OmniNavBench introduces three paradigm shifts: *(1) Compositional Complexity.* We propose composite instructions that interleave sub-tasks from 6 distinct categories (*i.e.*, PointNav, VLN, ObjectNav, SocialNav, Human Following and EQA), compelling agents to seamlessly transition between exploration, interaction, and social compliance within a single unified episode. *(2) Morphological Universality and Sensor Flexibility.* We present a simulation platform that breaks the reliance on single-morphology evaluation. This ecosystem empowers researchers to test generalization across different robot types, including humanoid, quadrupedal, and wheeled, while accommodating diverse algorithmic needs through a modular sensor interface and a hybrid suite of 170 environments blending synthetic assets with real-world scans. *(3) Demonstrations Quality.* Moving beyond mechanical shortest-path algorithms, we curate 1,779 expert trajectories via human teleoperation, capturing critical behavioral nuances, such as exploratory glance and anticipatory avoidance, essential for natural human-robot coexistence. Extensive evaluations demonstrate that current methods, despite their claimed unified design, struggle to adapt to the complex, interleaved nature of truly general-purpose navigation. This exposes a critical disparity between existing capabilities and the demands of real-world deployment, underscoring OmniNavBench as a crucial testbed for the next generation of generalist navigators. Dataset, code, and leaderboard are available at http://omninavbench.cloud-ip.cc.**

I. Introduction

Embodied navigation stands as a cornerstone of embodied AI and aims to endow agents with the autonomy to traverse unfamiliar physical environments through perception and reasoning. While the field began with pioneering tasks like Vision-and-Language Navigation (VLN) [2, 32] and point goal

†Equal contribution. ✉ Corresponding author.

navigation (PointNav) [1, 42], it has subsequently expanded into a diverse array of specialized formulations. These range from PointNav and VLN to more complex behaviors such as object goal navigation (ObjectNav) [6, 10], social navigation (SocialNav) [3, 24], human following [13, 29, 41], and embodied question answering (EQA) [19, 37]. Although recent efforts have sought to develop unified models capable of addressing multiple tasks within a single architecture [39, 40, 43, 44], the evaluation landscape remains fragmented.

This fragmentation arises because different navigation tasks are typically assessed using isolated benchmarks, where VLN relies on R2R [2] and RxR [16] while ObjectNav and EQA utilize HM3D-OVON [36] and OpenEQA [19] respectively. Such a setup creates artificial boundaries between skills by implicitly assuming that agents perform only one type of task at a time. Consequently, models claiming general navigation [39, 40, 43, 44] capabilities must undergo testing across disparate environments that fail to reflect the reality of deployment where agents must seamlessly switch between diverse behaviors like exploratory search and obstacle avoidance. Furthermore, current evaluation protocols often fail to capture the complexity of physical deployment at both the behavioral and embodiment levels. Most datasets rely on trajectory data generated by shortest-path algorithms [15, 30] which miss the nuances of natural human motion under uncertainty such as exploratory behaviors or anticipatory avoidance [8, 17, 23, 36]. This limitation is compounded by evaluations restricted to limited robot morphologies [28], meaning that results derived from a single configuration are insufficient to predict performance across the heterogeneous kinematic properties of different embodiments required for real-world applications.

To bridge the gap between fragmented academic tasks and the holistic demands of real-world operation, we introduce **OmniNavBench**. Particularly, our OmniNavBench makes the following efforts for developing general navigation:

(1) Cross-Skill Coordination Evaluation: We introduce a rigorous protocol to bridge isolated navigation skills, pairing 1,779 composite instructions with expert trajectories to compel seamless transitions between capabilities. This benchmark is augmented to 7,116 instructions with diverse linguistic variations to test robustness against natural language ambiguity. Each instruction weaves together a sequence of at least two primary tasks (*i.e.*, VLN, ObjectNav, PointNav, and Human-Follow), which must be executed while concurrently satisfying overarching constraints for SocialNav and EQA. This complex formulation ensures agents are evaluated on their ability to maintain spatiotemporal continuity while switching strategies, rather than on performance in isolated subtasks.

(2) Flexible and Unified Evaluation Platform: We present a high-fidelity simulation platform built upon NVIDIA Isaac Sim [21] that breaks the reliance on single-morphology evaluation by enabling generalization tests across humanoid, quadrupedal, and wheeled robots. To accommodate diverse algorithmic needs, the system features a highly modular interface for customizing sensor count, position, and orientation (*e.g.*, RGB-D cameras, LiDAR, panoramic vision systems). This is supported by a hybrid suite of 170 environments, blending 85 high-fidelity synthetic assets from GRScenes [27] with 85 photorealistic real-world scans from Matterport3D [5].

(3) Human Expert Demonstrations: We move beyond the limitations of mechanical shortest-path algorithms by curating a comprehensive dataset of 1,779 expert trajectories collected through human teleoperation. These demonstrations, which span an average length of 16.7 meters and a cumulative distance of 29.5 kilometers, encapsulate the subtle decision-making patterns and hesitation behaviors that characterize natural human navigation. The dataset includes rich egocentric RGB-D video sequences totaling 24 hours and 2.6M frames. By providing these high-quality supervision signals, we offer a resource that enables models to learn human-intuitive strategies for exploration and obstacle avoidance rather than merely optimizing geometric efficiency.

To probe the limits of existing unified navigation models, we conducted a rigorous evaluation on OmniNavBench. Our analysis indicates a substantial performance degradation when models confront complex, multi-stage tasks. Although they can often make progress on individual sub-tasks, they frequently fail to complete the full composite instruction. This suggests the principal challenge lies in sequential task coordination rather than isolated skill deficiency. This weakness extends to dynamic scenarios, where models display inadequate social awareness and struggle to maintain appropriate interaction with moving humans. Furthermore, we observe significant generalization issues, where performance is inconsistent across different robot morphologies, deteriorates in photorealistic scans compared to synthetic environments, and is highly sensitive to variations in instruction style.

In summary, our main contributions are: ♠ We introduce OmniNavBench, a unified benchmark for evaluating general-purpose navigation. ♠ We build a scalable simulation platform that supports heterogeneous robot embodiments, customizable sensor configurations, and multiple action interfaces, enabling standardized evaluation across embodiment and control settings. ♠ We conduct systematic evaluations and in-depth analyses of state-of-the-art models with unified navigation potential under a unified protocol, providing a strong reference for future unified navigation research.

