AMAP CV Lab | Alibaba Group

ABot-N0: Technical Report on the VLA Foundation Model for Versatile Embodied Navigation

AMAP CV Lab

See Contributions and Acknowledgments section for a full author list.

Abstract

Embodied navigation has long been fragmented by task-specific architectures. We introduce **ABot-N0**, a unified Vision-Language-Action (VLA) foundation model that achieves a "Grand Unification" across **5** core tasks: Point-Goal, Object-Goal, Instruction-Following, POI-Goal, and Person-Following. ABot-N0 utilizes a hierarchical "Brain-Action" architecture, pairing an LLM-based Cognitive Brain for semantic reasoning with a Flow Matching-based Action Expert for precise, continuous trajectory generation.

To support large-scale learning, we developed the **ABot-N0 Data Engine**, curating **16.9M** expert trajectories and **5.0M** reasoning samples across **7,802** high-fidelity 3D scenes (**10.7 km**^{2}). ABot-N0 achieves new **SOTA** performance across **7** benchmarks, significantly outperforming specialized models. Furthermore, our Agentic Navigation System integrates a planner with hierarchical topological memory, enabling robust, long-horizon missions in dynamic real-world environments.

Correspondence: chuzedong.czd@alibaba-inc.com, tenan.xsc@alibaba-inc.com

Project Page: https://amap-cvlab.github.io/ABot-Navigation/ABot-N0/

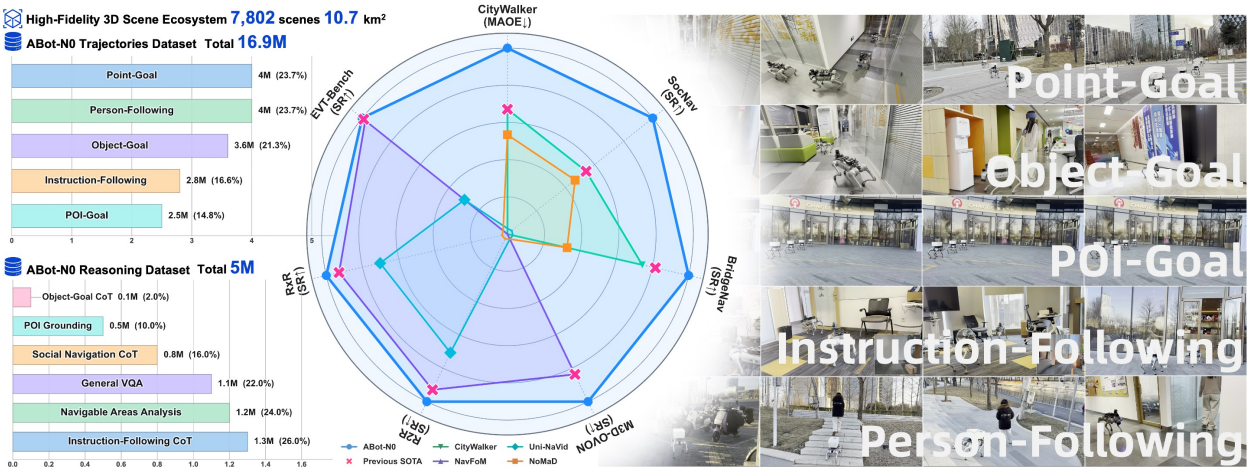

Figure 1 ABot-N0: A unified VLA foundation model for versatile embodied navigation. Powered by a massive dataset of 16.9M expert trajectories and 5M reasoning samples across diverse environments, the model achieves a "Grand Unification" across five core navigation tasks. It establishes new state-of-the-art performance across 7 challenging benchmarks and is successfully deployed in complex, dynamic real-world agentic navigation systems.

Contents

1 Introduction

The ultimate objective of Embodied AI is to develop general-purpose agents capable of perceiving complex human intentions and executing autonomous actions within non-structural, open-world environments. In this grand paradigm, navigation serves not only as a fundamental locomotion primitive but also as the critical bridge connecting high-level cognitive reasoning with low-level continuous motor control. However, contemporary navigation research remains deeply entrenched in a "fragmented paradigm". Existing methodologies [10, 20, 33, 34, 36, 37, 51, 56, 57, 62, 68, 70] often rely on isolated and specialized architectures tailored for specific tasks—such as Point-Goal, Object-Goal, or Instruction-Following navigation—which inherently limits cross-task generalization and prevents agents from extracting a unified physical prior from large-scale heterogeneous data.

To overcome these limitations, we introduce **ABot-N0**, a unified Vision-Language-Action (VLA) foundation model designed for universal embodied navigation. ABot-N0 challenges the traditional task-isolation approach by adopting a hierarchical "Brain-Action" design philosophy. The architecture comprises three pillars: a **Universal Multi-Modal Encoder** that unifies heterogeneous inputs into a shared latent space; a **Cognitive Brain** based on a pre-trained Large Language Model (LLM) for deep semantic understanding and spatial reasoning; and an **Action Expert** utilizing Flow Matching to generate precise, multi-modal trajectory distributions for continuous control.

Within the ABot-N0 framework, we achieve a "Grand Unification" of five core navigation tasks:

- **Point-Goal Navigation**: The agent must reach precise metric coordinates defined in a local frame, serving as the foundational primitive for robust locomotion and obstacle avoidance [10, 36].
- **Object-Goal Navigation**: The agent actively searches for and navigates to a specific object category in unseen environments, requiring sophisticated semantic reasoning and multimodal integration [57, 62, 70].
- **Instruction-Following Navigation**: The agent must execute long-horizon, complex natural language paths, focusing on the rigorous alignment between linguistic input and sequential action execution [33, 37, 56].
- **POI-Goal Navigation**: This task is defined by [68], which requires the agent to identify specific Points of Interest (POIs) and precisely navigate to their physical entrances, bridging outdoor and indoor environments while addressing the last-meters navigation challenge.
- **Person-Following Navigation**: This involves real-time tracking of dynamic human targets, representing a critical social capability for human-robot interaction [34, 51].

The core contributions of this work are summarized in three dimensions:

1. A Unified Embodied Foundation Model Achieving Comprehensive SOTA. ABot-N0 represents the first foundation model to achieve a "Grand Unification" across five core navigation tasks within a single architecture. Our evaluations demonstrate that ABot-N0 sets new performance benchmarks (SOTA) across a wide range of authoritative platforms, including CityWalker [36], SocNav [10], VLN-CE (R2R/RxR) [27, 29], HM3D-OVON [63], BridgeNav [68], and EVT-Bench [51].

2. The Largest Embodied Navigation Data Engine. The generalization capability of ABot-N0 is anchored by the ABot-N0 Data Engine, which has synthesized a standardized corpus of unprecedented scale. This includes a High-Fidelity 3D Scene Ecosystem comprising **7,802 scenes** covering **10.7 km^2** of the total area (6.25 km^2 indoor and 4.42 km^2 outdoor). Leveraging these assets, we curated approximately **16.9 million** expert training trajectories and **5.0 million** cognitive reasoning samples to enable robust and generalizable navigation skills.

3. A Deployable Agentic Navigation System. To bridge the gap between foundation models and practical utility, we propose the Agentic Navigation System. This framework augments ABot-N0 with an **Agentic Planner** that utilizes Chain-of-Thought (CoT) reasoning for intent decomposition and closed-loop error recovery, a hierarchical **Topo-Memory (Map-as-Memory)** for cross-scale spatial knowledge deposition. The system has been successfully deployed on the **Unitree Go2** quadrupedal robot using an **NVIDIA Jetson Orin NX** module, achieving 2Hz VLA inference and 10Hz closed-loop control in complex, dynamic real-world environments.

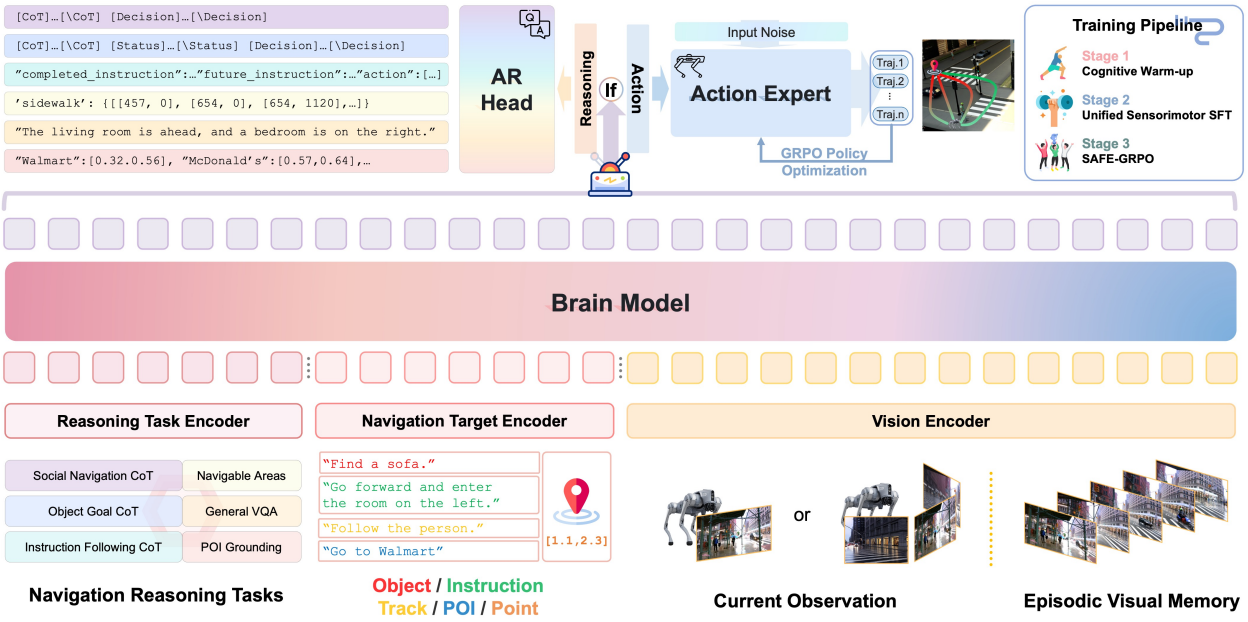

Figure 2 The Architecture of ABot-N0. The model adopts a hierarchical "Brain-Action" design. The Universal Multi-Modal Encoder unifies heterogeneous inputs (RGB observations, visual history, and goal specifications) into a shared token sequence. The Cognitive Brain (i.e., LLM) encodes these tokens and supports dual-mode operation: a Reasoning Head for high-level semantic understanding and an Action Head for motion planning. The Action Expert employs Flow Matching to generate trajectory distributions, enabling generalization across five navigation tasks.

2 ABot-N0

Existing navigation approaches often treat distinct tasks as separate domains requiring specialized architectures. We challenge this fragmented paradigm by introducing **ABot-N0**, a unified Vision-Language-Action foundation model. ABot-N0 is designed to combine high-level cognitive reasoning with low-level motion planning, seamlessly generalizing across these five core navigation tasks.

The architecture follows a hierarchical "Brain-Action" design philosophy (inspired by SocialNav [10]), comprising three key pillars (See Fig.2):

- A **Universal Multi-Modal Encoder** that unifies heterogeneous inputs into a shared latent space;
- A **Cognitive Brain** based on an Auto-Regressive Large Language Model that encodes generalizable semantic understanding;
- An **Action Expert** that uses Flow Matching [32] to generate precise, multi-modal trajectory distributions.

2.1 Universal Multi-Modal Encoder

To achieve a "Grand Unification" of embodied navigation tasks, ABot-N0 processes diverse sensor configurations and goal specifications through a flexible token-based encoding scheme.

2.1.1 Flexible Vision Interface

ABot-N0 supports agnostic visual inputs to accommodate different robot morphologies:

- **Current Observation (O_t):** We utilize a Vision Transformer (ViT) [12] to encode RGB observations. The interface supports dynamic configurations. In the panoramic mode, the left, front and right views are fed individually into the vision encoder, distinguished by view-specific special tokens. This design preserves

the spatial semantics of each view without the geometric distortion introduced by image stitching. In the front-view mode, only the single forward-facing image is processed.

- **Episodic Visual Memory (M_S):** ABot-N0 maintains an explicit visual history buffer. Relevant historical frames are encoded and appended to the context window, allowing the model to handle Partially Observable Markov Decision Processes (POMDPs).

2.1.2 Heterogeneous Navigation Target Encoder

ABot-N0 is able to interpret disparate goal definitions through a unified interface:

- **Semantic Goals (Text-based):** This category encompasses Instruction-Following, Object-Goal, POI-Goal, and Person-Following tasks. Whether the goal is a natural language command, a specific object category, a Point of Interest (POI) name, or a textual description of a target person, we employ the text tokenizer of LLM to embed the goal description directly.
- **Geometric Goals (Point-based):** For Point-Goal tasks, the target coordinates (x, y) are defined in the local Bird's Eye View (BEV) frame. These coordinates are projected into the shared embedding dimension via an MLP, allowing the LLM to process geometric vectors as pseudo-tokens.

2.1.3 Reasoning Task Encoder

To bridge the gap between perception and action, we introduce a Reasoning Task Encoder. This module injects specific task descriptions to activate relevant reasoning circuits within the LLM (e.g., *"Where is Luckin Coffee?"* or *"Identify the zebra crossing area"*). These tasks serve as auxiliary objectives during training, helping the Brain build a robust representation of the environment.

2.2 The Cognitive Brain

The backbone of ABot-N0 is a pre-trained LLM (Qwen3-4B [59]). It takes the reasoning instructions, navigation goals, visual history, and current observations as input.

Unlike sequential CoT approaches used in recent LLMs [15, 59], our architecture employs a task-conditional design. The Reasoning Head and the Action Head operate as distinct branches (an "If-Else" relationship based on the task token), rather than a strictly serial pipeline during inference:

- **Cognitive Activation:** During training, the model is supervised to perform explicit reasoning tasks—such as analyzing scene traversability, identifying social norms, or grounding distinct POIs. These auxiliary tasks are crucial for aligning the high-level semantic representation with the constraints of the physical world .
- **Navigation Decision Support:** For navigation tasks, the Brain leverages the rich, physically-grounded latent context cultivated by the reasoning tasks to directly condition the Action Expert.

2.3 The Action Expert

For precise motion control, we employ a Flow Matching Action Head, conditional on the context provided by the Brain. The Action Expert predicts a short-term trajectory plan consisting of a sequence of 5 waypoints in the local BEV frame. To ensure full control over the robot's pose, each waypoint includes both 2D coordinates and the yaw orientation:

$$\mathcal{W} = \{(x_1, y_1, \theta_1), (x_2, y_2, \theta_2), \ldots, (x_5, y_5, \theta_5)\} \tag{1}$$

where (x_i, y_i) denotes the position and θ_i represents the local heading angle at step i.

We adopt Flow Matching over deterministic regression for two primary reasons:

- **Continuous Precision:** It enables the generation of high-precision continuous values for both translation and rotation, which are essential for smooth robot control and stable heading adjustments.
- **Multi-Modal Distribution Modeling:** In large-scale navigation datasets, similar input conditions often correspond to multiple valid expert behaviors (e.g., bypassing an obstacle from either the left or the right).

