InternVLA-N1: An Open Dual-System Vision-Language Navigation Foundation Model with Learned Latent Plans

Intern Robotics, Shanghai AI Laboratory

Abstract: We introduce InternVLA-N1, the first open dual-system vision-language navigation foundation model. Unlike previous navigation foundation models that can only take short-term actions from a limited discrete space, InternVLA-N1 decouples the task as pixel-goal planning with System 2 and agile execution with System 1. A curriculum two-stage training paradigm is devised for this framework: First, two systems are pretrained with explicit pixel goals as supervision or condition. Subsequently, we freeze System 2 and finetune the newly added latent plans with System 1 in an asynchronous end-to-end manner. Such a paradigm relying on latent plans as the intermediate representation removes the ambiguity of pixel goal planning and provides new potentials for pretraining extensions with video prediction. To enable scalable training, we develop an efficient navigation data generation pipeline in simulation and introduce InternData-N1, the largest navigation dataset to date. InternData-N1 comprises over 50 million egocentric images collected from more than 3,000 scenes, amounting to 4,839 kilometers of robot navigation experience. We evaluate InternVLA-N1 across 6 challenging navigation benchmarks, where it consistently achieves state-of-the-art performance, with improvements ranging from 3% to 28%. In particular, although *only trained with simulation data*, it can be zero-shot generalized across diverse embodiments (wheeled, quadruped, humanoid) and in-the-wild environments, and demonstrates synergistic integration of long-horizon planning (>150m) and real-time decision-making (>30Hz) capabilities in the real world. All code, models, and datasets are publicly available.

Links: Code:InternNav | Model:InternVLA-N1 | Data:InternData-N1 | Homepage

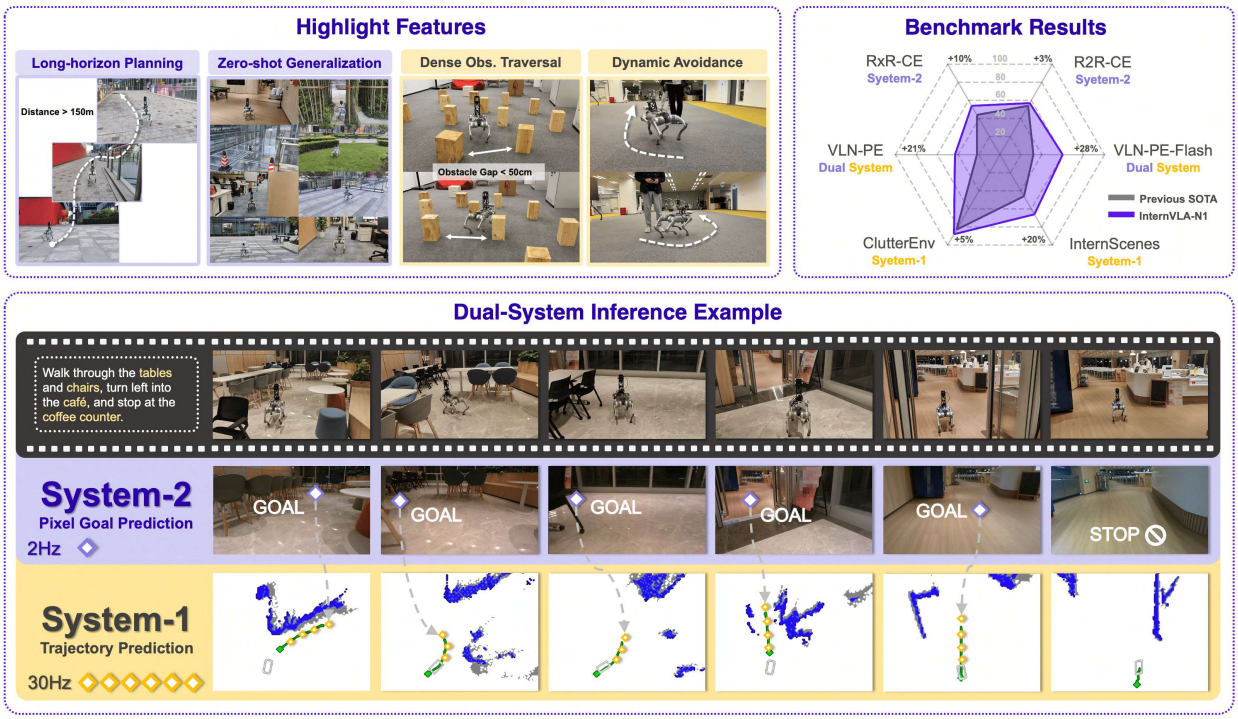

Figure 1 | Highlight features and inference examples of InternVLA-N1.

Contributors are listed in Appendix A.