II. Related Works

A. Embodied Navigation Benchmarks

Early embodied navigation research primarily treated tasks as isolated problems, which created a fragmented evaluation landscape. For instance, standard benchmarks like R2R [2] and RxR [16] focus on VLN, while HM3D-OVON [36] targets ObjectNav. HA-VLN [8] introduces dynamic human activities, which is still confined to the VLN domain. This separation prevents us from verifying whether an agent can coordinate multiple skills within a unified task flow. To address these limitations, recent works have introduced more complex settings. LHPR-VLN [25] extends task horizons with multi-step goals, but these are largely repetitions of homogeneous subtasks that do not require strategy switching. Other benchmarks

like GOAT-Bench [14] involve navigating to goals specified by different modalities. Several similar works attempted to unify multiple navigation tasks, including OctoNav [11] and VLNVerse [18]. However, their scope is limited as they operate primarily in static settings, failing to incorporate dynamic entities like humans. This prevents them from testing an agent's critical ability to switch strategies in response to real-time events. Furthermore, this issue is compounded by their reliance on trajectory data from shortest-path algorithms. Such simplistic paths do not reflect the complex, nuanced movements of human experts, which is a key feature captured by the manually collected data in our benchmark.

B. Trajectory Generation and Human Demonstrations

Most existing navigation benchmarks rely heavily on map-based shortest-path algorithms to generate trajectories [8, 11, 25]. These "oracle-view" paths exhibit idealized behavior [2, 15] and lack the natural characteristics humans display when facing uncertainty, such as exploration, anticipatory avoidance, and active scanning [8, 17, 23, 36]. This distributional gap prevents models from learning human-intuitive strategies. To address this, some works have explored incorporating human demonstration data. For instance, RxR [16] provides human follower trajectories with temporally aligned language grounding. Similarly, analysis in CityNav [17] reveals that human paths differ significantly from shortest paths, particularly in landmark utilization (36.3% vs. 24.6%). The necessity of human data is further supported by research in robot manipulation, such as Robomimic [20] and RH20T [9]. These studies demonstrate that even limited high-quality human demonstrations can capture behavioral diversity that algorithmically generated data cannot replicate. Despite these findings, this paradigm has not yet been adopted in multi-task navigation benchmarks. Current works either persist with shortest-path generation or limit human data collection to single task types [3], leaving a lack of expert demonstrations that cover complex task flows.

C. Unified Navigation Models

At the algorithmic level, significant progress has been made toward building unified navigation capabilities. Several works have successfully unified multiple tasks into a single framework. For example, NaviLLM [43] and SAME [44] formulate navigation as language generation or modular decision problems. Similarly, UniGoal [35] and OmniNav [33] manage multi-modal goals and diverse task types within one system, while Uni-NaVid [39] further integrates VLN, ObjectNav, EQA, and HumanFollow into a single architecture. These efforts demonstrate that current model architectures possess strong potential for cross-task generalization. In parallel, there is a growing trend toward developing generalist algorithms that adapt to different robot morphologies. NavFoM [40] has been deployed across diverse platforms, including quadrupeds, drones, and wheeled robots. InternVLA-N1 [31] showcases zero-shot cross-embodiment generalization, spanning wheeled to humanoid robots, and NaVILA [7] achieves transfer for legged robots by decoupling high-level planning from low-level control. Furthermore, VLN-PE [28] highlights the importance of physical embodiment, revealing how sensitive existing algorithms are to robot dynamics and camera configurations. Collectively, these advances indicate that unified navigation models have demonstrated promising capabilities in task generalization and continuous control. However, a critical gap remains. These sophisticated models are still evaluated on fragmented, single-task benchmarks. There is currently no unified platform capable of systematically verifying their ability to perform dynamic strategy switching across skills and adapt to different embodiments within a cohesive task flow.

III. OMNINAVBENCH

The goal of OmniNavBench is to offer a unified platform for evaluating cross-skill navigation capabilities across heterogeneous robot embodiments, utilizing high-quality human demonstrations as behavioral references. To achieve this, we adhere to the following four core design principles.

♠ **Compositional Task Structure.** To evaluate the coordination of multiple navigation skills, each instruction integrates at least two heterogeneous sub-tasks, which combine static goal reaching, dynamic target tracking, and semantic reasoning within a single episode. This design imposes spatio-temporal continuity constraints, necessitating adaptive strategy switching as task demands evolve.

♠ **Cross-Embodiment Coverage.** Addressing the diversity of real-world platforms, we ensure the same instruction set is executable across three morphologically distinct agents, including wheeled, quadrupedal, and humanoid robots. This design enables a direct assessment of algorithmic robustness against variations in kinematics, viewpoints, and dynamics.

♠ **Naturalistic Behavior Supervision.** Prioritizing behavioral realism over geometric optimality, all reference trajectories are collected via human teleoperation. Unlike shortest-path algorithms (*e.g.*, A*), this approach preserves human-specific navigation patterns, such as exploratory behaviors and anticipatory adjustments [8, 17, 23, 36].

♠ **Controlled Difficulty Levels.** To facilitate systematic evaluation, we establish a complexity curriculum by modulating the sequence length and task heterogeneity. The benchmark offers a difficulty gradient, ranging from two-task compositions to long-horizon four-or-more-task sequences, which allows for fine-grained analysis of agent capabilities.

A. Task Specification

We formalize the navigation mission as a composite instruction comprising an ordered sequence of sub-tasks defined by

$$\mathcal{M} = \{(l_1, \tau_1, g_1), (l_2, \tau_2, g_2), \ldots, (l_n, \tau_n, g_n)\} \quad (1)$$

where l_i represents the natural language instruction segment, τ_i specifies the sub-task type, and g_i defines the target goal. Our benchmark incorporates the following primitives to construct these sequences.

♠ **Sequential Sub-tasks.** We integrate four distinct navigation primitives to evaluate a full spectrum of embodied capabilities.

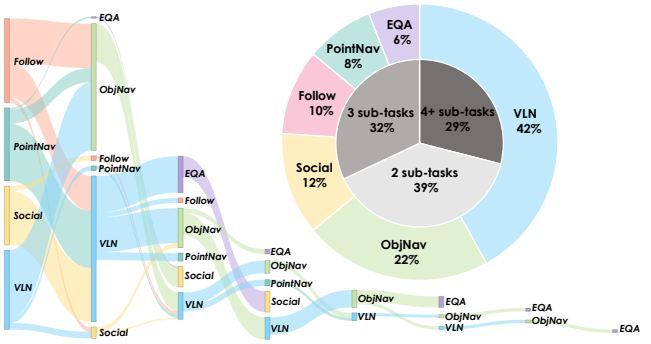

Fig. 2: **Statistics of Sub-instructions.** Sankey diagram depicting cross-skill compositional patterns (left), and distribution of sub-instruction types with task complexity breakdown (right).