Figure 3 High-Fidelity 3D Scene Ecosystem. Our data engine integrates diverse indoor and outdoor environments. *Left*: Indoor scenes span residential spaces to large-scale public venues (offices, malls, transit stations). *Right*: Outdoor scenes include real-world scans of intersections and parks, alongside the dynamic virtual city SocCity. All scenes are annotated with traversable navigation graphs for collision-free trajectories generation.

Deterministic regression tends to average these divergent modes, leading to invalid or collision-prone paths. Flow Matching effectively models this complex distribution, allowing the planner to sample valid, distinct trajectories that reflect the diversity of expert demonstrations.

3 Data Engine

To bridge the critical gap between perceptual understanding and actionable policy, we introduce the **ABot-N0 Data Engine**, a unified synthesis pipeline designed for scalable embodied navigation learning. This engine integrates three synergistic layers: a **High-Fidelity 3D Scene Ecosystem** (Sec. 3.1) offering photorealistic, semantically annotated environments; a **Universal Trajectories Dataset** (Sec. 3.2) aggregating expert demonstrations across five core navigation paradigms (Point-Goal, Object-Goal, Instruction-Goal, POI-Goal, and Person-Following); and a **Cognitive Reasoning Dataset** (Sec. 3.3) that grounds decision-making in explicit spatial-social logic. By systematically aligning high-level intent with precise execution, this pipeline provides the essential data fuel for training a generalist, socially compliant navigator.

3.1 High-Fidelity 3D Scene Ecosystem

To construct a foundation model capable of generalizing across diverse physical environments, we collect and build high-fidelity 3D scenes to support the closed-loop simulation. As shown in Fig.3, we categorize our environments into **Indoor** and **Outdoor** domains, integrating widely-used academic benchmarks with our large-scale proprietary 3D Gaussian Splatting (3DGS) scans [24]. Detailed statistics of our 3D scene collection are summarized in Fig.4.

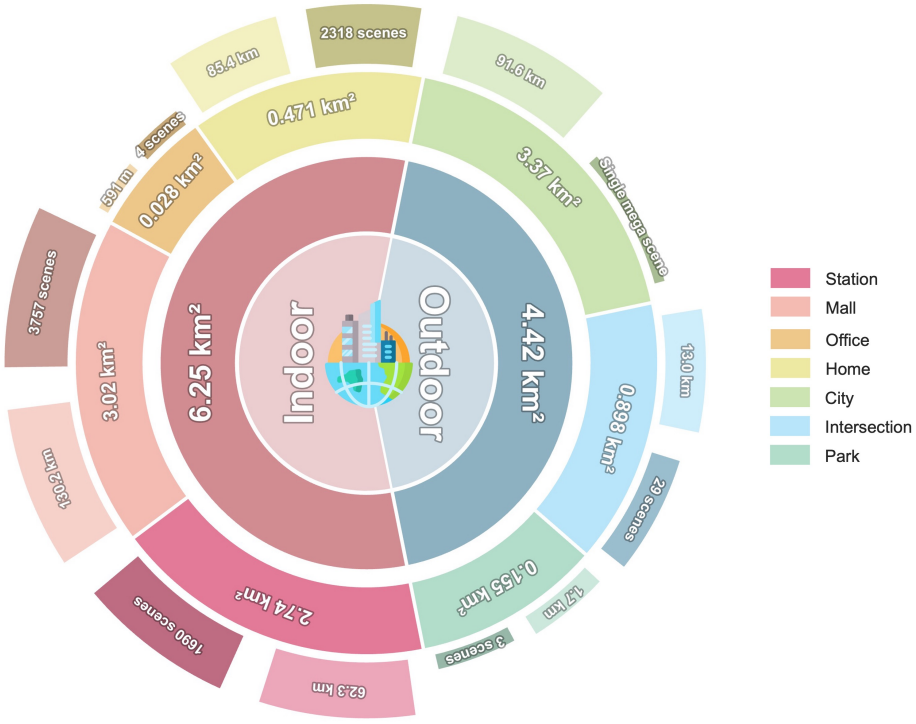

Figure 4 3D Scene Ecosystem Statistics. Our collection comprises **7,802** high-fidelity 3D scenes, covering **6.25** km^2 of indoor environments (offices, malls, stations, homes) and **4.42** km^2 of outdoor environments (intersections, parks, city). Navigation graphs totaling **384,754 meters** enable collision-free trajectory synthesis across diverse spatial scales—from compact residential units to expansive transit hubs and urban environments.

Crucially, we implement a unified pipeline for navigation graph annotation to transform these raw visual reconstructions into navigable environments. Every scene in our collection has been manually annotated with traversable road networks, ensuring that all generated expert trajectories are collision-free and socially compliant.

3.1.1 Indoor Environments: From Domestic to Public

Our indoor collection is designed to bridge the gap between narrow domestic spaces and unstructured public venues.

Residential Scenes (Home). Domestic environments are critical for embodied agents serving household roles. We integrate HM3D [38] and InteriorGS [45], comprising a total of **2,318** diverse residential units. These scenes provide rich semantic layouts (e.g., bedrooms, kitchens) essential for Object-Goal and Instruction-Following navigation. The annotated navigation graphs in these scenes strictly define walkable floor space, excluding furniture and obstacles to ensure precise interaction.

Public Spaces (Office, Mall, Station). Moving beyond the home, we introduce a large-scale proprietary dataset of public interiors reconstructed via 3DGS.

- **Offices:** Featuring narrow corridors, cubicles, and meeting rooms, challenging the agent's fine-grained motion control.
- **Shopping Malls:** Spanning vast areas (**3.02** km^2) with complex topology, multiple floor levels, and reflective surfaces (e.g., glass storefronts).
- **Transit Stations:** Characterized by channelized pedestrian traffic, navigational bottlenecks (e.g., turnstiles), and large-scale waiting halls interspersed with narrow transit corridors.

3.1.2 Outdoor Environments: Static Scans and Dynamic City

The outdoor domain introduces unique challenges such as traffic rules, unstructured terrain, and vast spatial scales.

Real-World Scans (Intersection, Park). We utilize high-precision LiDAR and photogrammetry to reconstruct **37** real-world outdoor scenes.

- **Intersections:** These scenes allow the agent to learn the geometry of crosswalks, vehicle lanes, and sidewalks. Our road network annotation strictly forbids the agent from entering vehicle lanes, enforcing traffic rule compliance.
- **Parks:** These unstructured park environments (**8 scenes**) cover narrow walking paths and open plazas.

Dynamic Virtual City (SocCity). To complement static scans, we employ SocCity [10], a large-scale (**3.37** $\boldsymbol{km^2}$) virtual urban environment. Unlike static 3DGS scans, SocCity supports dynamic simulation of vehicles and pedestrians. The occupancy map here is hierarchically provided to distinguish between pedestrian walkways, crosswalks, vehicle roads and dynamic obstacles, enabling the training of *Socially-Aware Flow Exploration* (SAFE-GRPO) [10].

3.2 ABot-N0 Trajectories Dataset

Leveraging the unified navigation graphs and high-fidelity environments established in Sec. 3.1, we construct the **ABot-N0 Trajectory Dataset**, a massive, standardized corpus of sensorimotor demonstrations comprising approximately **16.9 million** training trajectories across five navigation paradigms. Unlike prior works that rely on fragmented datasets with task-specific formats, this collection unifies Point-Goal, Instruction-Following, Object-Goal, POI-Goal, and Person-Following tasks into a shared action representation. By deploying optimal planners on navigation graphs and aggregating large-scale human/robot video logs, we align diverse high-level intent signals with precise, continuous trajectory distributions, providing the essential supervision for the Action Expert to learn a generalized motion policy.

3.2.1 Point-Goal Trajectories

Point-goal navigation serves as the foundational locomotion primitive in our data engine, focusing on the agent's ability to reach precise metric coordinates while maintaining environmental and social awareness. To cultivate a robust and generalizable navigation foundation, we curate a large-scale collection of approximately **4.0 million** point-goal navigation trajectories. This dataset is synthesized through a multi-source data engine that transforms raw 3D scenes, uncurated internet videos, and heterogeneous robotic logs into standardized navigation episodes (See Fig.5). The construction process focuses on ensuring visual diversity, metric accuracy, and kinematically feasible motion priors.

Deriving Pseudo-Trajectories from Internet Videos. To capture the long-tail distribution of global architectural styles and weather conditions, we process **2.0 million** pseudo-trajectories from a curated collection of first-person urban exploration videos. We employ a scalable vision-to-trajectory pipeline: (1) **Structure Recovery**: Utilizing π^3 [52] for dense 3D reconstruction from monocular streams; (2) **Metric Alignment**: Applying MoGe [50] to resolve scale ambiguity and align the reconstructed paths with real-world metric units; (3) **Episode Synthesis**: Sampling diverse point-goal pairs along the camera's motion manifold to generate kinematically consistent navigation samples. This approach allows the data engine to ingest "passive" observation data and convert it into "active" navigation knowledge.

Synthesis from High-Fidelity 3D Scenes. We leverage a vast repository of 3D scenes in our collection and the large-scale dynamic SocCity [10]. Point-goal episodes are generated by sampling reachable coordinate pairs (s, g) and computing optimal paths using path-finding algorithms. Crucially, we incorporate recovery trajectories by initializing agents in off-path or near-collision states, forcing the model to learn error-correction behaviors essential for real-world robustness. Finally, **1.7 million** training samples are generated through these 3D scene assets.

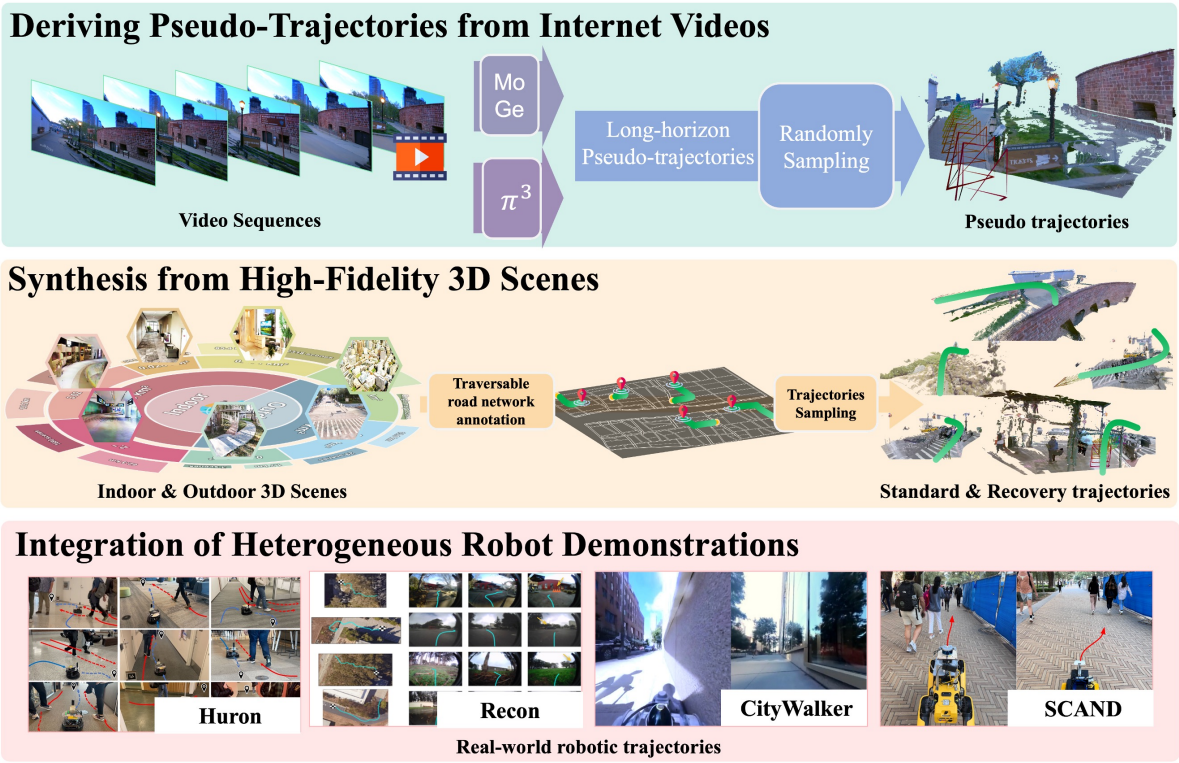

Figure 5 Point-Goal Navigation Data Pipeline. Our **4.0M** point-goal trajectory dataset aggregates three complementary streams: (Top) **2.0M** pseudo-trajectories from first-person internet videos via 3D structure recovery and metric alignment; (Middle) **1.7M** synthetic trajectories from high-fidelity 3D scenes with annotated navigation graphs; (Bottom) **340K** real-world robot demonstrations from heterogeneous datasets, providing ground-truth physical dynamics.

Integration of Heterogeneous Robot Demonstrations. We incorporate **340K** high-quality trajectories from diverse real-world robotic platforms (e.g., SCAND [22], HuRoN [18], Recon [41], and CityWalker [36] teleoperation data). Unlike simulated or video-derived data, these trajectories provide ground-truth physical dynamics and sensor noise characteristics. We standardize these multi-modal logs into a unified point-goal format, ensuring that the model internalizes the intricate constraints of real-world hardware, such as non-holonomic motion limits and sensor-to-actuation latency.

By aggregating these three streams, the data engine provides a comprehensive trajectory distribution that spans from idealistic optimal paths to complex, noise-perturbed real-world maneuvers, forming the backbone of our model's navigation intelligence. More details of this pipeline is reported in our SocialNav [10].

3.2.2 Instruction-Following Trajectories

Within the Visual Language Navigation (VLN) framework [11, 37, 56, 65–67], instruction following is defined as a task setting where an agent must navigate a 3D environment by continuously processing egocentric visual observations to reach a destination specified by natural language instructions. This paradigm emphasizes the alignment and execution of the navigation action derived from complex linguistic input. Its significance extends beyond simply reaching the target; it critically evaluates the agent's adherence to the specific paths and procedural milestones described in the instruction, providing a more authentic assessment of language grounding and instruction-following capabilities. To enable the model to comprehend natural language and plan the corresponding trajectories, we constructed the following datasets.

VLN-CE R2R/RxR. This dataset is built within the VLN-CE (Continuous Environment) framework [27] using Matterport3D [5] scenes. For each training instance, we recorded the navigation trajectory waypoints, the

Figure 6 Diverse Instruction-Following Tasks in InteriorGS [45]. We illustrate three representative navigation scenarios designed to enhance instruction-following capabilities: (a) Door-Traversal for navigating narrow passages, (b) Language-Guided Person Search, and (c) Short-Horizon navigation for atomic movement primitives.

corresponding natural language instruction, and panoramic observations comprising three rendered views (front, left, and right) at each step.

The VLN-CE R2R [27] component consists of approximately 10K continuous clips, each representing a shortest-path navigation task paired with an instruction. Employing a teacher forcing protocol, we unrolled these trajectories into individual time-step samples and applied data balancing to normalize the action distribution. This process filtered the initial pool of 600K samples down to 200K retained training trajectory samples.