1. Introduction

Navigation is a fundamental task in robotics. In practice, the navigation system typically takes language instructions with visual observations as input and executes the planned trajectory accordingly. Recent years have witnessed the progress of this field, from the exploration of building benchmarks based on discrete goal planning Anderson et al. (2018a); Ku et al. (2020) to continuous action space Krantz et al. (2020b) and physically realistic simulation with locomotion controllers Cheng et al. (2025); Wang et al. (2025b). On the other hand, multi-modal LLMs provide new potential to train such models in simulation while generalizing to the open real world, given their strong prior knowledge. The research community is showing growing interest Cheng et al. (2025); Wei et al. (2025); Zhang et al. (2025a); Zheng et al. (2024) and has made successful primary attempts along this direction, demonstrating them on diverse embodiments, including quadruped robots and humanoids.

However, although these models were developed on continuous environment benchmarks such as VLN-CE Krantz et al. (2020b), their action space is simplified into discrete choices and predicted in an end-to-end manner. As a result, they can only take short-term action steps from the limited space and struggle with the inference speed as well as fragmented navigation behavior. Intuitively, in contrast to such hard mapping from visual observations and language instruction to direct action output, a more native target type should be the midterm goals, especially on the image pixels, which indicate where the robot should go and can be aligned with the visual grounding capability of multi-modal LLMs. Meanwhile, another high-frequency local planner is devised to execute the path planning towards the midterm goal with the agility to avoid dynamic obstacles. Ultimately, the overall framework performs with a mechanism that is similar to the human cognitive theory Kahneman (2011) “System 1 Execution & System 2 Thinking". There have been several attempts at such ideas on building VLA models, such as Helix FigureAI (2025), GR00T Bjorck et al. (2025), Hi Robot Shi et al. (2025), and OneTwoVLA Lin et al. (2025b).

This paper presents InternVLA-N1, the first open dual-system vision-language navigation foundation model that incorporates learned latent plans as intermediate representation. Unlike planning in fully observable settings such as tabletop manipulation, System 2 in InternVLA-N1 is required to perform multi-round, precise planning based on language instructions, under conditions of partial observability and mobile exteroceptive perspectives. Meanwhile, System 1 is responsible for executing these plans in real-world environments, robustly handling dynamic disturbances such as pedestrians.

To address these challenges, we formulate System 2 as a pixel goal planner, leveraging multi-modal LLMs as the backbone to exploit their inherent commonsense knowledge and multi-modal perception capabilities. We define the pixel goal as the preferred navigation waypoint projected on the 2D image plane. Complementarily, System 1 is designed as a lightweight, diffusion-based visual navigation policy capable of real-time path planning conditioned on the goal generated by System 2. Both systems are first pre-trained to form fundamental navigation abilities. System 2 is trained to align the pixel grounding ability to the VLN domain, and System 1 is trained to condition on explicit goals, including the pixel goal coordinates, and generate collision-free navigation paths towards the goal.

Although the two systems can be cascaded into a complete VLN framework after pre-training, this design introduces several critical challenges. First, synchronizing System 2 planning with System 1 execution significantly increases overall latency, as System 1 must wait for responses from the multi-modal LLM. This delay compromises the system’s ability to respond effectively in real-time, reducing its feasibility for dynamic environments. Second, representing navigation goals using 2D pixel coordinates leads to ambiguity, often resulting in suboptimal or confused behavior by System 1.

To address these issues, we introduce an additional fine-tuning phase that enables asynchronous inference and enhances the spatial representation of the intermediate goal interface between two

systems. Specifically, during fine-tuning, System 1 continuously receives the latest observations, while System 2 operates on delayed inputs. This setup encourages System 1 to estimate goal completion dynamically and adapt to the asynchronous execution pace. Furthermore, we replace explicit pixel goals with learnable latent tokens, enabling potentially more informative implicit planning references through joint tuning. To enhance and validate the latent representations, we train a latent plan-based world model to predict the subsequent egocentric observation sequences as an extension. Experiments demonstrate that our world model can imagine consistent and high-quality egocentric observation sequences towards the planned goals. The video prediction objective facilitates the extraction of spatial information from the latent tokens and accelerates the efficiency of the joint-tuning process, also leading to a scalable training paradigm with real-world video data.

To support the aforementioned pre-training and joint fine-tuning, we develop a highly efficient simulation data generation pipeline capable of producing 50K navigation trajectories per day on a single machine. Integrated with automatic instruction labeling and data filtering processes, this pipeline enables the construction of a large-scale navigation dataset, InternData-N1, comprising over 53 million egocentric image observations and 800K language instructions across more than 3,000 indoor scenes. This corresponds to approximately 4,839 kilometers of robot navigation experience.