- **VLN** requires the agent to interpret natural language instructions grounded in visual landmarks, such as *"turning right after passing a sofa"*.
- **ObjectNav** compels the agent to conduct semantic exploration in unknown environments to locate instances of specific categories like *"find a trash can"*.
- **PointNav** tests precise geometric understanding by requiring movement to exact metric coordinates (x, y) relative to the starting pose.
- **HumanFollow** necessitates active visual tracking where the agent must maintain continuous line-of-sight with a moving human subject.

Notably, **Return** is a VLN sub-task that mandates navigating back to the starting position upon task completion, requiring the agent to leverage long-term spatial memory.

♠ **Parallel Sub-task.** We designate **SocialNav** and **EQA** as a concurrent operational constraint capable of being superimposed onto any of the aforementioned sequential sub-tasks.

- **SocialNav** requires the agent to minimize intrusion into human activity spaces during navigation.
- **EQA** fuses navigation with cognitive reasoning where the agent must gather visual evidence to answer questions regarding environmental attributes.

B. Data Acquisition

We adhere to a strict human teleoperation protocol to capture the nuances of biological navigation rather than relying on algorithmic planners. While shortest-path algorithms prioritize geometric optimality, human controllers exhibit organic behaviors like hesitation and exploratory glances [8, 17, 23, 36]. These subtle cues are essential for learning robust policies, particularly for dynamic tasks like HumanFollow where rule-based generators fail to produce convincing social compliance.

♠ **Rigorous Collection Pipeline.** We implemented a multi-stage workflow involving trained annotators to ensure high-fidelity data. Operators manually control the robot to execute composite instructions within the simulator while the system records pose and sensor data as waypoints. To guarantee semantic alignment, experts subsequently verify the trajectories by reviewing the egocentric video feeds. We mandate that any trajectory failing to capture specified landmarks or objects clearly must be revised by the original operator to maintain behavioral consistency throughout the iteration loop.

Fig. 3: **Cross-embodiment navigation across diverse environments.** Each row shows a different robot morphology (Carter, Aliengo, H1); each column shows a different environment source (synthetic, Matterport3D). Insets display egocentric RGB observations, highlighting viewpoint variations across embodiments.

♠ **Morphology-Agnostic Representation.** The final dataset comprises waypoint sequences (x, y, θ, t) paired with synchronized RGB-D streams. We deliberately exclude low-level action primitives because the high-fidelity physics in Isaac Sim introduces inertia and friction that necessitate continuous micro-corrections. These task-irrelevant adjustments constitute noise for high-level policy learning whereas our waypoint representation abstracts away these dynamics to facilitate transfer across different robot controllers.

C. Dataset Statistics

OmniNavBench contains 1,779 expert demonstration trajectories spanning 170 distinct scenes accompanied by 7,116 corresponding instructions, with an average trajectory length of 16.7 meters and duration of 48.9 seconds. The scenes are drawn from three sources with different visual characteristics: home environments account for 828 trajectories (46.5%), covering diverse residential layouts and furniture arrangements; commercial spaces contribute 213 trajectories (12.0%), including offices and retail environments; and Matterport3D scans provide 738 trajectories (41.4%) with photorealistic reconstructions of real-world buildings. This diversity in visual appearance and spatial structure enables evaluation of cross-domain generalization.

♠ **Task and Linguistic Complexity.** Fig. 2 illustrates the comprehensive statistics of our dataset. The task composition (outer ring) demonstrates a balanced distribution covering both static and dynamic objectives. Specifically, VLN accounts for the largest share at 42%, followed by ObjectNav at 22%. To ensure diversity in interaction, we incorporate significant proportions of SocialNav (12%), Follow (10%), PointNav (8%), and EQA (6%). Notably, we allocated 28.6% of the VLN category to Return episodes, a deliberate design choice to rigorously evaluate the agent's long-term spatial memory and path integration capabilities as previously discussed. The complexity distribution (inner ring) reveals a deliberate gradient of difficulty designed for systematic evaluation. Specifically, 39% of trajectories contain 2 sub-tasks, 32% extend to 3

TABLE I: **Comparisons between OmniNavBench and existing benchmarks.** OmniNavBench supports the most comprehensive set of tasks and utilize human-expert trajectories.

Benchmark	Task Capability							Cross	Trajectory	Avg. Instruction
	ObjNav	SocialNav	VLN	Follow	PointNav	EQA	Unified	Embodiment	Source	Length
R2R [2]	✗	✗	✓	✗	✗	✗	✗	✗	Algorithm	29
RxR [16]	✗	✗	✓	✗	✗	✗	✗	✗	Human Expert	-
VLN-CE [15]	✗	✗	✓	✗	✗	✗	✗	✗	Algorithm	30
OpenEQA [19]	✗	✗	✗	✗	✗	✓	✗	✗	-	-
HA-VLN [8]	✗	✓	✓	✗	✗	✗	✗	✗	Algorithm	-
GOAT-Bench [14]	✓	✗	✗	✗	✗	✗	✗	✗	Algorithm	-
REVERIE [22]	✗	✗	✓	✗	✗	✗	✗	✗	Algorithm	18
OVON [36]	✓	✗	✗	✗	✗	✗	✗	✗	Algorithm	-
LHPR-VLN [25]	✓	✗	✓	✗	✗	✗	✓	✓	Algorithm	18.17
VLN-PE [28]	✗	✗	✓	✗	✗	✗	✗	✓	Algorithm	-
VLNVerse [18]	✓	✗	✓	✗	✗	✗	✓	✓	Algorithm	-
OctoNav-Bench [11]	✓	✗	✓	✗	✓	✗	✓	✗	Algorithm	-
OmniNavBench (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	**Human Expert**	42

sub-tasks, and a substantial 29% involve sequences of 4 or more sub-tasks.The instructions exhibit rich linguistic diversity, encompassing spatial references (*e.g.*, bedroom, kitchen), target objects (*e.g.*, chair, sofa), and action verbs (*e.g.*, forward, follow), with lengths spanning from 9 to over 130 words (see Appendix Fig. 8).