VLN-CE RxR [29] represents a significant expansion over R2R in both scale and complexity. Compared to R2R, RxR features longer path lengths, more complex topological structures, and instructions characterized by higher linguistic density—incorporating frequent and semantically diverse references to landmarks and complicated spatial relationships. Consequently, it offers richer supervision for long-horizon language grounding and continuous decision-making. This subset contains approximately 20K continuous clips. Following a similar balancing protocol, we curated 1.3 million training samples from an initial pool of 1.8 million. In total, the combined VLN-CE R2R/RxR dataset comprises **1.5 million** training trajectory samples.

Door-Traversal. To address the specific challenges associated with navigating narrow passages and bottlenecks, we synthesized 6,000 door-traversal clips, which were subsequently processed into **0.3 million** training trajectory samples. Leveraging InteriorGS [45], we sampled initial poses in the vicinity of doorways while enforcing a diverse distribution of post-traversal orientations and travel distances. Additionally, we annotated each trajectory with multiple distinct instructions to maximize diversity in both geometric configurations and linguistic variability. A schematic illustration of the data acquisition process is presented in Fig. 6.

Language-Guided Person Search. We generated a diverse library of digital human assets, consisting of 200 3DGS avatars for the InteriorGS [45] environment and 2,000 polygonal mesh characters for the Habitat simulator [40]. Descriptions for these avatars were synthesized from attribute combinations (e.g., clothing, gender, pose) and subsequently refined using LLMs to ensure natural linguistic flow. For each training episode, we populated the scene with 2 to 5 randomly instantiated avatars, designating one as the navigation target based on a specific language query. We then initialized the agent at a random starting pose and computed ground-truth trajectories using the geodesic shortest path to the target character (Fig. 6). In total, this process yielded 4,000 training clips, culminating in a dataset of **0.2 million** trajectory samples.

Short-Horizon. Focusing on atomic navigation primitives characterized by short horizons and explicit metric

goals, we curated 12,000 clips within InteriorGS [45], which were subsequently processed into **0.8 million** training trajectory samples. This dataset encompasses rotational actions (e.g., turn left, turn around, turn left 60 degrees), translational movements (e.g., move forward 1 meter, move backward 3 meters), and compound sequences that integrate both rotation and translation as shown in Fig. 6. These samples are specifically designed to enhance the model's fine-grained execution capabilities. Furthermore, the inclusion of these fundamental short-horizon tasks serves to effectively regularize and stabilize the model training process.

3.2.3 Object-Goal Trajectories

Object-Goal Navigation (ObjectNav) requires an agent to actively search for and navigate to an instance of a specific object category within an unseen environment. Unlike traditional Point-Goal Navigation (PointNav) tasks, ObjectNav relies heavily on semantic reasoning and the effective integration of multimodal information. Furthermore, the heterogeneity of indoor layouts and the complexity caused by environmental occlusions exacerbate the difficulty of the task, posing significant challenges to the model's generalization and robustness in diverse real-world scenarios.

HM3D and OVON. For the HM3D ObjectNav task [63], we focused on 6 target object categories across 80 scenes, constructing a training set of 0.4 million trajectories derived from 10K expert clips. To further strengthen the model's generalization capabilities, we incorporated the Open-Vocabulary ObjectNav (OVON) benchmark [63]. Unlike standard ObjectNav, which operates on a closed set of categories, OVON requires the agent to locate targets conditioned on free-form language queries. This task rigorously assesses performance across unseen categories, synonyms, and semantically distant objects. Spanning 145 scenes, we sampled 10 starting poses for each object instance, generating a comprehensive navigation dataset comprising 35K episodes and a total of 1.4 million training trajectories. In total, we obtained **1.8 million** trajectories.

OVON-sub. To mitigate the complexity associated with long-horizon planning and object occlusion in standard OVON tasks, we constructed a short-range navigation subset specifically designed to enhance the model's visual-semantic alignment. We generated this subset by processing the original ground-truth trajectories with a specific truncation strategy: we identify the onset of target visibility—the first timestep where the target object enters the agent's Field of View (FoV); and retain the trajectory segment from this moment until the episode's termination. This process yielded 6,000 training clips including **0.2 million** training trajectory samples.

InteriorGS. Compared to HM3D [38], InteriorGS [45] provides superior photorealistic rendering fidelity, featuring 1,000 scenes annotated with bounding boxes for over 700 object categories. However, the absence of pre-defined visible position annotations precludes the direct application of the standard HM3D ObjectNav data generation protocol. To bridge this gap, we developed a custom trajectory generation pipeline driven by object visibility estimation as shown in Figure 7. Finally, we construct a training set of 40K clips with **1.6 million** trajectory samples.

3.2.4 POI-Goal Trajectories

We have proposed the POI-Goal task along with its corresponding dataset; detailed definitions of the task and specifics of dataset construction can be found in Zhao et al. [68]. This task aims to bridge outdoor and indoor environments, enabling embodied agents to navigate from open, coarse-grained outdoor scenes into structured, fine-grained indoor spaces.

Given a single input image, we first employ image segmentation and depth estimation techniques to generate a local occupancy grid map. We then use the A* algorithm to plan a trajectory from the starting position to the POI, which serves as the ground-truth path. Next, we feed both the input image and the generated trajectory into the Wan2.1-I2V Diffusion Transformer video generation model [48] to synthesize first-person-view observation videos that simulate the navigation process of an embodied agent. Finally, **2.5 million** training trajectories are sampled from these generated videos.

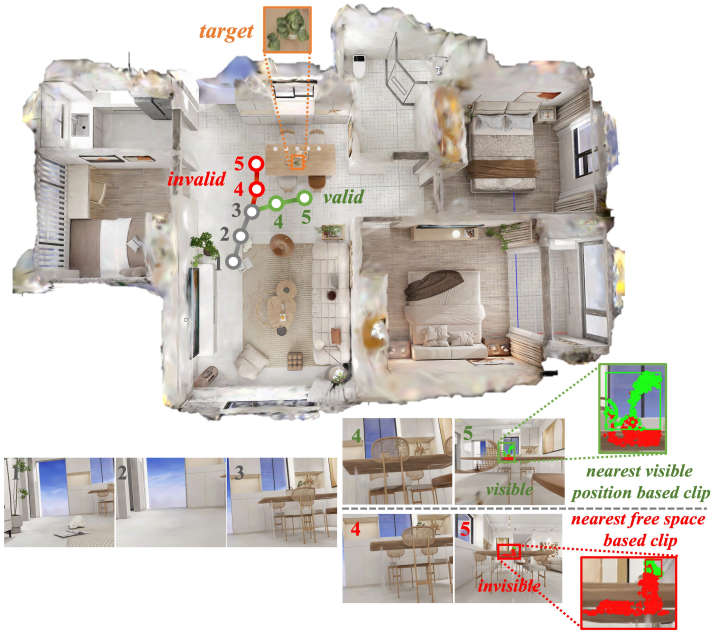

Figure 7 Trajectory generation pipeline for Object Goal tasks in InteriorGS [45]. Our method selects the nearest visible position as the clip endpoint, ensuring cleaner training data compared to the naive approach of using the nearest free space.

Figure 8 Trajectory generation pipeline for Person-Following tasks. We track human trajectories and sample viewpoints along the path to create training data for person-following navigation.

3.2.5 Person-Following Trajectories

We referenced the dataset creation method of TrackVLA [51]. However, since its training data was not fully open sourced, we generated our own training data based on its open-sourced humanoid avatar motion trajectories (See Fig.8). We generated synthetic tracking sequences through three distinct proximity configurations (2.0m, 1.5m, and 1.2m target distances) to emulate real-world tracking scenarios. For each distance configuration, we employed the A* algorithm to compute optimal waypoints while maintaining stable tracking observations. Following TrackVLA's established taxonomy, we generated three tracking challenge categories (Single-Target Tracking/STT, Distracted Tracking/DT, Ambiguity Tracking/AT) for each distance configuration. Each category-distance combination contains 400K samples, yielding 3.6 million base training instances (3 distances $\times$ 3 categories $\times$ 400K). To address target absence scenarios where no tracked subject appears in the field of view, we supplemented the dataset with 400K target-absent cases. The final composite dataset comprises **4.0 million** training samples, with each sample containing current frame images, historical frame sequences, their corresponding future trajectories, and associated navigation commands.

3.3 ABot-N0 Reasoning Dataset

While the Trajectory Dataset provides the sensorimotor experience (how to move), it lacks the explicit causal logic required for robust decision-making in unseen environments. To bridge this gap, we introduce the **ABot-N0 Reasoning Dataset**, a large-scale corpus comprising **5.0 million** samples designed to activate the high-level cognitive capabilities of the VLA Brain. It compels the model to explicitly verbalize its understanding of the scene geometry, semantic landmarks, and social norms when committing to an action. The dataset is structured hierarchically, progressing from general navigation skills to task-specific reasoning, and general world knowledge.

3.3.1 General Navigation Skills

This subset instills the fundamental spatial and social common sense required for any mobile agent, regardless of its specific mission. We adopt the data generation pipeline proposed in our SocialNav [10].

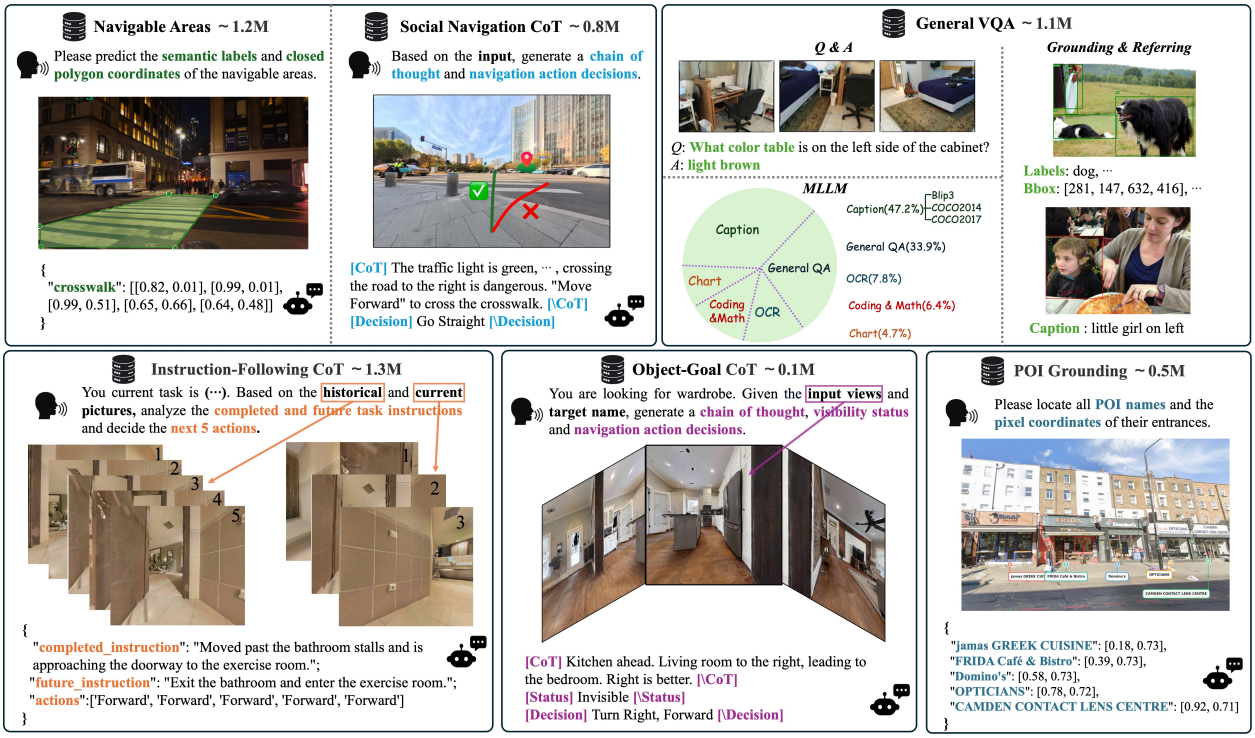

Figure 9 Examples from the ABot-N0 Reasoning Dataset. The dataset encompasses diverse reasoning tasks including Navigable Areas Analysis, Social Navigation CoT, Instruction-Following Reasoning, Object-Goal Reasoning, POI Grounding, and General VQA.

Navigable Areas Analysis (1.2M Samples). A core competency for any embodied agent is distinguishing between "physically traversable" and "socially traversable" areas. To ground the agent's visual perception in physical affordances, we construct a dataset comprising **1.2 million** ego-centric images collected from diverse outdoor video streams. For each image, we manually annotate precise polygons that delineate socially compliant zones (e.g., sidewalks, crosswalks) while strictly excluding non-navigable areas (e.g., roadways, lawns). The model is trained to output the semantic labels and polygon coordinates of these regions given a visual observation. This task helps ensure that the generated waypoints in downstream tasks always reside within safe boundaries, effectively constraining the action space with social norms.

Social Navigation Chain-of-Thought (0.8M Samples). To enable the model to handle complex social interactions (e.g., yielding to pedestrians, obeying traffic lights), we construct a massive dataset of navigation rationales. Leveraging the automated pipeline from SocialNav [10], we utilize a powerful VLM (Qwen-VL-Max [4]) as a teacher agent to generate structured Chain-of-Thought (CoT) rationales. Each sample contains the reasoning process that explicitly explains the logic behind a decision (e.g., *"The traffic light is red, so I must wait"*).

3.3.2 Instruction-Following Reasoning

In long-horizon navigation, embodied agents often struggle to align visual observations with instruction progress, leading to trajectory drift and task failure. Leveraging the sequential structure of global instructions, we decompose long-horizon tasks into atomic sub-instructions. For each sub-instruction, we identify its spatiotemporal termination point, defined as a "milestone node". These nodes form a Visual CoT that explicitly grounds the execution process. We automate this pipeline using MLLMs, specifically Gemini-3 Pro [14] and Qwen3-VL [59]. By processing 30,000 clips from the VLN-CE R2R/RxR datasets [27, 29], we generated **1.3 million** training samples. Each sample consists of current and historical observations, completed instruction, future instruction, and predicted action. More pipeline details are in our NavForesee [33].

3.3.3 Object-Goal Reasoning

End-to-end policies, typically trained with sparse action supervision, often overfit to implicit biases within the training distribution. They lack explicit constraints on intermediate reasoning steps, such as target visibility or spatial relationships. This limitation severely hampers generalization across open-vocabulary targets and unseen scenarios. To address this, we utilized the OVON dataset [63] to construct Object Reasoning, a large-scale CoT dataset.

Specifically, we prompt MLLMs to generate structured reasoning chains based on panoramic observations and target information. These chains encompass four key components: (1) target visibility assessment, (2) spatial relationship analysis, (3) path planning, and (4) the corresponding atomic action sequence. To ensure data quality, we implemented a consistency check mechanism. We filter out hallucinations by verifying that the generated action sequences align with ground-truth trajectories. Ultimately, we curated **0.1 million** high-quality samples. This design provides interpretable reasoning for decision-making and effectively bridges the gap between high-level semantic intent and low-level control.

3.3.4 POI Grounding

Standard object detection datasets often lack the precision required for navigation, specifically in linking semantic Point-of-Interest (POI) goals to physical entry points. To bridge this gap, we introduce the POI Grounding Dataset.