Experimental results demonstrate that InternVLA-N1 consistently outperforms previous state-of-the-art methods across six challenging benchmarks, achieving performance gains ranging from 3% to 28%. Moreover, real-world evaluations demonstrate strong long-horizon planning ability (>150m) and real-time decision-making ability (>30Hz) across multiple robot platforms in diverse scenarios, underscoring the adaptability in the dynamic open world.

2. Related Work

Vision-Language Navigation. Vision-Language Navigation (VLN) is a long-horizon instruction-following task that requires precise planning and following by robots. Early approaches simplify the problem by adopting a discrete setting Anderson et al. (2018a); Ku et al. (2020); Qi et al. (2020), in which the agent is teleported between predefined nodes in a navigation graph. This abstraction bypasses key real-world challenges such as obstacle avoidance and path planning. To better approximate real-world conditions, Vision-Language Navigation in Continuous Environments (VLN-CE) Krantz et al. (2020b); Savva et al. (2019) has been introduced, where the agent operates using low-level discrete control actions. A number of methods have since been proposed An et al. (2022, 2023); Hong et al. (2022); Irshad et al. (2022); Krantz and Lee (2022); Krantz et al. (2021); Raychaudhuri et al. (2021); Wang et al. (2023b), steadily improving navigation accuracy within simulated environments. However, the reliance on task-specific network architectures and limited training data continues to hinder zero-shot generalization and sim-to-real transfer. To address these limitations, recent agentic approaches Chen et al. (2024, 2025); Lin et al. (2025a); Long et al. (2024c,d); Qiao et al. (2024); Zhang et al. (2025b); Zhou et al. (2023, 2024) leverage general-purpose foundation models, demonstrating improved performance and robustness in real-world VLN tasks. However, without access to diverse data for downstream task fine-tuning, such general-purpose foundation models remain poorly aligned with the navigation domain. To address this, we propose an efficient data generation pipeline in simulation, resulting in a high-quality dataset, InternData-N1. Combined with an advanced training recipe and a refined network architecture, our model, InternVLA-N1, achieves state-of-the-art performance on multiple VLN benchmarks and demonstrates strong zero-shot generalization to real-world.

Visual Navigation Policy Learning. The visual navigation skill is responsible for reaching explicit goals and performing real-time obstacle avoidance. Traditional modular approaches Fox et al. (1997); Karaman and Frazzoli (2011); Kramer and Stachniss (2012); Williams et al. (2015); Zhou et al.

(2020) rely on explicit localization and mapping to accomplish navigation tasks. However, these systems often suffer from compounding errors and latency introduced by cascaded modules, and typically require extensive hyperparameter tuning to adapt to different robotic platforms. To address these challenges, recent work has explored end-to-end learning-based approaches. For instance, GNM Shah et al. (2023a), X-Nav Wang et al. (2025a), RING Eftekhar et al. (2024), and X-Mobility Liu et al. (2024) focus on improving zero-shot policy generalization across different embodiments. Other methods such as iPlanner Yang et al. (2023), ViPlanner Roth et al. (2024), FDM Roth et al. (2025), and S2E He et al. (2025) investigate efficient training paradigms and enhance sim-to-real transfer in point-goal navigation. Meanwhile, approaches like SLING Wasserman et al. (2023), ViNT Shah et al. (2023b), NoMad Sridhar et al. (2024), and NaviDiffuser Zeng et al. (2025) focus on image-goal navigation. Our model incorporates strong pretraining for both components of its dual-system architecture. Notably, the pretrained System 1 represents the first sim-to-real visual navigation policy that supports no-goal exploration, point-goal and image-goal navigation within a unified framework.

Vision-Language-Action Model for Navigation. Recent studies increasingly leverage multi-modal large models as pretrained backbones for navigation tasks, with the goal of utilizing the commonsense knowledge inherent in backbones to enhance the navigation performance. A common approach is to formulate navigation actions as texts, thereby unifying the task as a next-token prediction problem within large language models (LLMs). For example, a line of work Gao et al. (2025); Wang et al. (2025c); Wei et al. (2025); Zhang et al. (2024, 2025a); Zheng et al. (2024) adopts the discrete action space same as VLN-CE and defines the corresponding vocabulary list, using it as the response labels for LLMs. In contrast, RoboPoint Yuan et al. (2025) and NaviMaster Luo et al. (2025) circumvent the limitations of discrete action spaces by framing navigation as a pixel grounding task. However, the action execution still requires additional modules, such as camera calibration and a point-goal navigation policy. Recent methods such as UniVLA Bu et al. (2025) and TrackVLA Wang et al. (2025d) adopt an end-to-end paradigm, directly mapping latent features extracted from large language models (LLMs) to continuous trajectories executable by the robot. However, these approaches typically rely on a synchronized framework, which limits their ability to make high-frequency decisions to deal with the dynamic open world. Although recent efforts have explored slow-fast dual-system architectures Bu et al. (2024); FigureAI (2025); Shi et al. (2025), these approaches primarily target tabletop manipulation tasks, leaving the challenges of long-context memory modeling and exploration in unknown scenarios unaddressed. Our proposed InternVLA-N1 is the first asynchronous dual-system architecture capable of long-horizon instruction following, accurate planning, and cross-building navigation in unseen environments.