♠ **Instruction and Linguistic Diversity** Our benchmark further provides three supplementary instruction variants which are *Concise*, *First Person*, and *Verbose* to enhance linguistic diversity, yielding a total of 7,116 stylized instructions. These variants were generated using Qwen3-Max [34] to produce semantically consistent paraphrases of the original instructions. *Concise* instructions retain only essential navigation intents by removing redundant descriptions. *First Person* instructions reframe the original into a first-person voice. *Verbose* instructions incorporate additional contextual elaborations while preserving the target and action sequence. Details about stylized instructions are shown in the appendix due to space limitations.

♠ **Morphology and Evaluation Protocols.** The benchmark encompasses three robot morphologies including wheeled (NVIDIA Carter V1 and Kitt15), quadrupedal, and bipedal models where each category accounts for approximately one-third of the trajectories. Notably, the quadrupedal subset specifically includes stair-climbing sequences that are largely absent from the other two groups. To facilitate rigorous assessment, the trajectory split partitions the dataset by scene and contains 1,187 training trajectories across 112 scenes along with 592 test trajectories from 58 scenes to evaluate generalization to novel environments.

D. Evaluation Platform

We established an integrated simulation platform based on NVIDIA Isaac Sim to facilitate unified navigation evaluation and cross-embodiment algorithm development. Distinct from some modern simulators (*e.g.*, Matterport3D [4], Habitat [26]) that employ discrete state transitions or teleportation-based navigation, our platform performs continuous physical stepping with full dynamics simulation. We adopt the RL-based locomotion controllers from VLN-PE [28] to ensure physically grounded robot dynamics. This design provides high-fidelity physics and photorealistic rendering which ensures that evaluated agents must handle realistic motion dynamics, sensor noise, and visual complexity that closely approximate real-world deployment conditions.

Our platform integrates three morphologically distinct robots including the wheeled NVIDIA Carter V1, the quadrupedal Unitree Aliengo, and the humanoid Unitree H1 within a unified framework. Fig. 3 visualizes the three embodiments navigating in both synthetic and photorealistic environments, highlighting the viewpoint variations arising from different morphologies. Despite fundamental differences in kinematic constraints and viewpoint heights, the platform achieves unified orchestration through abstracted control interfaces. Algorithms can output waypoint sequences for low-level path tracking or directly issue velocity commands (v, ω) for end-to-end control. Furthermore, the system supports discrete actions such as step forward or rotate left and right that are mapped to physically-simulated movements. These actions reach the same target poses as teleportation-based simulators while preserving realistic dynamics, which enables the same algorithm to be evaluated across different morphologies without modification. Our platform also facilitates modular sensor configuration where users can specify sensor types including RGB, depth, panoramic camera, LiDAR, and IMU along with their quantities and mounting poses via configuration files.

To enable dynamic scene interaction, the platform leverages the built-in human character system of Isaac Sim which offers diverse pedestrian models with various appearances. Pedestrians follow predefined trajectories while maintaining the capability for autonomous obstacle avoidance to create realistic social navigation scenarios for SocialNav and HumanFollow tasks. The platform supports both synthetic environments with controllable lighting and real-world Matterport scans for cross-domain generalization evaluation. To ensure scientific rigor, scene initial states, pedestrian trajectories, and random seeds can be precisely reproduced via configuration files which guarantees verifiable and reproducible experimental results.

TABLE II: **Sub-task Definitions and Completion Rules in OmniNavBench.** A composite instruction consists of sub-tasks $\mathcal{M} = \{(l_i, \tau_i, g_i)\}$. For each τ_i, we define its input l_i, evaluator goal g_i, and the sub-task completion rule.

Task Type (τ_i)	Instruction Input (l_i)	Goal (g_i)	Completion Rule
State-Goal Navigation (Agent must reach a specific terminal state)			
VLN	Language with Landmarks & Rooms	Landmarks set $\mathcal{L}_{goal}$ and Region $\mathcal{R}_{goal}$	$\exists t:\ \big(\min_{\ell\in\mathcal{L}_{goal}} d_{geo}(x_t,\ell) \le 3.0\,\text{m}\big)\ \wedge\ \big(x_t \in \mathcal{R}_{goal}\big)$
PointNav	Relative Displacement Δp	Evaluator goal coordinate p_{goal}	$\exists t:\ d(x_t, p_{goal}) \le 0.36\,\text{m}$
ObjectNav	Open-vocabulary Query (text)	Target object set $\mathcal{O}$(query)	$\exists t:\ \min_{o\in\mathcal{O}(\text{query})} d_{geo}(x_t, o) \le 1.0\,\text{m}$
Process-Maintenance (Agent must maintain a state over time)			
HumanFollow	"Follow [Description]"	Moving person H_{id}	$\text{Cond}_{t_e} = \text{True}$, where $\text{Cond}_t := (1.0 < d_t \le 3.0) \wedge (\angle_{err,t} \le 60°)$
Cognitive Reasoning (Agent must generate an answer)			
EQA	Visual Question Q	Ground Truth Answer A	$\text{SubstringMatch}(A_{\text{pred}}, A) = \text{True}$
Global Constraint (Applies to all tasks)			
SocialNav	(Implicit)	All avoided humans $\mathcal{H}$	Constraint: $d(x_t, \mathcal{H}) \ge 1.2\,\text{m}$ throughout the episode

Notation. x_t: robot center position; $d_{geo}(\cdot,\cdot)$: geodesic distance computed via NavMesh; $d(\cdot,\cdot)$: 2D Euclidean distance; $\angle_{err,t}$: angle between robot heading and robot-to-human vector; H_{id}: identity of the person to follow; t_e: timestep when the person stops moving; For State-Goal tasks, completion is determined at the **first** timestep satisfying the predicate. Episode-level metrics (SR, SPL, *etc.*) are defined in Tab. III.

TABLE III: **Evaluation Metrics in OmniNavBench.** All reported metrics represent the average over test episodes (using the relevant subset N_{task} for task-specific metrics). **SR**, **SPL**, and **OSR** are standard navigation metrics. **SGC** measures continuous progress. **HFS**, **HFR**, **ACC**, and **SII** evaluate human-robot interaction and safety.