We curated a subset of street-view images from the BridgeNav Dataset [68], covering diverse urban scenes. Unlike traditional facade-level labeling, we focus on precise entry point localization. Utilizing Qwen3-VL-Plus [59], we automatically generate structured annotations as tuples: $\langle \text{POI Name}, \text{Entrance Coordinates } (u, v) \rangle$. This process yields **0.5 million** Visual Question Answering (VQA) pairs, providing the spatial grounding necessary for the ABot-N0 to translate semantic POI targets into executable waypoints.

3.3.5 General VQA

To preserve general visual-linguistic representations and enhance generalization in unseen environments, we incorporated a diverse set of VQA and Embodied Question Answering (EQA) datasets. Specifically, we utilized Blip3 [7], COCO2014 [30], COCO2017 [31] and MAmmoTH-VL [16] to provide general commonsense knowledge. To bolster the model's grounding and referring capabilities, we included the RefCOCO series [23] and Objects365 [44]. Furthermore, we integrated R2R-EnvDrop [47] and ScanQA [3]. These indoor embodied datasets are critical for developing spatial awareness and trajectory planning skills. Collectively, we curated a total of **1.1 million** training samples from these sources.

4 Training Recipe

Training a unified foundation model across five distinct navigation tasks presents a significant challenge: *multimodal alignment*. The model must balance high-level semantic reasoning (e.g., understanding the relationship between observed images and navigation instructions) with low-level waypoints-based control.

To address this, we propose a three-stage curriculum learning pipeline: **Cognitive Warm-up**, **Unified Sensorimotor SFT**, and **Post-Training Value Alignment via SAFE-GRPO**. This progressive approach ensures that ABot-N0 first acquires a robust understanding of the world before learning to act within it.

4.1 Phase 1: Cognitive Warm-up

Before learning strictly "how to move", the agent must learn "what to see" and "how to reason." In this phase, we utilize the **ABot-N0 Reasoning Dataset** to align the LLM backbone with the embodied domain. By freezing the Vision Encoder and text tokenizer, we fine-tune the LLM core using Next Token Prediction (NTP) loss across diverse tasks (e.g., general navigation skills, specific navigation reasoning tasks and general VQAs). This process enables the model to interpret visual scenes, ground textual instructions to physical entities, and perform complex spatial reasoning. Throughout this phase, the Action Expert remains frozen, ensuring that

gradients are dedicated exclusively to optimizing the visual-linguistic representations and cognitive foundations of the model.

4.2 Phase 2: Unified Sensorimotor SFT

Once the cognitive backbone is established, we introduce the **ABot-N0 Trajectory Dataset**. This phase unifies all five navigation tasks into a single multi-task training regime. To preserve the reasoning capabilities acquired in the previous phase and prevent catastrophic forgetting, we employ a mixed-training strategy by augmenting the trajectory data with a replay buffer of reasoning data at an approximate ratio of 20%. During this phase, we perform **Dual-Head Optimization** to jointly optimize the AR Head (autoregressive reasoning) and the Action Expert (flow matching). The training objective is defined by the joint loss function:

$$\mathcal{L}_{Phase2} = \lambda_{txt}\mathcal{L}_{NTP}(\theta_{brain}) + \lambda_{flow}\mathcal{L}_{CFM}(\theta_{action}|\theta_{brain}) \tag{2}$$

where $\mathcal{L}_{NTP}$ denotes the cross-entropy loss for text generation and $\mathcal{L}_{CFM}$ represents the Conditional Flow Matching loss for trajectory generation. This integrated approach grounds high-level semantic plans into precise continuous physical actions, resulting in a unified policy capable of executing diverse and complex embodied tasks.

4.3 Phase 3: Post-Training Value Alignment via SAFE-GRPO

Although Phase 2 equips the model with general navigation capabilities via Imitation Learning (IL), IL inherently suffers from a limitation: it captures the surface-level statistics of expert behaviors but fails to master the causal structure underlying normative conduct in complex social environments. For instance, an agent might learn to walk on a sidewalk, but fail to understand why deviating into a flowerbed or a vehicle lane is strictly prohibited.

To bridge this gap, we employ **SAFE-GRPO** (Socially-Aware Flow Exploration GRPO) [10], a flow-based reinforcement learning framework designed to explicitly enforce social compliance.

During this phase, we freeze the Brain model to preserve the semantic understanding acquired in previous stages, and exclusively fine-tune the Action Expert. The training is conducted with the expert trajectories from SocCity [10], which provides the rich annotations necessary to calculate precise rewards.

The policy is optimized to maximize a composite reward function $\mathcal{R}$ that balances social compliance with navigation efficiency:

$$\mathcal{R} = w_{soc}\mathcal{R}_{social} + w_{exp}\mathcal{R}_{expert} + w_{sm}\mathcal{R}_{smooth} + w_{eff}\mathcal{R}_{eff} \tag{3}$$

- **Social Compliance Reward ($\mathcal{R}_{social}$):** This is the primary signal for the alignment of values. It is derived from a ground-truth *semantic occupancy map* provided by the SocCity environment. The agent incurs a heavy penalty if its predicted trajectory traverses non-navigable or socially restricted zones (e.g., lawns, restricted lanes, walking humans), forcing the model to internalize the causal rule: "navigable geometry $\neq$ socially acceptable path".
- **Expert Similarity Reward ($\mathcal{R}_{expert}$):** To prevent reward hacking, we maintain a regularization term that encourages the policy to stay within a reasonable distribution of the expert demonstrations.
- **Smoothness & Efficiency ($\mathcal{R}_{smooth}, \mathcal{R}_{eff}$):** These terms penalize jerky movements and encourage progress toward the goal, ensuring the generated trajectories remain feasible for real-world robot execution.

Through this value-aligned post-training, ABot-N0 ensures strict adherence to social norms when executing all navigation tasks.

5 Evaluation

To empirically validate ABot-N0 as a generalist foundation model, we conduct a comprehensive evaluation across five distinct navigation paradigms. By benchmarking against task-specific state-of-the-art methods

Table 1 Open-Loop Point-Goal Evaluation on CityWalker Benchmark [36]. We evaluate MAOE metric in each critical scenario for all methods. Percentages under scenarios indicate their data proportions. The "Mean" column shows scenario means averaged over six scenarios; "All" shows sample means over all data samples.

Method	Mean	Turn 8%	Crossing 12%	Detour 12%	Proximity 6%	Crowd 7%	Other 55%	All 100%
GNM [42]	16.2	31.1	14.8	12.5	14.7	12.8	11.0	12.1
ViNT [43]	16.5	31.1	15.4	12.9	14.8	13.3	11.6	12.6
NoMaD [46]	19.1	35.1	18.5	15.6	18.1	14.3	12.8	12.1
CityWalker [36]	15.2	26.6	14.1	13.9	14.3	12.0	10.4	11.5
ABot-N0	**11.2**	**21.3**	**9.8**	**12.8**	**8.1**	**8.8**	**6.3**	**7.6**

Table 2 Point-Goal Performance Comparison on the Closed-Loop SocNav Benchmark [10]. ABot-N0 demonstrates superior navigation performance and social compliance across all metrics.

Method	Navigation Performance			Social Compliance	
	SR↑	RC↑	SPL↑	DCR↑	TCR↑
GNM* [42]	43.3	62.4	37.0	26.5	28.7
ViNT* [43]	45.6	66.2	39.5	31.4	33.8
NoMaD* [46]	41.1	60.5	35.4	29.5	31.6
CityWalker [36]	47.8	64.7	44.7	36.1	36.6
ABot-N0	**88.3**	**92.1**	**79.2**	**85.1**	**85.4**

on standard platforms, we demonstrate that ABot-N0 not only unifies these disparate tasks within a single architecture but also achieves superior performance through its reasoning-driven policy.

5.1 Point-Goal Task

Following the rigorous evaluation protocol established in SocialNav [10], we assess the Point-Goal navigation performance of ABot-N0 through a comprehensive combination of open-loop benchmarks and closed-loop simulations.

Evaluation Setup and Metrics. We conduct the open-loop evaluation on the **CityWalker Benchmark** [36]. The primary metric is Maximum Average Orientation Error (MAOE), which measures the maximum angular deviation between the predicted and ground-truth actions over the future horizon, serving as a proxy for human-likeness. For closed-loop evaluation, we utilize the **SocNav Benchmark** [10], deploying the agent in the unseen and high-fidelity environments. We report standard navigation metrics: Success Rate (SR), Route Completion (RC), and Success weighted by Path Length (SPL). Crucially, to quantify social awareness, we employ Distance Compliance Rate (DCR) and Time Compliance Rate (TCR), which calculate the percentage of distance and time the agent spends strictly within socially traversable zones (e.g., sidewalks) versus restricted areas (e.g., lawns or roadways).

Results and Analysis. The quantitative results are presented in Table 1 and Table 2. In the open-loop setting, ABot-N0 achieves a mean MAOE of **11.2**, significantly outperforming the previous state-of-the-art CityWalker (15.2) and other baselines. In the closed-loop setting, ABot-N0 demonstrates remarkable robustness and social adherence. It attains an **88.3%** Success Rate, nearly doubling the performance of the baseline (47.8%). More importantly, in terms of social compliance, ABot-N0 achieves a DCR of **85.1%** compared to 36.1% for the baseline. This confirms that our model does not merely reach the goal but does so by actively respecting social norms, verifying the effectiveness of our hierarchical "Brain-Action" architecture and the SAFE-GRPO alignment strategy.

Table 3 Instruction-following evaluation on VLN-CE benchmarks [27, 29]. We compare ABot-N0 with state-of-the-art methods on R2R-CE and RxR-CE Val-Unseen splits. Methods marked with * use the waypoint predictor from Hong et al. [19]. ABot-N0 achieves superior performance using panoramic observations, demonstrating strong generalization in continuous navigation environments.

Method	Observation				R2R-CE Val-Unseen				RxR-CE Val-Unseen		
	S.RGB	Pano.	Depth	Odo.	NE↓	OS↑	SR↑	SPL↑	NE↓	SR↑	SPL↑
HPN+DN* [28]		✓	✓	✓	6.31	40.0	36.0	34.0	-	-	-
CMA* [19]		✓	✓	✓	6.20	52.0	41.0	36.0	8.76	26.5	22.1
Sim2Sim* [25]		✓	✓	✓	6.07	52.0	43.0	36.0	-	-	-
GridMM* [53]		✓	✓	✓	5.11	61.0	49.0	41.0	-	-	-
DreamWalker* [49]		✓	✓	✓	5.53	59.0	49.0	44.0	-	-	-
Reborn* [1]		✓	✓	✓	5.40	57.0	50.0	46.0	5.98	48.6	42.0
ETPNav* [2]		✓	✓	✓	4.71	65.0	57.0	49.0	5.64	54.7	44.8
HNR* [54]		✓	✓	✓	4.42	67.0	61.0	51.0	5.50	56.3	46.7
AG-CMTP [8]		✓	✓	✓	7.90	39.0	23.0	19.0	-	-	-
R2R-CMTP [8]		✓	✓	✓	7.90	38.0	26.0	22.0	-	-	-
InstructNav [37]		✓	✓	✓	6.89	-	31.0	24.0	-	-	-
LAW [39]	✓		✓	✓	6.83	44.0	35.0	31.0	10.90	8.0	8.0
CM2 [13]	✓		✓	✓	7.02	41.0	34.0	27.0	-	-	-
WS-MGMap [9]	✓		✓	✓	6.28	47.0	38.0	34.0	-	-	-
AO-Planner [6]		✓	✓		5.55	59.0	47.0	33.0	7.06	43.3	30.5
Seq2Seq [26]	✓		✓		7.77	37.0	25.0	22.0	12.10	13.9	11.9
CMA [26]	✓		✓		7.37	40.0	32.0	30.0	-	-	-
NaVid [66]	✓				5.47	49.0	37.0	35.0	-	-	-
Uni-NaVid [65]	✓				5.58	53.5	47.0	42.7	6.24	48.7	40.9
NaVILA [11]	✓				5.22	62.5	54.0	49.0	6.77	49.3	44.0
StreamVLN [56]	✓				4.98	64.2	56.9	51.9	6.22	52.9	46.0
InternVLA-N1 (S2)† [55]	✓				4.89	60.6	55.4	52.1	6.41	49.5	41.8
InternVLA-N1 (S1+S2)† [55]	✓		✓		4.83	63.3	58.2	54.0	5.91	53.5	46.1
NavFoM (Four views) [67]		✓			4.61	**72.1**	61.7	55.3	4.74	64.4	56.2
ABot-N0		✓			**3.78**	70.8	**66.4**	**63.9**	**3.83**	**69.3**	**60.0**

5.2 Instruction-Following Task

Following the evaluation standards established in previous methods [11, 37, 56, 65–67], we assess the instruction-following capabilities of ABot-N0 within continuous environments.

Evaluation Setup and Metrics. We conduct evaluations on the Val-Unseen splits of two distinct datasets of the **VLN-CE**: **R2R-CE** [27], which focuses on short, goal-oriented commands, and **RxR-CE** [29], which involves longer, temporally grounded path descriptions. We report three standard metrics to measure precision and efficiency: Navigation Error (NE), the average geodesic distance to the target; Success Rate (SR), the percentage of episodes where the agent stops within 3 meters of the goal; and Success weighted by Path Length (SPL), which penalizes inefficient paths.

Results and Analysis. As shown in Table 3, ABot-N0 demonstrates superior performance across both benchmarks. On the **R2R-CE** Val-Unseen split, ABot-N0 achieves an SR of **66.4%**, outperforming the previous state-of-the-art method, NavFoM (Four views), by a margin of **4.7%**. Notably, we observe a **8.6%** improvement in SPL. This significant gain indicates that ABot-N0 has superior environmental understanding and path planning capabilities, allowing it to navigate more efficiently without compromising success. A consistent performance trend is observed on the **RxR-CE** Val-Unseen benchmark, where SR and SPL reach **69.3%** and **60.0%**, respectively. We attribute these improvements primarily to our model architecture and the synergy derived from joint training on diverse indoor-outdoor, multi-scene, and multi-task datasets.

Table 4 Object-Goal Navigation Performance on HM3D-OVON [63]. We evaluate ABot-N0 on the Open-Vocabulary ObjectNav benchmark across three validation splits: Val-Seen, Val-Seen-Synonyms, and Val-Unseen. Our method achieves superior performance using only panoramic RGB input, without requiring depth sensors or odometry.

Method	Observation				Val-Seen		Val-Seen-Synonyms		Val-Unseen	
	S.RGB	Pano.	Depth	Odo.	SR↑	SPL↑	SR↑	SPL↑	SR↑	SPL↑
BC	✓				11.1	4.5	9.9	3.8	5.4	1.9
DAgger	✓				11.1	4.5	9.9	3.8	5.4	1.9
RL	✓				18.1	9.4	15.0	7.4	10.2	4.7
DAgRL	✓				41.3	21.2	29.4	14.4	18.3	7.9
BCRL	✓				39.2	18.7	27.8	11.7	18.6	7.5
VLFM [62]	✓		✓	✓	35.2	18.6	32.4	17.3	35.2	19.6
DAgRL+OD [63]	✓		✓	✓	38.5	21.1	39.0	21.4	37.1	19.8
Uni-NaVid [65]	✓				41.3	21.1	43.9	21.8	39.5	19.8
MTU3D [70]	✓		✓	✓	55.0	23.6	45.0	14.7	40.8	12.1
NavFoM (Four views) [67]		✓			40.1	27.1	45.4	32.6	45.2	**31.9**
ABot-N0		✓			**55.3**	**32.1**	**55.4**	**33.2**	**54.0**	30.5

5.3 Object-Goal Task

Following the evaluation standards established in previous methods [62, 63, 65, 67], we assess the object-goal capabilities of ABot-N0 on the HM3D-OVON dataset [63].