3. InternData-N1 Dataset

For the navigation task, most real-world datasets Hirose et al. (2018, 2023); Karnan et al. (2022); Shah et al. (2021) are constrained by scene diversity and scale. Meanwhile, Internet video datasets Lin et al. (2023); Liu et al. (2025) suffer from imprecise localization and mapping information, which limits their feasibility as reliable navigation datasets for trajectory prediction. In contrast, we propose three efficient pipelines for generating navigation datasets in simulation, aiming to facilitate scalable training. Specifically, the **InternData-N1** dataset comprises subsets **VLN-N1**, **VLN-CE** and **VLN-PE**, which owns complementary features:

- **VLN-N1** is collected from large-scale open-source 3D assets with extensive domain randomization to enhance generalization to diverse real-world scenes.
- **VLN-CE** offers high-quality, fine-grained instruction annotations, which improve performance on long-horizon downstream navigation tasks.
- **VLN-PE** incorporates low-level motion controllers within physics-based simulation, supporting effective sim-to-real transfer by modeling realistic robot dynamics during navigation.

Real-World Navigation Dataset							
Dataset	**Scene**	**Distance (Km)**	**Hour**	**Image**	**Instruction**	**Action**	**Collection**
GoStanford Hirose et al. (2018)	27	25.5	16.7	178K	N/A	Trajectory	Teleoperated
RECON Shah et al. (2021)	9	152.5	40	610K	N/A	Trajectory	Autonomous
SCAND Karnan et al. (2022)	1	40	8.7	100K	N/A	Trajectory	Teleoperated
SACSoN Hirose et al. (2023)	5	58	75	241K	N/A	Trajectory	Autonomous
Internet-Video Navigation Dataset							
Dataset	**Scene**	**Distance**	**Hour**	**Image**	**Instruction**	**Action**	**Collection**
Youtube-HT Chang et al. (2020)	1387	-	119	550K	N/A	Trajectory	Internet
Youtube-VLN Lin et al. (2023)	**4078**	-	433	587K	14K	Trajectory	Internet
Simulation Navigation Dataset							
Dataset	**Scene**	**Distance**	**Hour**	**Image**	**Instruction**	**Action**	**Collection**
AMR Meng et al. (2025)	54	-	N/A	7.5M	N/A	Trajectory	Autonomous
REVERIE Qi et al. (2020)	86	-	N/A	10.6K	21.7K	Discrete	Autonomous
Habitat-Web Ramrakhya et al. (2022)	81	-	N/A	19.5M	N/A	Discrete	Teleoperated
R2R-CE Hong et al. (2022)	61	103.9	N/A	647K	10.8K	Discrete	Autonomous
RxR-CE Anderson et al. (2018b)	59	303.3	N/A	1.9M	20K	Discrete	Autonomous
R2R-EnvDrop-CE Tan et al. (2019)	60	1630.5	N/A	146.2K	21.6K	Discrete	Autonomous
ScaleVLN Yu et al. (2023)	800	-	N/A	-	**4.9M**	Discrete	Autonomous
InternData-N1	3154	**4839.9**	**1344**	**53.5M**	0.8M	**Trajectory + Joint**	**Autonomous**

Table 1 | Comparison of InternData-N1 with other navigation datasets.

3.1. VLN-N1

Abundant open-source scene assets provides an ideal playground for generating indoor navigation trajectories. We use Replica Straub et al. (2019), Matterport3D Chang et al. (2017), Gibson Xia et al. (2018), 3D-Front Fu et al. (2021), HSSD Khanna et al. (2024) and HM3D Ramakrishnan et al. (2021) as the scene repository. To generate realistic navigation process with egocentric observations, we generate a batch of collision-free and smooth trajectories with a multi-stage path-planning process. We first build Euclidean Signed Distance Field (ESDF) for each floor based on the mesh structure, then the global path-planning contains three steps similar to the previous work Cai et al. (2025): (1) Initialize the global path for randomly sampled starting points and goals with A-star algorithm. (2) Trajectory waypoints optimization with ESDF map. (3) Trajectory smoothing. The collected trajectories are used for rendering RGB and depth observations in BlenderProc Denninger et al. (2020).