Dimension	Metric	Sym.	Definition / Formula (Averaged over relevant episodes)	Significance
Completion	Success Rate	SR	$\frac{1}{N}\sum_{i=1}^{N} S_i(\delta_g)$	Fraction of fully successful episodes.
	Completion Success Rate	CSR	$\frac{1}{N}\sum_{i=1}^{N} \frac{K_{\text{completed},i}}{K_{\text{total},i}}$	Avg. ratio of strictly completed sub-instructions.
	Sub-Goal Completion	SGC	$\frac{1}{N}\sum_{i=1}^{N} \frac{1}{K_i}\sum_{j=1}^{K_i} \mathcal{P}(\tau_{i,j})$	Average continuous progress across sub-tasks.
	Oracle Success Rate	OSR	$\frac{1}{N}\sum_{i=1}^{N} OS_i(\delta_g)$	Success if goal was *ever* reached.
Efficiency	Success weighted by Path Length	SPL	$\frac{1}{N}\sum_{i=1}^{N} S_i \frac{\ell_{shortest,i}}{\max(\ell_{actual,i}, \ell_{shortest,i})}$	Efficiency weighted by success.
Interaction	Human Follow Success	HFS	$\frac{1}{N_{\text{follow}}}\sum_{i=1}^{N_{\text{follow}}} \mathbb{I}(\text{pos}_{\text{end},i} \in \text{Valid}_i)$	Binary success at the final step.
	Human Follow Ratio	HFR	$\frac{1}{N_{\text{follow}}}\sum_{i=1}^{N_{\text{follow}}} \frac{T_{\text{valid},i}}{T_{\text{follow},i}},\ T_{\text{valid},i} = \sum_{t=1}^{T_{\text{follow},i}} \mathbb{I}(\text{Cond}_{i,t})$	Coverage ratio during following phase.
	EQA Accuracy	EQA-ACC	$\frac{1}{N_{\text{eqa}}}\sum_{i=1}^{N_{\text{eqa}}} \mathbb{I}(a_{\text{pred},i} \cong a_{\text{gt},i})$	Accuracy of numeric-normalized substring match.
Safety	Social Intrusion Index	SII	$\frac{1}{N_{\text{social}}}\sum_{i=1}^{N_{\text{social}}} \big(\frac{1}{T_i}\sum_{t=1}^{T_i} \mathbb{I}(d_t < 1.2\,\text{m})\big)$	Ratio of steps violating personal space.

Notation. N: total test episodes ($N_{\text{follow}}, N_{\text{social}}, N_{\text{eqa}}$: task-specific subsets); $\mathbb{I}(\cdot)$: indicator function; $K_{\text{total},i}$: total number of sub-instructions in episode i; K_i: number of sub-tasks for progress calculation; $\tau_{i,j}$: the j-th sub-task of episode i; $\mathcal{P}(\cdot)$: continuous progress function (see Appendix); S_i: binary success ($S_i = 1$ if agent issues `Stop` within δ_g of the final goal); OS_i: oracle success ($OS_i = 1$ if agent ever reaches within δ_g of the final goal); δ_g: goal distance threshold (2.0 m); ℓ: path length; $T_{\text{follow},i}$: duration of the following phase; $a_{\text{gt}}, a_{\text{pred}}$: ground truth and predicted answers ($\cong$ denotes a success condition via numeric-normalized substring matching); d_t: distance between robot and human, SII threshold of 1.2 m corresponds to Hall's personal distance zone [12].

E. Comparison to Other Benchmarks

As detailed in Tab. I, OmniNavBench significantly advances upon previous benchmarks. While recent works like OctoNav-Bench [11] and VLNVerse [18] move towards task unification, they are largely confined to static environments and neglect dynamic interaction. In contrast, OmniNavBench uniquely integrates six tasks spanning the full spectrum from static navigation (*e.g.*, PointNav) to dynamic human-agent interaction (*e.g.*, SocialNav). Most critically, OmniNavBench distinguishes itself through its behavioral realism. Unlike virtually competing unified benchmarks (*e.g.*, LHPR-VLN [25], VLN-Verse [18], OctoNav-Bench [11]) that rely on geometrically optimal paths from algorithmic planners, it is built entirely on human-expert teleoperated data. This foundation allows it to capture naturalistic nuances like exploratory behaviors and social compliance, which are absent in algorithmic trajectories [8, 17, 23, 36]. Furthermore, by incorporating diverse robot morphologies and the highest linguistic complexity (avg. 42 words), OmniNavBench provides a more faithful approximation of real-world challenges.

IV. EXPERIMENTS

Successful completion of a composite instruction requires valid execution of all sequential sub-tasks and fulfillment of end-of-task criteria. Task definitions and completion rules are detailed in Tab. II, with episode-level metrics in Tab. III. We provide more details about the metrics definition in the appendix due to space limitation.

A. Evaluating Unified Navigation Models

We evaluate four recent popular unified navigation models representing different unification mechanisms. PoliFormer [38] relies on scaled transformer-based reinforcement learning to induce general navigation behaviors across tasks while Uni-NaVid [39] formulates diverse navigation tasks as a single vision–language–action generation problem from egocentric video. MTU3D [45] unifies navigation and exploration by explicitly optimizing movement for spatial understanding and OmniNav [33] adopts a shared continuous waypoint abstraction to support multiple navigation task formulations. All models are evaluated zero-shot with publicly released weights,

TABLE IV: **Quantitative comparison with state-of-the-art baselines across different robot embodiments.** All metrics are reported in %. **CSR** (Completion Success Rate) and **SGC** (Sub-Goal Completion) indicate partial progress. **SII** (Social Intrusion Index) is a safety metric where **lower is better** (↓). **EQA-ACC** denotes the Embodied Question Answering accuracy. Best results are bolded.

Method	Embodiment	SR ↑	CSR ↑	SGC ↑	OSR ↑	SPL ↑	HFS ↑	HFR ↑	SII ↓	EQA-ACC ↑
MTU3D	NVIDIA Carter	0.00	1.08	29.65	5.38	0.00	20.00	24.66	4.46	–
	Unitree Aliengo	0.00	0.98	30.95	11.76	0.00	11.11	16.91	5.65	–
	Unitree H1	0.00	2.08	27.42	8.33	0.00	22.22	27.83	6.74	–
Poliformer	NVIDIA Carter	0.98	0.98	29.12	**10.78**	0.73	14.81	14.70	**2.72**	–
	Unitree Aliengo	1.00	0.00	29.41	16.00	0.29	3.70	9.68	**2.94**	–
	Unitree H1	**1.08**	0.00	25.27	3.23	0.31	8.00	9.32	4.14	–
Uni-NaVid	NVIDIA Carter	1.94	**2.91**	**38.93**	9.71	1.45	**45.16**	**43.67**	9.39	0.00
	Unitree Aliengo	1.96	**2.94**	**44.54**	**26.47**	1.56	**23.33**	**32.75**	18.46	6.25
	Unitree H1	1.06	**2.13**	**29.02**	**8.51**	**1.06**	**24.14**	**25.99**	10.09	7.69
OmniNav	NVIDIA Carter	**1.96**	0.00	32.34	8.82	**1.71**	33.33	31.04	5.23	–
	Unitree Aliengo	**8.74**	1.94	35.30	15.53	**7.01**	14.29	17.49	3.40	–
	Unitree H1	0.00	0.00	24.82	2.13	0.00	14.81	13.93	**3.31**	–

adapting only camera count, resolution, and FOV to each algorithm's specifications.