Evaluation Setup and Metrics. We report results across the three validation splits based on semantic proximity: Val-Seen (categories seen during training), Val-Seen-Synonyms (unseen synonyms of training categories), and Val-Unseen (semantically distinct unseen categories). We report Success Rate (SR) and Success Weighted by Path Length (SPL) as our primary metrics.

Results and Analysis. As shown in Table 4, our method outperforms current SOTA approaches using exclusively RGB inputs, eliminating the need for depth sensors or pose. We attribute this success to our proposed high-level cognitive reasoning with low-level motion planning architecture, and the extensive diversity of our training data. Quantitatively, ABot-N0 surpasses MTU3D [70] by **13.2%** in SR on the challenging Val-Unseen split. Furthermore, while MTU3D suffers a significant performance drop of 14.2% from Val-Seen to Val-Unseen, ABot-N0 exhibits a negligible decline of only 1.3%. This minimal generalization gap underscores the consistency of our model's performance across diverse scenarios.

5.4 POI-Goal Task

Adhering to the evaluation framework proposed in our BridgeNav [68], we assess the performance of ABot-N0 on the BridgeNav Dataset. This benchmark is specifically designed to test the "last-meters" navigation capability, requiring the agent to identify and precisely enter specific POIs from outdoor environments.

Evaluation Setup and Metrics. The primary objective of this task is to pass through designated entrances, which vary significantly in width across real-world scenes. To rigorously evaluate navigation precision, we report the **Success Rate (SR)** under three strict distance thresholds to the entrance center: 0.1m, 0.2m, and 0.3m. Furthermore, **Navigation Efficiency** is measured by calculating the spatial deviation from optimal paths. We report the deviation metrics across average, best-case, and worst-case scenarios to provide a holistic view of path fidelity.

Results and Analysis. As presented in Table 5, ABot-N0 outperforms existing approaches across all metrics. Most notably, it achieves a substantial **70.1%** improvement in SR at the strictest **0.1m** threshold. This result highlights our model's exceptional fine-grained control capabilities, enabling it to successfully negotiate narrow entrances where baselines often fail. In terms of efficiency, our approach reduces the average trajectory deviation by **30.5%** and yields up to a **55.6%** improvement in the best-case scenario. These gains suggest that ABot-N0's reasoning-driven planning not only finds the target but does so by generating smooth, quasi-optimal paths that closely mirror geometric ground truth.

Table 5 POI-Goal navigation evaluation on BridgeNav Dataset [68]. We compare ABot-N0 with state-of-the-art methods on task success rate under varying distance thresholds (0.1m, 0.2m, 0.3m) and the navigation efficiency. ABot-N0 demonstrates superior success rate in POI entrance navigation and achieves more efficient path planning.

Method	Task Success Rate			Navigation Efficiency		
	SR (0.1m) ↑	SR (0.2m) ↑	SR (0.3m) ↑	TR (mean) ↓	TR (best) ↓	TR (worst) ↓
NoMaD [46]	4.13	15.07	29.20	31.35	5.45	85.91
Citywalker [36]	13.79	41.02	65.96	15.58	0.76	56.47
OmniNav [58]	18.78	46.99	72.39	14.16	0.99	53.79
ABot-N0	**32.14**	**71.50**	**88.68**	**9.84**	**0.44**	**51.38**

Table 6 Quantitative results on the EVT-Bench [51] (single view). We evaluate performance using Success Rate (SR), Tracking Rate (TR), and Collision Rate (CR). The symbols † and ‡ indicate the adoption of GroundingDINO [35] and the SoM [60]+GPT-4o [21] pipeline, respectively. Within each category, the top performance is highlighted in **bold**, and the runner-up is underlined.

Method	Single-Target Tracking (STT)			Distracted Tracking (DT)			Ambiguity Tracking (AT)		
	SR↑	TR↑	CR↓	SR↑	TR↑	CR↓	SR↑	TR↑	CR↓
IBVS‡ [17]	42.9	56.2	3.75	10.6	28.4	6.14	15.2	39.5	**4.90**
PoliFormer† [64]	4.67	15.5	40.1	2.62	13.2	44.5	3.04	15.4	41.5
EVT [69]	24.4	39.1	42.5	3.23	11.2	47.9	17.4	21.1	45.6
EVT‡ [69]	32.5	49.9	40.5	15.7	35.7	53.3	18.3	21.0	44.9
Uni-NaVid [65]	25.7	39.5	41.9	11.3	27.4	43.5	8.26	28.6	43.7
TrackVLA [51]	85.1	78.6	**1.65**	57.6	63.2	5.80	50.2	63.7	17.1
NavFoM [67]	85.0	80.5	-	61.4	68.2	-	-	-	-
TrackVLA++ [34]	86.0	81.0	2.10	66.5	68.8	**4.71**	51.2	63.4	15.9
ABot-N0	**86.9**	**87.6**	8.54	**66.7**	**75.4**	11.6	**67.3**	**79.5**	7.05

5.5 Person-Following Task

To assess the model's ability to follow a dynamic person, we evaluate it on the EVT-Bench [51]. This benchmark simulates realistic, dynamic social scenarios, requiring the agent not only to detect the target but also to predict their motion under occlusion and distinguish them from distractors.

Evaluation Setup and Metrics. The evaluation is stratified into three difficulty levels: Single-Target Tracking (STT) (nominal following), Distracted Tracking (DT) (presence of similar-looking pedestrians causing identity confusion), and Ambiguity Tracking (AT) (frequent severe occlusions requiring target re-identification). We report three standard metrics: Success Rate (SR), the percentage of episodes where the agent successfully follows the target for the full duration without losing track; Tracking Rate (TR), the proportion of time steps where the agent maintains the target within a valid distance and field of view; and Collision Rate (CR), measuring navigation safety.

Results and Analysis. As detailed in Table 6, ABot-N0 establishes a new state-of-the-art across all categories. In the nominal STT task, it achieves an SR of **86.9%**, surpassing the previous best, TrackVLA++, by 0.9%. The performance gap widens significantly in complex scenarios: in the highly challenging AT task, our model achieves a **16.1%** improvement in both SR and TR. This suggests that in complex environments, ABot-N0's superior reasoning and target prediction capabilities effectively contribute to safer, more robust tracking planning compared to reactive baselines.

6 Application

In this section, we bridge the gap between foundation model capabilities and practical utility by integrating our model into a comprehensive autonomous system. We first introduce the **Agentic Navigation System**

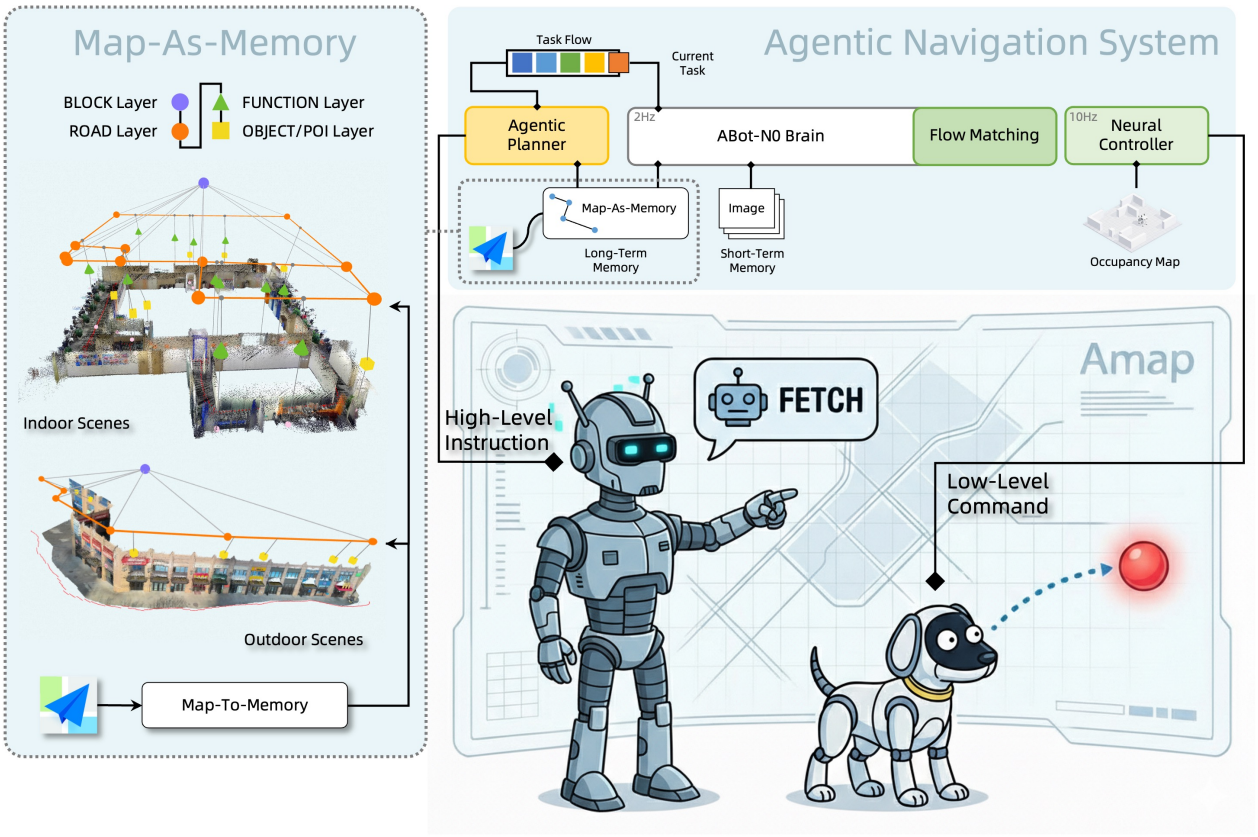

Figure 10 Agentic Navigation System Overview. Our system integrates ABot-N0 with an agentic framework comprising Agentic Planner, Actor, short-term Episodic Memory, and long-term Topo-Memory to handle complex real-world navigation tasks.

(Sec. 6.1), an agentic framework that augments ABot-N0 with planning, topological memory, and self-reflection to address the heterogeneity of real-world directives. Subsequently, we detail the physical deployment setup (Sec. 6.2.1) and demonstrations that substantiate the system's operational robustness and capacity in executing complex, long-horizon missions across diverse open-world environments.

6.1 Agentic Navigation System

To enable effective deployment of the ABot-N0 foundation model in complex real-world scenarios, we propose the Agentic Navigation System (Fig. 10). Real-world user instructions are inherently heterogeneous, requiring distinct execution strategies based on the spatial context. A single mission often necessitates a combination of *Approaching* (Point-Goal) for traversing known spaces, *Reaching* (Object-Goal, POI-Goal) for precise target discovery, and *Interaction* (Instruction-Following, Person-Following) for dynamic engagement. Recognizing that different navigational phases demand different primitives, our agentic framework systematically decomposes high-level intents into these specific, executable sub-tasks, thereby enabling robust and adaptive long-horizon autonomy.

6.1.1 Architecture

Departing from traditional end-to-end control strategies, our system adopts an Agentic VLA framework comprising four core modules: the Agentic Planner, the Actor, the short-term Episodic Memory, and the long-term Topo-Memory. We formulate the task execution process as a Partially Observable Markov Decision Process (POMDP).

Acting as the high-level controller, the Agentic Planner leverages the reasoning capabilities of VLMs to

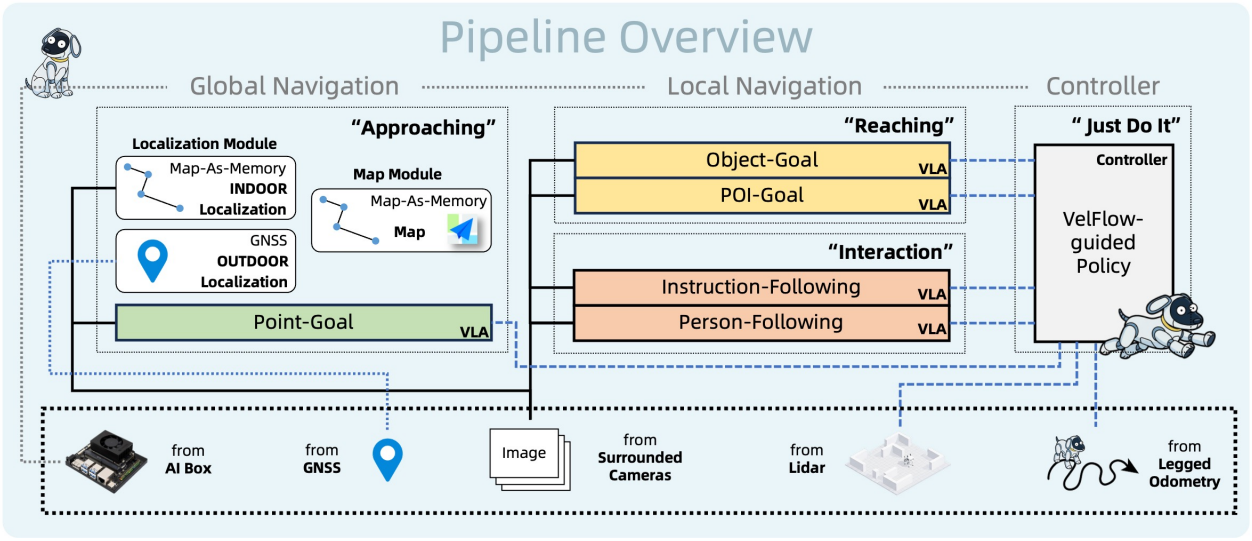

Figure 11 Task Execution Pipeline in Agentic Navigation System. The system orchestrates a hierarchical navigation workflow: (1) Global Navigation employs *Approaching* (Point-Goal) to traverse known spaces using topological memory, (2) Local Navigation utilizes *Reaching* (Object-Goal/POI-Goal) for precise target discovery and *Interaction* (Instruction-Following, Person-Following) for dynamic engagement through visual grounding, and (3) Neural Controller executes low-level velocity-based motion control to realize the planned trajectory.

interpret high-level intents. It systematically decomposes these intents into a sequence of sub-tasks executable by the Actor. The Actor module incorporates the ABot-N0 and a Neural Controller to execute these specific sub-tasks. The short-term Episodic Memory maintains recent observations within the current episode, enabling context-aware decision-making and immediate error recovery. The long-term Topo-Memory serves as a persistent spatial knowledge repository, maintaining hierarchical topological representations that enable cross-scale navigation and experience accumulation. Upon the conclusion of a sub-task, the Self-Reflector of the planner assesses the completion status. In the event of failure, it diagnoses the cause and triggers a re-planning process.