To generate both fine-grained or long-horizon task language instructions, we first extract key frames based on the trajectory geometry information, such as the corresponding frames when a sharp turn happens. Based on the extracted key frames, the entire trajectory is split into several sub-clips. Then, we deploy an open-source multi-modal large model LLaVa-OneVision Li et al. (2024) to generate fine-grained language instructions for every sub-clip. We find the linguistic style of the generated instructions is limited, therefore, we adapt another language model - Qwen3-72b Yang et al. (2025) to rewrite the language instructions for every clip and summarize all the sub-clips into one instruction for long-horizon task. Following the above pipeline illustrated in Figure 3, we represent a new large-scale navigation dataset VLN-N1. The dataset proportion details and the statistical metrics are shown in Figure 2.

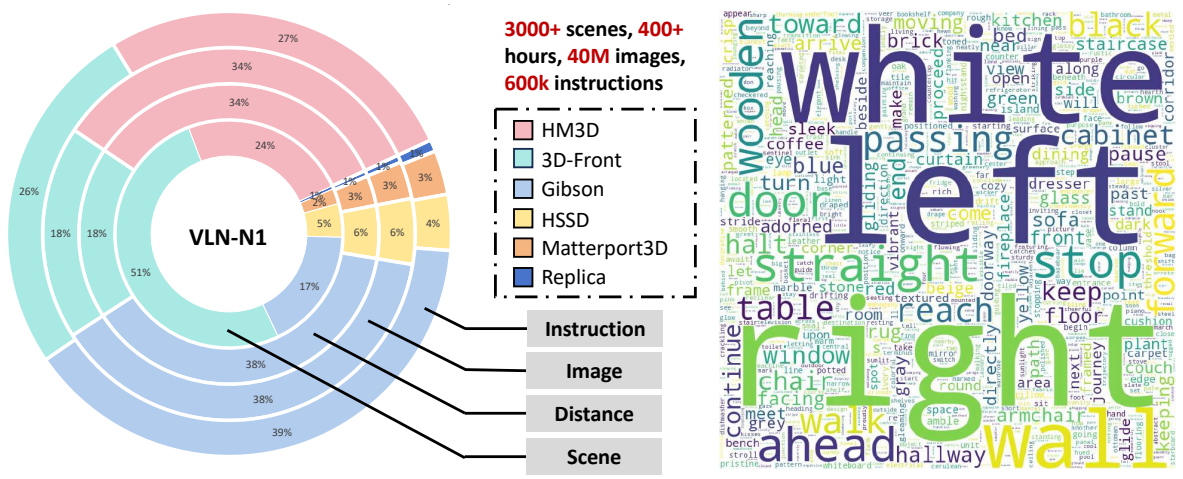

Figure 2 | An overview of VLN-N1 dataset. The left figure shows the dataset proportion while the right demonstrates keywords in the annotated instructions.

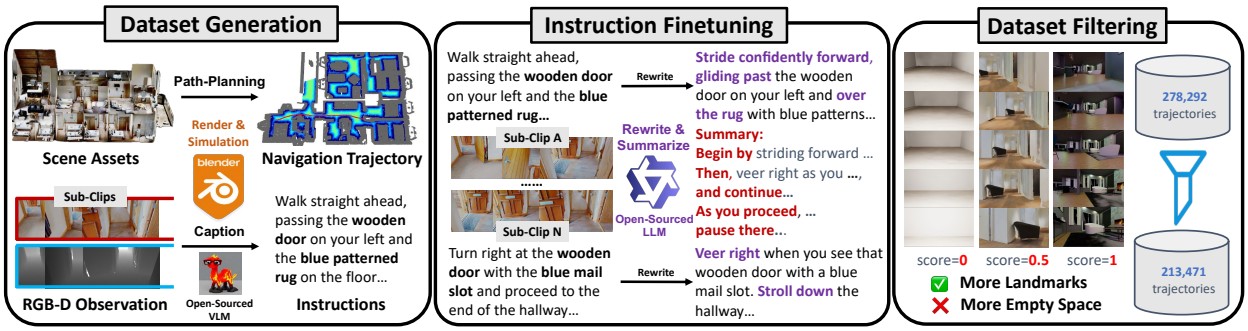

Figure 3 | Data processing pipeline of the VLN-N1 dataset.