Findings1: Universal Performance Degradation. Tab. IV presents the quantitative evaluation of four unified navigation models across three robot embodiments. Despite their claimed multi-task capabilities demonstrated on isolated benchmarks, all evaluated models exhibit substantial performance degradation on our OmniNavBench. The highest `Success Rate` (`SR`), which measures whether the agent issues a stop action at goal area, achieved is merely 8.74% by OmniNav on the Unitree Aliengo platform while most configurations fall below 2%. Crucially, this underperformance is consistent across all three morphologies which indicates a fundamental limitation rather than platform-specific issues. Consequently, the finding underscores the necessity of OmniNavBench as a rigorous testbed to bridge the gap between specialized success and true cross-task cross-embodiment generalization.

Findings2: Disparity Between Isolated Skills and Sequential Execution. `Sub-Goal Completion` (`SGC`) is defined as the mean progress across sub-tasks, where each sub-task progress score ranges from 0 to 1. The contrast between `SGC` and the `Success Rate` (`SR`) reveals a critical insight where models can achieve moderate progress on individual sub-tasks with `SGC` exceeding 44% for Uni-NaVid on Aliengo yet fail to complete full episodes. This disparity indicates that the primary bottleneck lies in the inability to execute multiple diverse navigation tasks sequentially within a single continuous episode. This limitation largely stems from current training paradigms which rely on datasets composed of isolated single task and fail to model the realistic transition between diverse navigation behaviors, which validates the necessity of our proposed dataset for fostering true multi-task sequential capability.

Findings3: The Last Centimeter Gap in End-to-End Navigation. `Oracle Success Rate` (`OSR`) captures instances where the agent physically reaches goal once without issuing a stop action. Interestingly, Uni-NaVid achieves 26.47% `OSR` on the Unitree Aliengo yet secures only a 1.96% `SR`. This significant disparity between higher `OSR` values and the lower `SR` suggests that agents frequently navigate correctly to the vicinity of target goals but fail to issue correct termination signals. These results reflect that current end-to-end navigation models successfully handle path planning but fail at the final decision stage which indicates they are still missing the "last centimeter" capability required for real-world deployment.

Findings4: Deficiencies in Dynamic Social Interaction. Performance on dynamic interaction tasks exposes significant limitations. `Human Follow Success` (`HFS`), the task completion rate, and `Human Follow Ratio` (`HFR`), the proportion of valid following, remain low across all configurations. The best `HFS` of 45.16% achieved by Uni-NaVid on Carter degrades substantially on legged robots. This persistent violation of personal space constraints reveals a lack of social awareness and poses critical challenges for real-world deployment.

Findings5: Sensitivity to Morphological Variations. Performance varies dramatically across robot morphologies even for identical algorithms. OmniNav achieves 8.74% `SR` on the quadrupedal Aliengo but drops to 1.96% on the wheeled Carter, while the humanoid platform consistently produces the lowest performance due to its complex dynamics. This sensitivity can be attributed to the inherent differences in embodiment complexity where distinct kinematic constraints influence trajectory feasibility and the increased control difficulty of bipedal locomotion introduces greater physical instability compared to wheeled or quadrupedal platforms.

B. Failure Mode Analysis

To provide a more granular characterization of the performance bottlenecks, we categorize the observed failures into four primary modes (Fig. 4 top): (i) `Goal-reaching Drift` (`Drift`), where the agent reaches the target but leaves again; (ii) `Termination Signal Failure` (`Stuck`), where the agent reaches the goal but fails to claim completion; (iii) `Premature Termination` (`Prem. Stop`), involving episodes that stop before reaching the target vicinity; (iv) `Fall`, which accounts for failures resulting from falling. To further disentangle the failure patterns, we extend analysis to the interplay between scenario types and task performance,

TABLE V: **Ablation study of language prompt strategies.** All metrics are reported in %. Best results are bolded.

Model	Variant	SR ↑	CSR ↑	SGC ↑	OSR ↑	SPL ↑	HFS ↑	HFR ↑	SII ↓	EQA-ACC ↑
OmniNav	Origin	8.74	1.94	**35.30**	15.53	7.01	14.29	**17.49**	**3.40**	–
	Concise	13.65	1.94	26.85	**16.51**	10.41	7.86	8.91	8.98	–
	First Person	**14.47**	**3.88**	25.02	14.32	**11.88**	**25.72**	13.06	13.18	–
	Verbose	8.65	2.94	25.09	9.47	6.69	**25.72**	11.88	7.88	–
Uni-NaVid	Origin	1.96	2.94	**44.54**	26.47	1.56	23.33	**32.75**	18.46	6.25
	Concise	4.55	1.82	38.74	25.45	**3.72**	24.24	27.85	**14.46**	**10.53**
	First Person	**5.83**	1.94	39.83	**27.18**	3.28	**35.71**	31.10	19.34	5.26
	Verbose	3.88	**3.88**	39.47	18.45	2.85	32.14	31.59	15.50	5.26

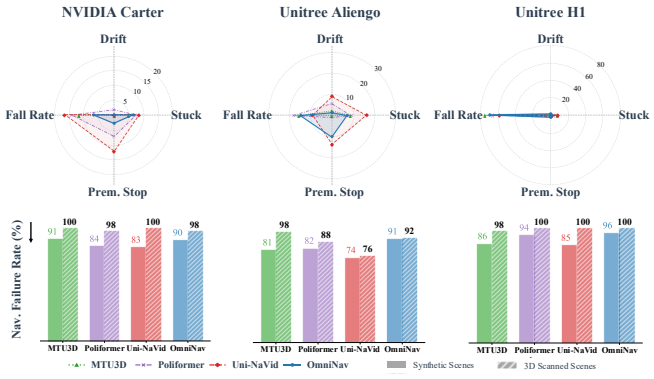

Fig. 4: **Failure Mode Analysis across Embodiments.** Radar charts comparing four failure types for four models. Bar chats comparing total failure on two scenes.

investigating how the structural complexity of different environments influences the distribution of failure modes.