6.1.2 Map as Memory

To endow Vision-Language Navigation model with human-like environmental familiarity—progressively refining their spatial knowledge across scales ranging from residential interiors to urban environments—we propose Topo-Memory, a spatio-temporal topological memory module. Rather than treating maps as static backdrops, this framework redefines them as a sustained, augmented external memory, enabling the structured deposition and dynamic updating of spatial cognition.

Multi-layer Topology. Topo-Memory employs a hierarchical topological graph to organize spatial knowledge. By abstracting environments at multiple granularities, the framework decouples task complexity and ensures cross-scale capabilities from macro-planning to micro-perception:

- **Block Layer:** Defines indoor rooms and outdoor neighborhoods or campus districts. This layer facilitates coarse-grained cross-regional localization and long-range task decomposition.
- **Road Layer:** Characterizes physical connectivity via intersections, doorways, etc., providing rigid reachability constraints for path planning.
- **Function Layer:** Annotates critical semantic nodes such as rest areas, kitchens, and elevator halls, helping to translate abstract linguistic intentions into functional reachability targets.
- **Object/POI Layer:** Identifies specific entities, storefronts, etc. These serve as precise visual-semantic anchors for "last-meters" Object/POI-Goal navigation.

Unified Indoor-Outdoor Spatial Representation. We adopt a "One Map" foundation to unify indoor and outdoor navigation. This strategy integrates AMAP-based global routing with real-time visual decision-making. It enables the system to execute complex, multi-stage missions: leaving a home, exiting a residential complex, crossing urban blocks, and finally reaching a specific restaurant inside a shopping mall 3 kilometers away. This creates a comprehensive and robust autonomous navigation system.

Dynamic Maintainability. Departing from traditional static mapping, Map-as-Memory centers on a "maintenance-in-the-loop" mechanism, mitigating cognitive obsolescence caused by dynamic environmental shifts. Observations, trajectories, and interaction outcomes from every task execution are back-propagated as evidence into the topological graph. This sustained knowledge deposition enables experience-driven efficiency gains. In response to environmental dynamics—such as temporary road closures, construction barriers, or corridor congestion—the system adaptively updates graph structures or edge weights. This prevents recursive failures that stem from outdated spatial memory.

6.1.3 Agentic Planner

The Planner $\mathcal{P}$ takes the ambiguous user instruction I, current observation O_t, and historical context M_L (Topo-Memory), M_S (Episodic Visual Memory) as inputs. Leveraging CoT reasoning, it decomposes the instruction into a sequence of sub-tasks $G = \{g_1, g_2, \ldots, g_n\}$ (e.g., Object-Goal, Point-Goal) executable by the ABot-N0. This process is formalized as:

$$G = \mathcal{P}(I, M_L, M_S, O_t)$$

By harnessing the commonsense reasoning capabilities of VLMs, the Planner ensures the following:

Ambiguity Resolution. Free-form natural language instructions are frequently ambiguous and unstructured. VLA often fails to interpret such high-level semantics directly. Acting as the reasoning engine $\mathcal{P}$, our planner leverages the commonsense knowledge embedded within VLMs to parse the ambiguous instruction I and decompose it into a structured, executable sequence of sub-tasks.

Memory-Aware Planning. Our planner integrates a map memory M_L. During the planning phase, instead of blindly initiating a search, the system first retrieves information from M_L. For instance, if the instruction is *"find an oven"* and the memory indicates the *"kitchen"* is located at coordinates (x, y), the planner prioritizes generating a deterministic Point-Goal(x, y) task. This rapidly guides the robot to the target region, avoiding inefficient global Object-Goal exploration. Upon arrival, the system transitions to a local Object-Goal policy to precisely locate the specific target instance.

Strategic Coarse-to-Fine Decomposition. Motivated by the rapid degradation of *Reaching* tasks (Object-goal, POI-goal) performance over long horizons versus the inherent robustness of *Approaching* tasks (Point-Goal), we adopt a "Coarse-to-Fine" decomposition strategy. Instead of a single prone-to-failure command, the planner structures the task flow into a reliable sequence: utilizing Point-Goal for the long-range *Approaching* phase, transitioning to Object/POI-Goal for precise local *Reaching*, and concluding with task-specific *Interaction*:

$$P(W|I, M_L, M_S) = \underbrace{P(G|I, M_L, M_S)}_{\text{Agentic Planning}} \cdot \prod_{j=1}^{N} \underbrace{P(\mathcal{W}_j|g_j, M_L, M_S)}_{\text{ABot-N0}}$$

Closed-loop Reflection and Self-Correction. To bolster system robustness and establish a closed-loop control mechanism, we incorporate a VLM-based Self-Reflector $\mathcal{S}$. Upon the conclusion of a sub-task g_i, the Self-Reflector assesses the memory M_L, M_S, generating a completion status r and natural language feedback f:

$$(r, f) = \mathcal{S}(M_L, M_S, g_i)$$

In the event of task failure (i.e., r = False), the system leverages the provided feedback f to initiate a re-planning mechanism within the planner, thereby emulating human-like self-correction capabilities:

$$G' = \mathcal{P}(I, M_L, M_S, O_t, f)$$

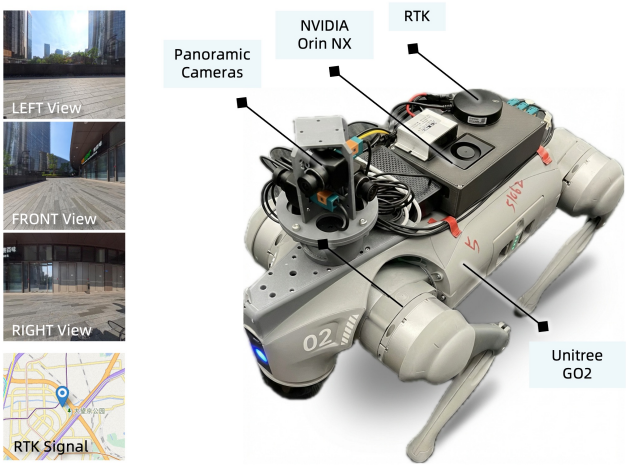

Figure 12 Hardware Platform. Our deployment platform is built on a Unitree Go2 X quadrupedal robot equipped with three monocular RGB cameras (270° horizontal FOV), a Unitree 4D LiDAR L2 for occupancy mapping, an RTK-GNSS receiver for global localization, and an onboard NVIDIA Jetson Orin NX module for real-time inference.

6.1.4 Neural Controller

To bridge the latency gap between high-level reasoning and real-time execution, we incorporate a Neural Controller based on the CE-Nav framework [61]. Although ABot-N0 generates strategic waypoints at a frequency of 2Hz, this rate is insufficient to guarantee stable obstacle avoidance in dense or dynamic environments. The Neural Controller serves as a high-speed reactive layer, operating at over 10Hz on our edge device.

Specifically, we leverage the pre-trained VelFlow expert—a conditional normalizing flow model—as a guiding prior to fine-tune a dynamics-aware refiner via Reinforcement Learning within the Isaac Sim environment. This refinement process is tailored to the specific robot hardware and its underlying locomotion policy. During inference, the Neural Controller takes the latest waypoints provided by the VLA and a real-time occupancy map derived from LiDAR scans as primary inputs. This perception-action loop enables the controller to translate abstract waypoints into precise and dynamically feasible body velocity commands (v_x, v_y, v_{yaw}). This hierarchical decoupling ensures both high-fidelity waypoint tracking and robust safety responses during complex missions.

6.2 Real-world Deployment

To validate the practical applicability of our proposed system, we deploy the complete Agentic Navigation System on a Unitree Go2 quadrupedal robot and conduct extensive real-world experiments in diverse indoor and outdoor environments.

6.2.1 Hardware setting

The experimental platform is built upon the Unitree Go2 X quadrupedal robot (Fig. 12), a dynamically stable system featuring 12 actuated degrees of freedom. For geometric environmental sensing, the robot is equipped with a Unitree 4D LiDAR L2, providing omnidirectional point cloud data.

Multi-Modal Sensor Suite. To provide the VLA model with comprehensive environmental context, we integrate a heterogeneous sensor array. Global localization is maintained via a Real-Time Kinematic GNSS (RTK-GNSS) receiver, ensuring geospatial consistency across long-horizon missions. For local scene perception, we deploy three rolling-shutter monocular RGB cameras (FOV: $H120°, V90°$ for each), rigidly mounted to cover the forward, left, and right sectors. This configuration yields a combined horizontal field of view of approximately 270°, functioning as the "eyes" of the robot for precise egocentric navigation and target identification.

Perception and Occupancy Mapping. To bridge the spatiotemporal gap between the low-frequency VLA waypoints planner and the high-frequency velocity controller, we implement a tightly coupled BEV perception framework. This framework fuses proprioceptive legged odometry with a self-developed real-time local mapping pipeline. Departing from the default height-map provided by the Unitree SDK, we developed a dedicated occupancy grid construction pipeline. By filtering transient obstacle noise and integrating temporal consistency, this pipeline generates a robust occupancy representation. This approach significantly enhances navigation reliability in crowded, human-centered environments while ensuring the state feedback remains spatially grounded for terrain-adaptive execution.

Onboard Computation. All critical perception and execution tasks, including the real-time inference of the optimized VLA model, are performed onboard an NVIDIA Jetson Orin NX module (157 TOPS, 16GB RAM).

6.2.2 Deployment Strategy

To balance high-level reasoning with real-time execution, we adopt a hybrid cloud-edge deployment architecture. The Agentic Planner is deployed on a cloud server equipped with an NVIDIA RTX 4090 GPU, leveraging its extensive computational resources for complex intent decomposition and self-reflection. In contrast, the ABot-N0 and the Neural Controller are deployed locally on an NVIDIA Jetson Orin NX module. This local deployment is critical for ensuring low-latency response and operational safety, enabling the robot to maintain autonomous navigation and obstacle avoidance even in environments with unstable or no network connectivity.

To facilitate the deployment of the VLA on edge hardware, we utilize a lightweight vision encoder and an aggressive visual token compression strategy. Specifically, we employ a compact 93M SigLIP-B/16 as the vision backbone, followed by an MLP projector to map visual features into the language token space. Furthermore, we introduce a token merging mechanism (with a merge size of 4) to compress visual tokens. This design significantly reduces visual forward-pass overhead and shortens the prefill sequence length for the LLM, thereby minimizing computational costs and terminal latency. Ultimately, the edge-side VLA model achieves an onboard inference speed of 2Hz while suffering only a 3% reduction in performance.

6.2.3 Deployment Visualization

To comprehensively evaluate the real-world applicability of ABot-N0, we design a progressive validation protocol from basic instruction execution to long-horizon complex task handling. We first validate the model's execution precision on core navigation primitives through single-task experiments, including Point-Goal (Fig. 13), Object-Goal (Fig. 14), POI-Goal (Fig. 15), Instruction Following (Fig. 16) and Person-Following (Fig. 17). We then assess the system's capability to execute long-horizon missions in open-world scenarios. As shown in Fig. 18–20, these visualizations demonstrate how the system handles comprehensive instructions spanning indoor-outdoor transitions with multi-stage intents. Results indicate that the Agentic architecture enables seamless transitions between navigation phases and exhibits strong closed-loop robustness against environmental uncertainties and task failures. We also demonstrate the system's versatility and feasibility in real-world settings through proof-of-concept experiments across diverse downstream scenarios (Fig. 21).

7 Conclusion

In this report, we presented **ABot-N0**, a unified Vision-Language-Action (VLA) foundation model that achieves a "Grand Unification" across five core embodied navigation tasks: Point-Goal, Object-Goal, Instruction-Following, POI-Goal, and Person-Following. By adopting a hierarchical "Brain-Action" architecture, our model effectively bridges high-level cognitive reasoning with low-level continuous motor control. This synergy is powered by the ABot-N0 Data Engine, the largest corpus of its kind, comprising 16.9 million expert trajectories and 5.0 million reasoning samples. Comprehensive evaluations demonstrate that ABot-NO sets new state-of-the-art benchmarks across seven authoritative platforms, while real-world deployment of the agentic navigation system on quadrupedal hardware confirms its robustness, generalization and social compliance in dynamic environments.

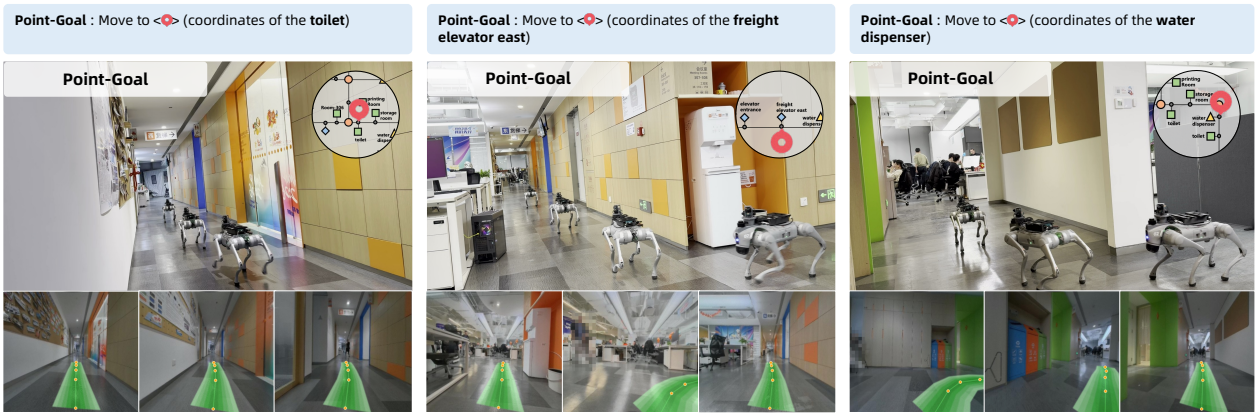

Figure 13 Visualization of Point-Goal in real-world deployment.

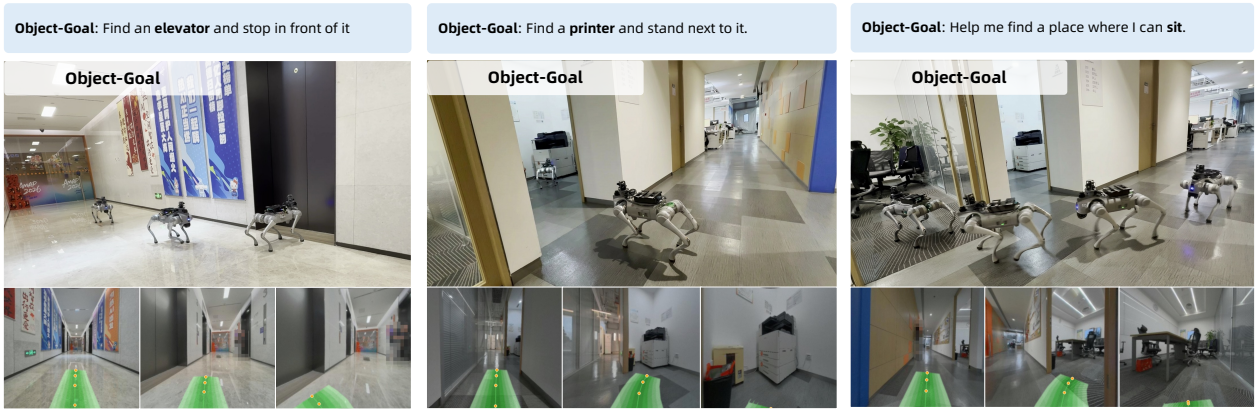

Figure 14 Visualization of Object-Goal in real-world deployment.