3.2. VLN-CE

The VLN-CE dataset is derived from established Vision-and-Language Navigation benchmarks, including VLN-CE Krantz et al. (2020b), EnvDrop Tan et al. (2019) and ScaleVLN Wang et al. (2023a), which are designed for training general-purpose indoor navigation models. Using the Habitat simulator Szot et al. (2021), we render scenes from Matterport3D Chang et al. (2017) and HM3D Ramakrishnan et al. (2021), then replay the episodes to collect our dataset. Specifically, we utilize the built-in `ShortestPathFollower` agent in Habitat to generate trajectories by following predefined reference paths, with each path corresponding to an aligned fine-grained natural language instruction. The action space adheres to Habitat's default VLN task configuration, comprising four discrete actions: `MOVE_FORWARD` (0.25m), `TURN_LEFT` (15°), `TURN_RIGHT` (15°), and `STOP`. For each episode, we recorded RGB observations paired with their corresponding action sequences. In total, we collected 332,179 episodes spanning 856 unique scenes across both Matterport3D and HM3D datasets. To make the dataset applicable for training the System 2, we segment the raw trajectories into multiple clips and project the agent's position onto the 2-D image plane to serve as pixel goal labels. Further details can be found in Section 4.2.

3.3. VLN-PE

The VLN-PE dataset is designed to bridge the sim-to-real gap in the Vision-and-Language Navigation (VLN) task by collecting data that reflects realistic robot motion within the physical simulation platform InternUtopia Wang et al. (2024a). Unlike the previous VLN-N1 and VLN-CE, VLN-PE

explicitly incorporates both robot embodiment and locomotion policy into its data collection process. We employ a diverse set of robotic platforms, including quadruped (Unitree AlienGo), humanoid (Unitree H1 and G1), and wheeled (Jetbot) robots, and use existing learning-based locomotion controllers Long et al. (2024a,b); Pan et al. (2025) to govern their movement. Each robot is tasked with following a predefined navigation path aligned with a natural language instruction, resulting in corresponding egocentric observations. The language instructions and paths are primarily sourced from the R2R dataset Anderson et al. (2018a), with modifications. Specifically, we exclude episodes that involve stair traversal (i.e., going upstairs or downstairs), which current locomotion policies cannot robustly handle. The final VLN-PE dataset comprises 8,679 episodes across 61 scenes from the Matterport3D dataset Chang et al. (2017).

4. Approach

4.1. Overview

As illustrated in Figure 4, InternVLA-N1 adopts a compositional architecture featuring a dual-system design that synergistically combines high-level instruction interpretation with low-level action execution. Specifically, our system integrates:

- **System 2:** A vision-language model (VLM)-based planning module that interprets navigation instructions to predict mid-term waypoint goals through image-grounded reasoning. By predicting pixel coordinates in the image space, it effectively connects instruction understanding with spatial reasoning, enabling long-horizon navigation instruction following.
- **System 1:** A multi-modal goal-conditioned diffusion policy guided by latent plan or supported explicit goals, which generates executable short-horizon trajectories conditioned on current observations and the asynchronous latent features from System 2. It enables robust, real-time control and local decision-making in complex environments.

To fully unlock open-world generalization and asynchronous inference capabilities in the dual-system architecture, we design a curriculum training scheme. Initially, each system is trained separately to acquire basic navigation skills using explicit goals in a synchronized setting. Then, a joint fine-tuning phase is introduced. In this phase, we incorporate learnable tokens into System 2 as implicit midterm goals to reduce the ambiguity of pixel-based targets. Additionally, System 2 is fed delayed observations, which forces System 1 to adapt to asynchronous execution. Further technical details are provided in the following section.

4.2. System2: Vision-Language Model based Planning via Pixel Grounding.

We build our goal planning module upon Qwen-VL-2.5 Bai et al. (2025), a strong open-source vision-language model capable of spatial grounding. Qwen-VL-2.5 consists of three main components: a vision encoder, a language model, and a lightweight multimodal connector for modality fusion. It supports grounding tasks by directly predicting pixel coordinates in response to spatial queries, making it particularly well-suited for tasks requiring fine-grained localization, such as referring expression comprehension and visual question answering.

To adapt Qwen-VL-2.5 for vision-and-language navigation (VLN), we formulate the high-level planning as a farthest pixel goal prediction problem. The model takes as inputs a sequence of egocentric images and the language instruction, and predicts a 2-D coordinates within the image that corresponds to the next preferred navigation waypoint. We fine-tune Qwen-VL-2.5 with the InternData-N1 VLN-CE subset. By measuring the visibility between the agent's position and the camera view, we divide each original VLN-CE trajectory into multiple farthest pixel prediction training