Findings6: Failure Modes Various across Embodiments. The distribution of failure modes differs substantially across robot morphologies (Fig. 4 top). On the wheeled Carter, `Stuck` and `Prem. Stop` constitute the dominant failure modes, while `Fall` remains minimal. Falls on the wheeled Carter are caused by collisions rather than balance loss. In contrast, the humanoid H1 exhibits an elevated rate of `Fall` (60–80%), which suppresses the relative proportion of other failure types. The quadrupedal Aliengo presents a more balanced distribution. We attribute this divergence to the inherent physical properties and control stability margins of each embodiment. The wheeled platform possesses a statically stable base, meaning that policy uncertainty or sub-optimal actions typically manifest passively as hesitation or idling (`Stuck`, `Prem. Stop`). Conversely, the bipedal H1 requires continuous active balancing and is highly sensitive to control noise. Consequently, similar policy imperfections propagate into dynamic instability, causing immediate falls rather than benign stops. Thus, the variation in failure modes reflects how distinct dynamic constraints amplify different consequences of the same policy limitations.

Findings7: Failure Modes Between Synthetic and Real Scenes. A comparative analysis (Fig. 4 bottom) between synthetic environments and 3D scanned scenes reveals a consistent performance degradation within photorealistic reconstructions. `Failure Rate` is defined as $1 - \texttt{SR}$. Across all model-embodiment configurations, Matterport3D scenes induce significantly higher failure rates than their synthetic counterparts. We attribute this decline to the inherent structural complexity of real-world scans, characterized by irregular geometries, dense clutter, and narrow passages, which impose stricter demands on both perceptual understanding and precise kinematic planning. Consequently, while synthetic scenes with predictable layouts tend to mask policy limitations, the unstructured nature of real-world environments exposes the models' fragility and highlights the critical gap between simplified training proxies and realistic deployment scenarios.

C. Instruction Style Analysis

Findings8: Sensitivity to Instruction Variations. To investigate the impact of linguistic variability on model robustness, we conducted systematic tests on the OmniNav and Uni-NaVid using the Unitree Aliengo platform. The results in Tab. V reveal that navigation performance is highly sensitive to instruction phrasing, with distinct styles eliciting varied responses across different metrics. Notably, the *Concise* and *First Person* variants consistently improve `Success Rate` (`SR`) compared to the original instructions for both models, whereas the *Verbose* style yields comparable or degraded performance. Furthermore, while `SGC` and `HFR` consistently decrease across all stylized variants, `HFS` peaks specifically under the *First Person* style. This observed sensitivity suggests that current models overfit to specific syntactic patterns and struggle to generalize across diverse linguistic formulations, highlighting a critical need for greater linguistic diversity during training to achieve robust natural language understanding.

V. Conclusion

We present OmniNavBench, a comprehensive platform that bridges the divide between specialized navigation primitives and general-purpose embodiment. Through the integration of composite instructions and diverse robot morphologies, we provide a robust testbed for cross-skill and cross-embodiment generalization. Crucially, our analysis reveals a systemic failure in current unified models to handle sequential skill transitions. These findings underscore two critical directions: robust progress tracking for sequential coordination, and reliable termination decisions to close the gap between goal reaching and task completion. More broadly, our results highlight urgent challenges in dynamic interaction, morphological adaptability, and instruction robustness, underscoring the need for next-generation agents capable of long-horizon reasoning.

Acknowledgement. This work was supported by the Natural Science Foundation of China (Grant No. 62503323).

References

[1] Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018.

[2] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683, 2018.

[3] Abhijat Biswas, Allan Wang, Gustavo Silvera, Aaron Steinfeld, and Henny Admoni. Socnavbench: A grounded simulation testing framework for evaluating social navigation. *ACM Transactions on Human-Robot Interaction (THRI)*, 11(3):1–24, 2022.

[4] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017.

[5] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

[6] Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems*, 33: 4247–4258, 2020.

[7] An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Zaitian Gongye, Xueyan Zou, Jan Kautz, Erdem Bıyık, Hongxu Yin, Sifei Liu, and Xiaolong Wang. Navila: Legged robot vision-language-action model for navigation. *arXiv preprint arXiv:2412.04453*, 2024.

[8] Yifei Dong, Fengyi Wu, Qi He, Heng Li, Minghan Li, Zebang Cheng, Yuxuan Zhou, Jingdong Sun, Qi Dai, Zhi-Qi Cheng, et al. Ha-vln: A benchmark for human-aware navigation in discrete-continuous environments with dynamic multi-human interactions, real-world validation, and an open leaderboard. *arXiv preprint arXiv:2503.14229*, 2025.

[9] Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. *arXiv preprint arXiv:2307.00595*, 2023.

[10] Chen Gao, Si Liu, Jinyu Chen, Luting Wang, Qi Wu, Bo Li, and Qi Tian. Room-object entity prompting and reasoning for embodied referring expression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):994–1010, 2023.

[11] Chen Gao, Liankai Jin, Xingyu Peng, Jiazhao Zhang, Yue Deng, Annan Li, He Wang, and Si Liu. Octonav: Towards generalist embodied navigation. *arXiv preprint arXiv:2506.09839*, 2025.

[12] Edmund T Hall and Edward T Hall. *The hidden dimension*, volume 609. Anchor, 1966.

[13] Yulong Huang, Yonggang Zhang, Peng Shi, Zhemin Wu, Junhui Qian, and Jonathon A Chambers. Robust kalman filters based on gaussian scale mixture distributions with application to target tracking. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 49(10):2082–2096, 2017.

[14] Mukul Khanna, Ram Ramrakhya, Gunjan Chhablani, Sriram Yenamandra, Theophile Gervet, Matthew Chang, Zsolt Kira, Devendra Singh Chaplot, Dhruv Batra, and Roozbeh Mottaghi. Goat-bench: A benchmark for multi-modal lifelong navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16373–16383, 2024.

[15] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *European Conference on Computer Vision*, pages 104–120. Springer, 2020.

[16] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. *arXiv preprint arXiv:2010.07954*, 2020.

[17] Jungdae Lee, Taiki Miyanishi, Shuhei Kurita, Koya Sakamoto, Daichi Azuma, Yutaka Matsuo, and Nakamasa Inoue. Citynav: Language-goal aerial navigation dataset with geographic information. *arXiv preprint arXiv:2406.14240*, 2024.