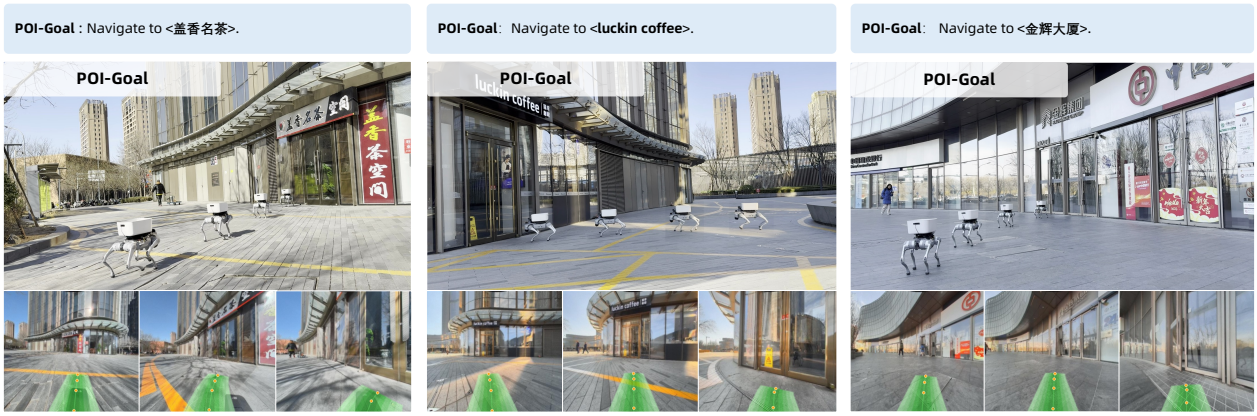

Figure 15 Visualization of POI-Goal in real-world deployment.

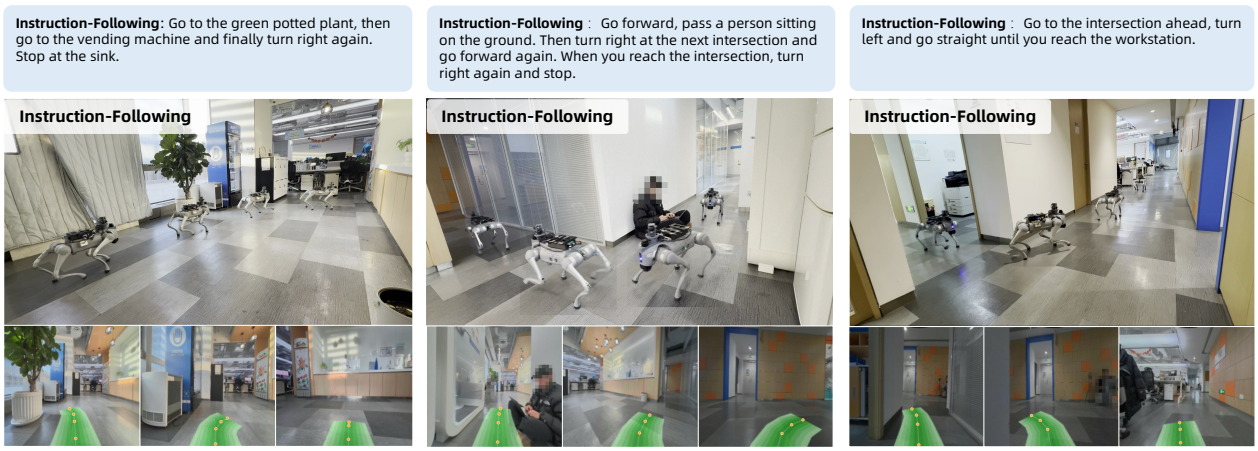

Figure 16 Visualization of Instruction-Following in real-world deployment.

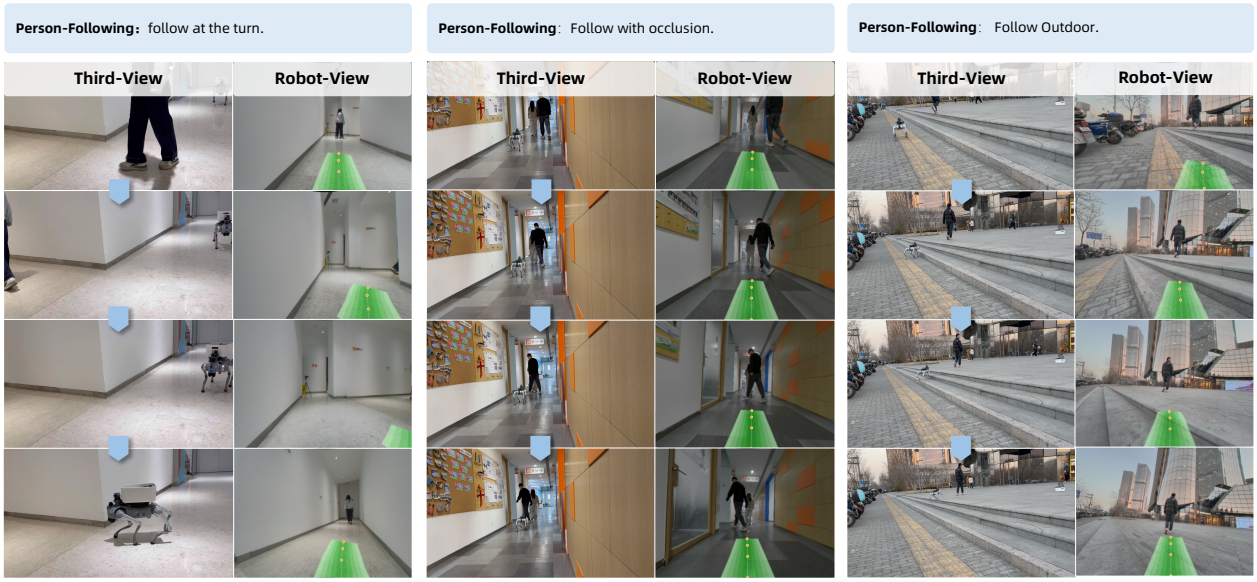

Figure 17 Visualization of Person-Following in real-world deployment.

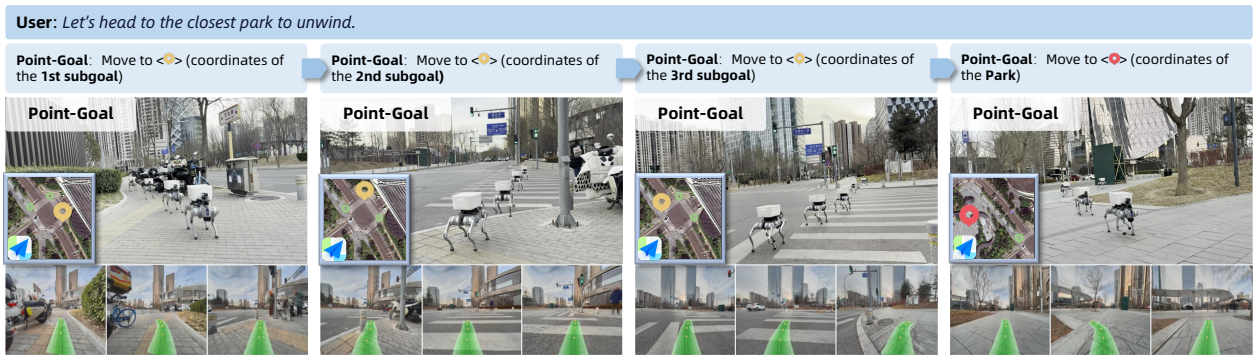

Figure 18 Visualization of a long-horizon outdoor navigation task. Given the user instruction "Let's head to the closest park to unwind" the Planner first interprets the user's intent and queries the Topo-Memory to retrieve the target coordinates. Subsequently, it decomposes the global route into a sequence of intermediate sub-goals. These sub-goals are then sequentially executed by the ABot-N0 using its Point-Goal primitive, successfully guiding the robot to the final destination.

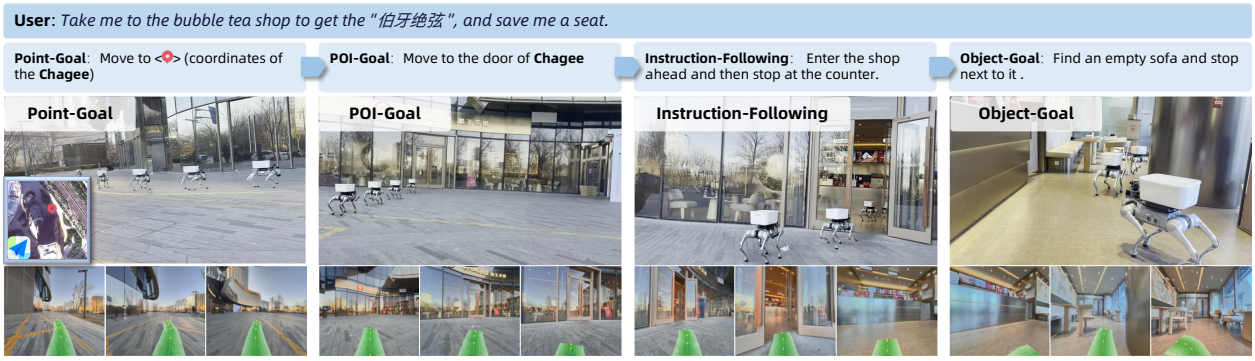

Figure 19 Visualization of a complex indoor-outdoor navigation task. Given the user instruction "Take me to the bubble tea shop to get 'Bo Ya Jue Xian' and save me a seat," the Planner first queries the Topo-Memory to identify the target POI. Subsequently, it decomposes the mission into a hierarchical sequence of sub-tasks: (1) Point-Goal to *Approach* the vicinity of the shop; (2) POI-Goal to precisely *Reach* the storefront; (3) Instruction-Following to navigate through the door; and (4) Object-Goal to locally *Reach* an available sofa. The figure depicts the model's predicted trajectory and the corresponding visual observations.

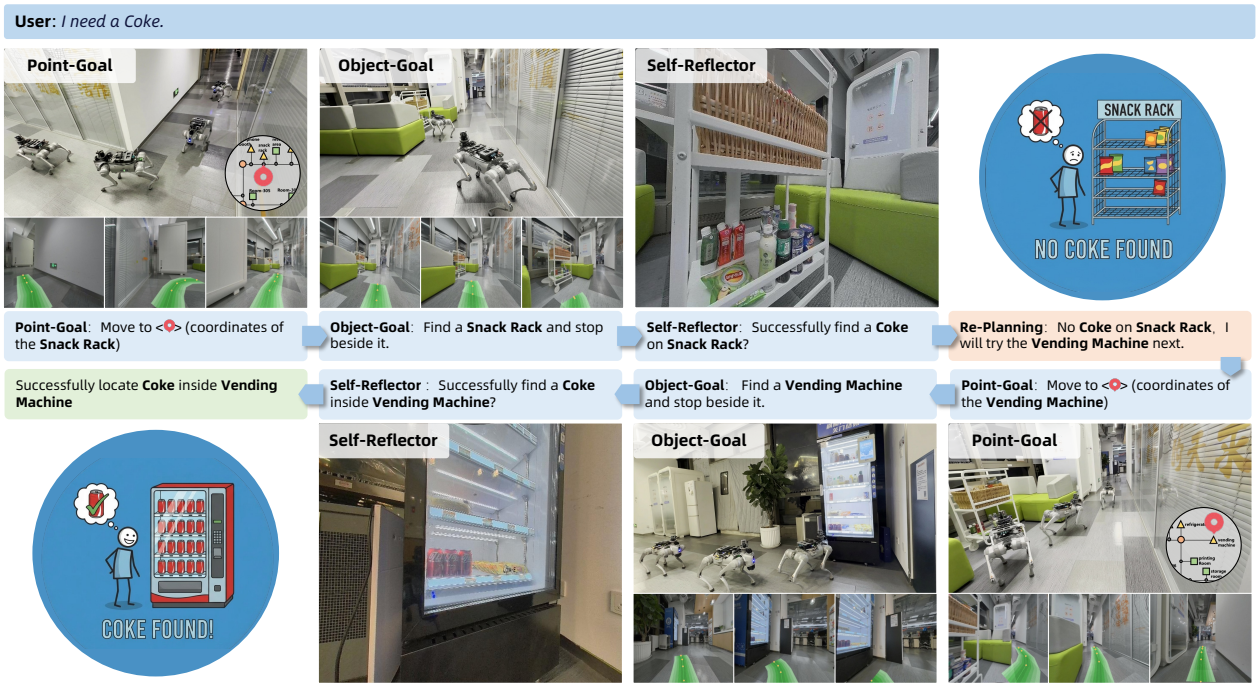

Figure 20 Visualization of the Self-Reflector mechanism and adaptive re-planning. Given the user instruction "I need a Coke," the Planner initially queries the Topo-Memory and generates a sequence to reach the *snack rack* (Point-Goal then Object-Goal). Upon execution, the Self-Reflector confirms the absence of the target, returning a failure status (r = False) with feedback f = "Coke not found on the snack rack." Leveraging this feedback, the system triggers a re-planning process to redirect the agent to the *vending machine* (Point-Goal then Object-Goal). Finally, the Self-Reflector successfully identifies the Coke in the frontal view (r = True) and terminates the task.