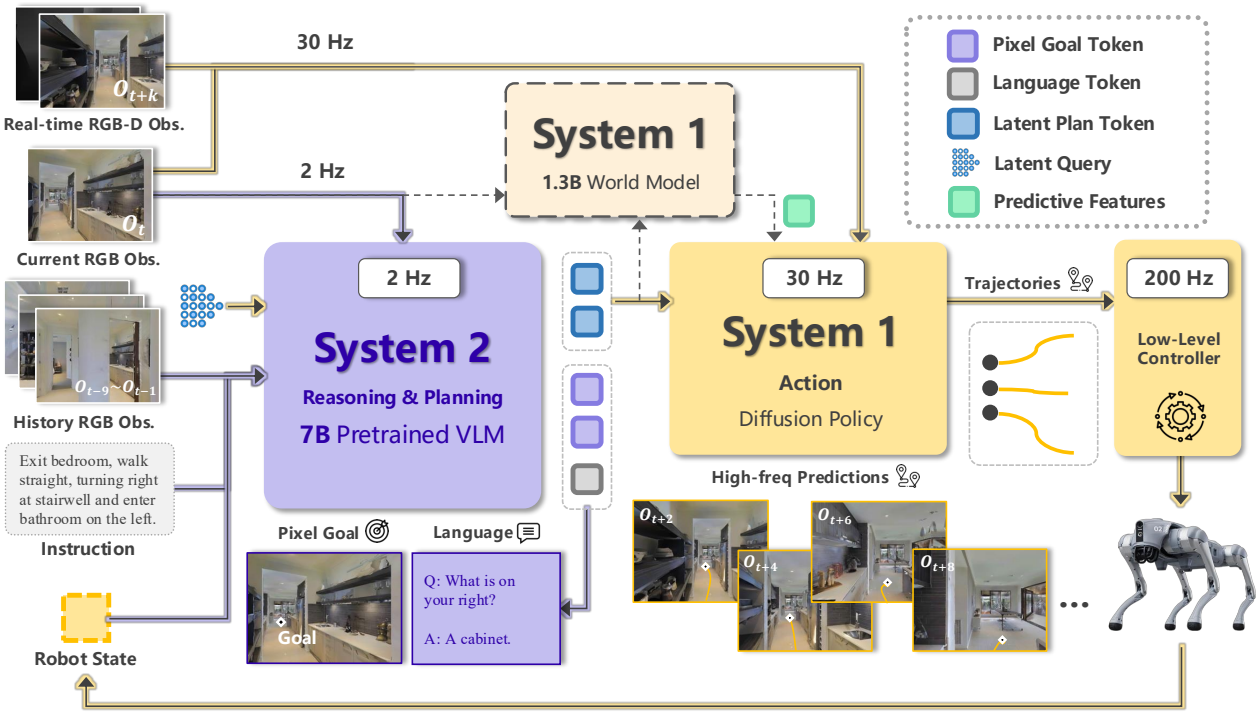

Figure 4 | An overview of InternVLA-N1 framework. System 2 perceives the long-horizon multi-modal inputs and translate into mid-term latent plans at 2 Hz, while System 1 processes the asynchronous latent plans along with short-term visual observations to enable real-time decision making.

samples, ultimately generating over 5 million samples for alignment with the navigation planning task. In addition, System 2 is responsible for deciding when to stop upon task completion and for performing on-the-spot rotations when no suitable navigation waypoints are detected in the image. Compared to direct action prediction, our approach offers a more efficient mechanism for bridging multi-modal understanding with spatial decision-making.

4.3. System 1: A Multi-Goal conditioned Diffusion Policy

Our System 1 model is a diffusion-based local navigation policy designed for real-time collision avoidance and path planning. It adopts a similar architecture to our previous work, NavDP Cai et al. (2025), which predicts both navigation trajectories and their corresponding safety scores for trajectory selection. To improve navigation performance across different types of goals, we introduce an explicit goal embedding alignment as an additional training objective.

In particular, we treat the point-goal as a general and unambiguous form of goal specification. Two auxiliary prediction heads are incorporated, which take image-goal and pixel-goal embeddings as inputs and are supervised using the point-goal as label. The goal alignment loss, combined with the action loss and critic loss, forms the overall training objective. By introducing the goal alignment objective, all types of navigation tasks are implicitly transformed into point-goal navigation tasks, thereby significantly reducing the learning complexity. The System 1 is trained with VLN-N1 subset.

4.4. Hierarchical Joint Training

Stage 1: Single-System Pre-training. The training process of System2 begins with a vision-language model (Qwen-VL-2.5 7B model) that has been pretrained on large-scale image-text corpora. We adapt this model for navigation-specific planning via task-adaptive supervised fine-tuning. Specifically, we

use paired trajectories consisting of navigation instructions, egocentric observations, and mid-term waypoints. In this setup, each mid-term waypoint is represented as a 2-D coordinate in the image pixel space of the current observation. During training, all components—including the vision encoder, the cross-modal connector, and the language model—are jointly optimized for one epoch with our curated SFT dataset. The model learns to interpret the instruction in context and predict the pixel-level goal location on the image that aligns with the intended navigation waypoint.