[18] Sihao Lin, Zerui Li, Xunyi Zhao, Gengze Zhou, Liuyi Wang, Rong Wei, Rui Tang, Juncheng Li, Hanqing Wang, Jiangmiao Pang, Anton van den Hengel, Jiajun Liu, and Qi Wu. Vlnverse: A benchmark for vision-language navigation with versatile, embodied, realistic simulation and evaluation, 2025. URL https://arxiv.org/abs/2512.19021.

[19] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16488–16498, 2024.

[20] Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. *arXiv preprint arXiv:2108.03298*, 2021.

[21] NVIDIA. Isaac Sim. URL https://github.com/isaac-sim/IsaacSim.

[22] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den

Hengel. 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, 2020.

[23] Zhiwen Qiu, Ziang Liu, Wenqian Niu, Tapomayukh Bhattacharjee, and Saleh Kalantari. Egocognav: Cognition-aware human egocentric navigation. *arXiv preprint arXiv:2511.17581*, 2025.

[24] Tim Schreiter, Tiago Rodrigues de Almeida, Yufei Zhu, Eduardo Gutierrez Maestro, Lucas Morillo-Mendez, Andrey Rudenko, Luigi Palmieri, Tomasz P Kucner, Martin Magnusson, and Achim J Lilienthal. Thör-magni: A large-scale indoor motion capture recording of human movement and robot interaction. *The International Journal of Robotics Research*, 44(4):568–591, 2025.

[25] Xinshuai Song, Weixing Chen, Yang Liu, Weikai Chen, Guanbin Li, and Liang Lin. Towards long-horizon vision-language navigation: Platform, benchmark and method. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12078–12088, 2025.

[26] Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. Habitat 2.0: Training home assistants to rearrange their habitat. *Advances in neural information processing systems*, 34:251–266, 2021.

[27] Hanqing Wang, Jiahe Chen, Wensi Huang, Qingwei Ben, Tai Wang, Boyu Mi, Tao Huang, Siheng Zhao, Yilun Chen, Sizhe Yang, et al. Grutopia: Dream general robots in a city at scale. *arXiv preprint arXiv:2407.10943*, 2024.

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

[29] Shaoan Wang, Jiazhao Zhang, Minghan Li, Jiahang Liu, Anqi Li, Kui Wu, Fangwei Zhong, Junzhi Yu, Zhizheng Zhang, and He Wang. Trackvla: Embodied visual tracking in the wild. *arXiv preprint arXiv:2505.23189*, 2025.

[30] Zun Wang, Jialu Li, Yicong Hong, Yi Wang, Qi Wu, Mohit Bansal, Stephen Gould, Hao Tan, and Yu Qiao. Scaling data generation in vision-and-language navigation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12009–12020, 2023.

[31] Meng Wei, Chenyang Wan, Jiaqi Peng, Xiqian Yu, Yuqiang Yang, Delin Feng, Wenzhe Cai, Chenming Zhu, Tai Wang, Jiangmiao Pang, et al. Ground slow, move fast: A dual-system foundation model for generalizable vision-and-language navigation. *arXiv preprint arXiv:2512.08186*, 2025.

[32] Meng Wei, Chenyang Wan, Xiqian Yu, Tai Wang, Yuqiang Yang, Xiaohan Mao, Chenming Zhu, Wenzhe Cai, Hanqing Wang, Yilun Chen, et al. Streamvln: Streaming vision-and-language navigation via slowfast context modeling. *arXiv preprint arXiv:2507.05240*, 2025.

[33] Xinda Xue, Junjun Hu, Minghua Luo, Xie Shichao, Jintao Chen, Zixun Xie, Quan Kuichen, Guo Wei, Mu Xu, and Zedong Chu. Omninav: A unified framework for prospective exploration and visual-language navigation. *arXiv preprint arXiv:2509.25687*, 2025.

[34] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[35] Hang Yin, Xiuwei Xu, Linqing Zhao, Ziwei Wang, Jie Zhou, and Jiwen Lu. Unigoal: Towards universal zero-shot goal-oriented navigation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19057–19066, 2025.

[36] Naoki Yokoyama, Ram Ramrakhya, Abhishek Das, Dhruv Batra, and Sehoon Ha. Hm3d-ovon: A dataset and benchmark for open-vocabulary object goal navigation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5543–5550. IEEE, 2024.

[37] Licheng Yu, Xinlei Chen, Georgia Gkioxari, Mohit Bansal, Tamara L Berg, and Dhruv Batra. Multi-target embodied question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6309–6318, 2019.

[38] Kuo-Hao Zeng, Zichen Zhang, Kiana Ehsani, Rose Hendrix, Jordi Salvador, Alvaro Herrasti, Ross Girshick, Aniruddha Kembhavi, and Luca Weihs. Poliformer: Scaling on-policy rl with transformers results in masterful navigators. *arXiv preprint arXiv:2406.20083*, 2024.

[39] Jiazhao Zhang, Kunyu Wang, Shaoan Wang, Minghan Li, Haoran Liu, Songlin Wei, Zhongyuan Wang, Zhizheng Zhang, and He Wang. Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks. *arXiv preprint arXiv:2412.06224*, 2024.

[40] Jiazhao Zhang, Anqi Li, Yunpeng Qi, Minghan Li, Jiahang Liu, Shaoan Wang, Haoran Liu, Gengze Zhou, Yuze Wu, Xingxing Li, et al. Embodied navigation foundation model. *arXiv preprint arXiv:2509.12129*, 2025.

[41] Zhen Zhang, Jiaqing Yan, Xin Kong, Guangyao Zhai, and Yong Liu. Efficient motion planning based on kinodynamic model for quadruped robots following per-

sons in confined spaces. *IEEE/ASME Transactions on Mechatronics*, 26(4):1997–2006, 2021.

[42] Xiaoming Zhao, Harsh Agrawal, Dhruv Batra, and Alexander G Schwing. The surprising effectiveness of visual odometry techniques for embodied pointgoal navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16127–16136, 2021.

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

[44] Gengze Zhou, Yicong Hong, Zun Wang, Chongyang Zhao, Mohit Bansal, and Qi Wu. 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, 2025.

[45] Ziyu Zhu, Xilin Wang, Yixuan Li, Zhuofan Zhang, Xiaojian Ma, Yixin Chen, Baoxiong Jia, Wei Liang, Qian Yu, Zhidong Deng, et al. Move to understand a 3d scene: Bridging visual grounding and exploration for efficient and versatile embodied navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8120–8132, 2025.