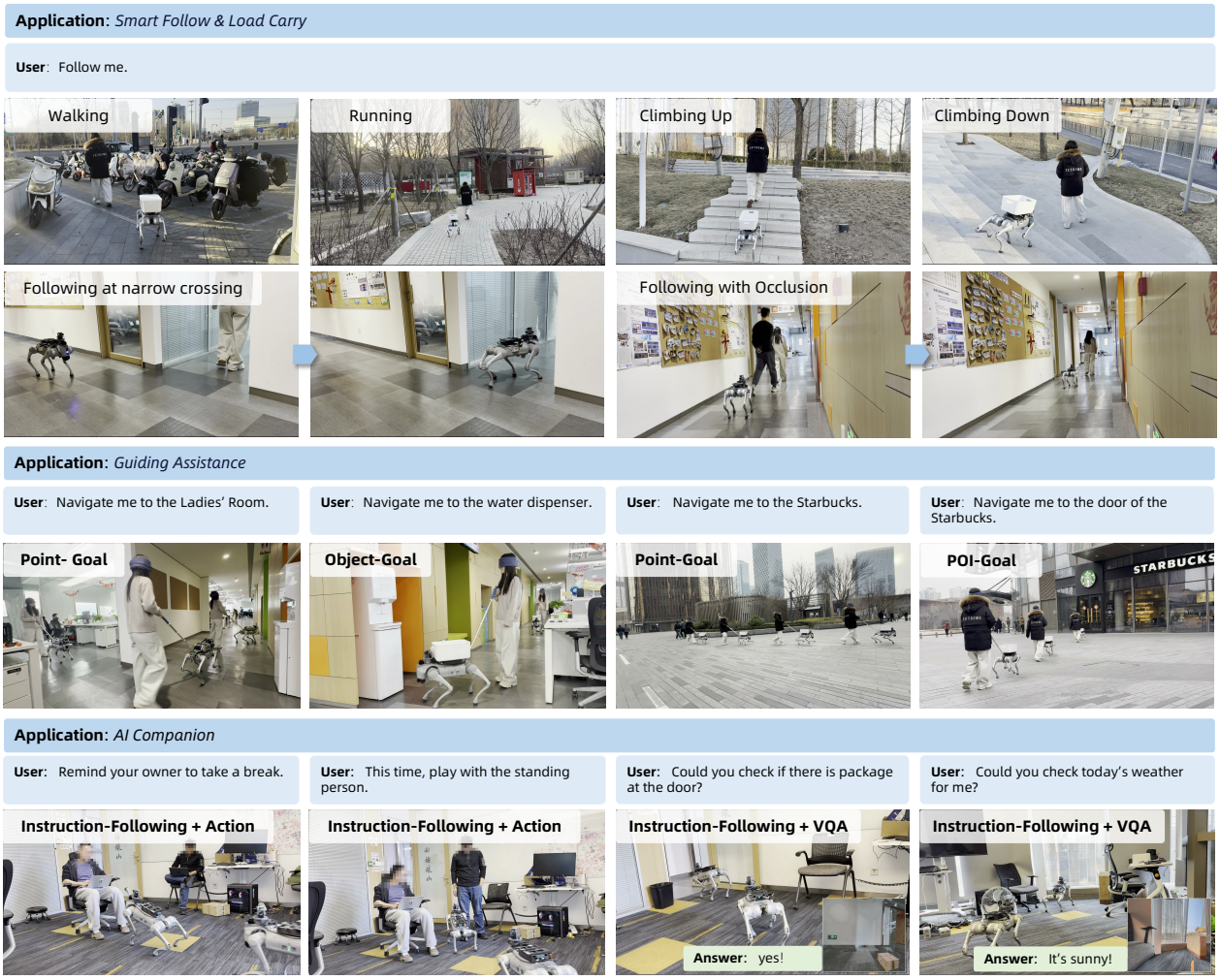

Figure 21 Real-world deployment across diverse downstream scenarios. We validate the system's versatility and robustness through three representative downstream applications: **Smart Follow & Load Carry**, **Guiding Assistance** and **AI Companion** in both indoor and outdoor environments. The visualizations demonstrate ABot-N0 adaptability to varying operational requirements.

8 Contributions and Acknowledgments

Paper Writing

- Zedong Chu†
- Shichao Xie†
- Xiaolong Wu

Research

- Xiaolong Wu
- Zedong Chu
- Shichao Xie
- Yanfen Shen
- Minghua Luo
- Zhengbo Wang
- Fei Liu
- Xiaoxu Leng
- Junjun Hu
- Mingyang Yin
- Jia Lu
- Yingnan Guo
- Kai Yang
- Jiawei Han
- Xu Chen
- Yanqing Zhu
- Yuxiang Zhao
- Xin Liu
- Yirong Yang
- Ye He
- Jiahang Wang
- Yang Cai
- Tianlin Zhang
- Li Gao
- Liu Liu
- Mingchao Sun
- Fan Jiang
- Chiyu Wang
- Zhicheng Liu
- Hongyu Pan

Data

- Honglin Han
- Zhining Gu
- Kuan Yang
- Jianfang Zhang
- Di Jing

Engineering

- Honglin Han
- Zihao Guan
- Wei Guo
- Guoqing Liu
- Di Yang
- Xiangpo Yang

Hardware

- Menglin Yang
- Hongguang Xing
- Weiguo Li

Project Leader

- Mu Xu
- Xiaolong Wu

Acknowledgments

We extend our sincere gratitude to the broader team for their support, particularly Ziqiao Li, Zixiao Tang and Jingjing Ma.

† Corresponding authors.

References

[1] Dong An, Zun Wang, Yangguang Li, Yi Wang, Yicong Hong, Yan Huang, Liang Wang, and Jing Shao. 1st place solutions for rxr-habitat vision-and-language navigation competition (cvpr 2022). arXiv preprint arXiv:2206.11610, 2022.

[2] Dong An, Hanqing Wang, Wenguan Wang, Zun Wang, Yan Huang, Keji He, and Liang Wang. Etpnav: Evolving topological planning for vision-language navigation in continuous environments. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.

[3] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 19129–19139, 2022.

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966, 2023.

[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] Jiaqi Chen, Bingqian Lin, Xinmin Liu, Lin Ma, Xiaodan Liang, and Kwan-Yee K Wong. Affordances-oriented planning using foundation models for continuous vision-language navigation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, pages 23568–23576, 2025.

[7] Jiuhai Chen, Le Xue, Zhiyang Xu, Xichen Pan, Shusheng Yang, Can Qin, An Yan, Honglu Zhou, Zeyuan Chen, Lifu Huang, et al. Blip3o-next: Next frontier of native image generation. arXiv preprint arXiv:2510.15857, 2025.

[8] Kevin Chen, Junshen K Chen, Jo Chuang, Marynel Vázquez, and Silvio Savarese. Topological planning with transformers for vision-and-language navigation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11276–11286, 2021.

[9] Peihao Chen, Dongyu Ji, Kunyang Lin, Runhao Zeng, Thomas Li, Mingkui Tan, and Chuang Gan. Weakly-supervised multi-granularity map learning for vision-and-language navigation. Advances in Neural Information Processing Systems, 35:38149–38161, 2022.

[10] Ziyi Chen, Yingnan Guo, Zedong Chu, Minghua Luo, Yanfen Shen, Mingchao Sun, Junjun Hu, Shichao Xie, Kuan Yang, Pei Shi, et al. Socialnav: Training human-inspired foundation model for socially-aware embodied navigation. arXiv preprint arXiv:2511.21135, 2025.

[11] 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.

[12] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

[13] Georgios Georgakis, Karl Schmeckpeper, Karan Wanchoo, Soham Dan, Eleni Miltsakaki, Dan Roth, and Kostas Daniilidis. Cross-modal map learning for vision and language navigation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 15460–15470, 2022.

[14] Google. Gemini api docs: Models (Gemini 3 Pro). URL https://ai.google.dev/gemini-api/docs/models?hl=zh-cn#gemini-3-pro. Online documentation.

[15] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948, 2025.

[16] Jiawei Guo, Tianyu Zheng, Yizhi Li, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Graham Neubig, Wenhu Chen, and Xiang Yue. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13869–13920, 2025.

[17] Meenakshi Gupta, Swagat Kumar, Laxmidhar Behera, and Venkatesh K Subramanian. A novel vision-based tracking algorithm for a human-following mobile robot. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 47(7):1415–1427, 2016.

[18] Noriaki Hirose, Dhruv Shah, Ajay Sridhar, and Sergey Levine. Sacson: Scalable autonomous control for social navigation. IEEE Robotics and Automation Letters, 9(1):49–56, 2023.

[19] Yicong Hong, Zun Wang, Qi Wu, and Stephen Gould. Bridging the gap between learning in discrete and continuous environments for vision-and-language navigation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 15439–15449, 2022.

[20] Junjun Hu, Jintao Chen, Haochen Bai, Minghua Luo, Shichao Xie, Ziyi Chen, Fei Liu, Zedong Chu, Xinda Xue, Botao Ren, Xiaolong Wu, Mu Xu, and Shanghang Zhang. Astranav-world: World model for foresight control and consistency, 2025. URL https://arxiv.org/abs/2512.21714.

[21] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. arXiv preprint arXiv:2410.21276, 2024.

[22] Haresh Karnan, Anirudh Nair, Xuesu Xiao, Garrett Warnell, Sören Pirk, Alexander Toshev, Justin Hart, Joydeep Biswas, and Peter Stone. Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation. IEEE Robotics and Automation Letters, 7(4):11807–11814, 2022.

[23] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 787–798, 2014.

[24] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph., 42(4):139–1, 2023.

[25] Jacob Krantz and Stefan Lee. Sim-2-sim transfer for vision-and-language navigation in continuous environments. In European conference on computer vision, pages 588–603. Springer, 2022.

[26] 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.

[27] Jacob Krantz, Erik Wijmans, Arjun Majundar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision and language navigation in continuous environments. In European Conference on Computer Vision (ECCV), 2020.

[28] Jacob Krantz, Aaron Gokaslan, Dhruv Batra, Stefan Lee, and Oleksandr Maksymets. Waypoint models for instruction-guided navigation in continuous environments. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 15162–15171, 2021.

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

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In European conference on computer vision, pages 740–755. Springer, 2014.

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

[32] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. arXiv preprint arXiv:2210.02747, 2022.

[33] Fei Liu, Shichao Xie, Minghua Luo, Zedong Chu, Junjun Hu, Xiaolong Wu, and Mu Xu. Navforesee: A unified vision-language world model for hierarchical planning and dual-horizon navigation prediction. arXiv preprint arXiv:2512.01550, 2025.

[34] Jiahang Liu, Yunpeng Qi, Jiazhao Zhang, Minghan Li, Shaoan Wang, Kui Wu, Hanjing Ye, Hong Zhang, Zhibo Chen, Fangwei Zhong, et al. Trackvla++: Unleashing reasoning and memory capabilities in vla models for embodied visual tracking. arXiv preprint arXiv:2510.07134, 2025.

[35] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In European conference on computer vision, pages 38–55. Springer, 2024.

[36] Xinhao Liu, Jintong Li, Yicheng Jiang, Niranjan Sujay, Zhicheng Yang, Juexiao Zhang, John Abanes, Jing Zhang, and Chen Feng. Citywalker: Learning embodied urban navigation from web-scale videos. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 6875–6885, 2025.

[37] Yuxing Long, Wenzhe Cai, Hongcheng Wang, Guanqi Zhan, and Hao Dong. Instructnav: Zero-shot system for generic instruction navigation in unexplored environment. arXiv preprint arXiv:2406.04882, 2024.

[38] Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. arXiv preprint arXiv:2109.08238, 2021.

[39] Sonia Raychaudhuri, Saim Wani, Shivansh Patel, Unnat Jain, and Angel Chang. Language-aligned waypoint (law) supervision for vision-and-language navigation in continuous environments. In Proceedings of the 2021 conference on empirical methods in natural language processing, pages 4018–4028, 2021.

[40] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In Proceedings of the IEEE/CVF international conference on computer vision, pages 9339–9347, 2019.

[41] Dhruv Shah, Benjamin Eysenbach, Nicholas Rhinehart, and Sergey Levine. Rapid exploration for open-world navigation with latent goal models. In Conference on Robot Learning, pages 674–684. PMLR, 2022.

[42] Dhruv Shah, Ajay Sridhar, Arjun Bhorkar, Noriaki Hirose, and Sergey Levine. Gnm: A general navigation model to drive any robot. In ICRA, pages 7226–7233. IEEE, 2023.

[43] Dhruv Shah, Ajay Sridhar, Nitish Dashora, Kyle Stachowicz, Kevin Black, Noriaki Hirose, and Sergey Levine. ViNT: A foundation model for visual navigation. In CoRL, 2023. URL https://openreview.net/forum?id=-K7-1WvKO3F.

[44] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In Proceedings of the IEEE/CVF international conference on computer vision, pages 8430–8439, 2019.

[45] Manycore Tech Inc. SpatialVerse Research Team. Interiorgs: A 3d gaussian splatting dataset of semantically labeled indoor scenes. https://huggingface.co/datasets/spatialverse/InteriorGS, 2025.

[46] Ajay Sridhar, Dhruv Shah, Catherine Glossop, and Sergey Levine. Nomad: Goal masked diffusion policies for navigation and exploration. In 2024 IEEE International Conference on Robotics and Automation (ICRA), pages 63–70. IEEE, 2024.

[47] Hao Tan, Licheng Yu, and Mohit Bansal. Learning to navigate unseen environments: Back translation with environmental dropout. arXiv preprint arXiv:1904.04195, 2019.

[48] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314, 2025.

[49] Hanqing Wang, Wei Liang, Luc Van Gool, and Wenguan Wang. Dreamwalker: Mental planning for continuous vision-language navigation. In Proceedings of the IEEE/CVF international conference on computer vision, pages 10873–10883, 2023.

[50] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 5261–5271, 2025.

[51] 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.

[52] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Scalable permutation-equivariant visual geometry learning. arXiv e-prints, pages arXiv–2507, 2025.

[53] Zihan Wang, Xiangyang Li, Jiahao Yang, Yeqi Liu, and Shuqiang Jiang. Gridmm: Grid memory map for vision-and-language navigation. In Proceedings of the IEEE/CVF International conference on computer vision, pages 15625–15636, 2023.

[54] Zihan Wang, Xiangyang Li, Jiahao Yang, Yeqi Liu, Junjie Hu, Ming Jiang, and Shuqiang Jiang. Lookahead exploration with neural radiance representation for continuous vision-language navigation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 13753–13762, 2024.

[55] 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.

[56] 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.

[57] Wentao Xiang, Haokang Zhang, Tianhang Yang, Zedong Chu, Ruihang Chu, Shichao Xie, Yujian Yuan, Jian Sun, Zhining Gu, Junjie Wang, Xiaolong Wu, Mu Xu, and Yujiu Yang. Nav-r^2 dual-relation reasoning for generalizable open-vocabulary object-goal navigation, 2025. URL https://arxiv.org/abs/2512.02400.

[58] 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.

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

[60] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. arXiv preprint arXiv:2310.11441, 2023.

[61] Kai Yang, Tianlin Zhang, Zhengbo Wang, Zedong Chu, Xiaolong Wu, Yang Cai, and Mu Xu. Ce-nav: Flow-guided reinforcement refinement for cross-embodiment local navigation. arXiv preprint arXiv:2509.23203, 2025.

[62] Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. Vlfm: Vision-language frontier maps for zero-shot semantic navigation. In 2024 IEEE International Conference on Robotics and Automation (ICRA), pages 42–48. IEEE, 2024.

[63] 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.

[64] 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.

[65] 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.

[66] Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and He Wang. Navid: Video-based vlm plans the next step for vision-and-language navigation. arXiv preprint arXiv:2402.15852, 2024.

[67] Jiazhao Zhang, Anqi Li, Yunpeng Qi, Minghan Li, Jiahang Liu, Shaoan Wang, Haoran Liu, Gengze Zhou, Yuze Wu, Xingxing Li, Yuxin Fan, Wenjun Li, Zhibo Chen, Fei Gao, Qi Wu, Zhizheng Zhang, and He Wang. Embodied navigation foundation model, 2025. URL https://arxiv.org/abs/2509.12129.

[68] Yuxiang Zhao, Yirong Yang, Yanqing Zhu, Yanfen Shen, Chiyu Wang, Zhining Gu, Pei Shi, Wei Guo, and Mu Xu. Bridging the indoor-outdoor gap: Vision-centric instruction-guided embodied navigation for the last meters, 2026. URL https://arxiv.org/abs/2602.06427.

[69] Fangwei Zhong, Kui Wu, Hai Ci, Churan Wang, and Hao Chen. Empowering embodied visual tracking with visual foundation models and offline rl. In European Conference on Computer Vision, pages 139–155. Springer, 2024.

[70] 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.