All components of the System 1 model are trained from scratch except for the DepthAnything Yang et al. (2024) RGB encoder. The System 1 model is trained with three major objectives: embedding alignment among different goals, noise prediction for the diffusion policy, and critic prediction. For embedding alignment, we add two auxiliary point-goal prediction tasks, with either the image-goal encoding or pixel-goal encoding as input. This helps the goal encoder trained from scratch to capture important representations for the navigation task. Concretely, denote the image-goal as $I_g \in \mathbb{R}^{C\times H\times W}$, the pixel-goal as $c_g = (u, v)$, the current RGB observation as $I_t \in \mathbb{R}^{C\times H\times W}$ and the point-goal as $p_g = (x, y, \theta)$. To encode the pixel-goal c_g, we first convert c_g into a image mask M_g with only local areas around (u, v) is set to one and leaving other pixels to zeros. Then, we use two ViT encoder training from scratch to fuse (I_g, I_t) and (M_g, I_t). The encoded embeddings $z_i = f_{img}(I_g, I_t)$ and $z_p = f_{pix}(M_g, I_t)$ are attached with an addition MLP layer to predict the estimated point-goal. Then, the goal alignment loss can be written as:

$$\mathcal{L}^{goal} = \frac{1}{N}\sum_{i=1}^{N}\left\|\text{MLP}(z_{img}) - p_g\right\|^2 + \frac{1}{N}\sum_{i=1}^{N}\left\|\text{MLP}(z_{pix}) - p_g\right\|^2 \tag{1}$$

Additionally, the training losses for both the diffusion process and critic prediction follow the methodology introduced in NavDP. We jointly optimize the action loss, critic loss, and goal alignment loss, balancing them using weighting coefficients. We set the coefficients as $\alpha = 0.8$, $\beta = 0.2$ and $\gamma = 0.5$. The overall training objective is defined as:

$$\mathcal{L}^{system1} = \alpha \cdot \mathcal{L}^{act} + \beta \cdot \mathcal{L}^{critic} + \gamma \cdot L^{goal} \tag{2}$$

Stage 2: Multi-System Joint-tuning. It is ambigious to represent an accurate 3-D navigation target by a 2-D pixel and challenging to perform high-speed inference for a 7B VLM on embedded devices. Therefore, the design of the intermediate feature connections that can bridge different systems is a critical factor. Such intermediates should preserve the advantages of the original systems—without degrading their efficiency or representational capacity—while at the same time enabling effective information flow across these systems with complementary functionalities. Instead of directly using the VLM's hidden states, which contain a mixture of abundant heterogeneous information, we introduce a set of learnable latent queries. The output latent features serve as compact intermediates that bridge the vision-language model (VLM) and the diffusion policy model via prompt tuning. Additionally, we adjust the temporal alignment of the two system inputs to accommodate asynchronous execution. Specifically, System 1 receives the most recent observation at timestep T, while the RGB memory input to System 2 is sampled from an earlier timestep in the range $(0, T - K)$, where K is a randomly selected interval drawn from the range $(0, 12)$. This temporal decoupling allows the dual-system framework to better adapt to asynchronous execution.

4.5. Extension: Learning Better Latent Plans with World Model

For building a better representation of latent plan, we introduce an extension of our model by using a predictive world model decoder to generate the egocentric observation sequence towards the mid-term goal. This paradigm potentially leads to scalable training with Internet videos and implicitly enhances

forecasting capabilities in dynamic environments. Specifically, we adopt the pre-trained 1.3B Wan2.1 model Wan et al. (2025) as our backbone, replacing its original T5-based encoder Raffel et al. (2020) with latent plan tokens generated by System 2. After fine-tuning on the InternData-N1 navigation dataset, the world model is able to simulate future outcomes conditioned on System 2 output latent plans with high predictive accuracy.

5. Experiments

5.1. System2 Evaluation

Dataset & Evaluation Metrics. We evaluate System 2 on the R2R-CE Anderson et al. (2018a) and RxR-CE Ku et al. (2020) benchmarks, both established under the VLN-CE Krantz et al. (2020b) setting using the Habitat simulator. These benchmarks simulate realistic indoor navigation in Matterport3D environments, where agents are required to follow natural language instructions under continuous control. R2R-CE provides English-only instructions with relatively short paths, while RxR-CE is a large-scale multilingual benchmark with longer and more diverse trajectories.

To evaluate the generalization ability of System 2, we conduct all experiments on the validation unseen splits of both benchmarks. Following prior work, we adopt standard VLN metrics: **Navigation Error (NE)**, which measures the final distance to the goal; **Success Rate (SR)**, the percentage of episodes where the agent stops within 3 meters of the goal; **Oracle Success Rate (OSR)**, where the best point along the path is considered; and **Success weighted by Path Length (SPL)**, which penalizes unnecessarily long trajectories. These metrics provide a comprehensive evaluation of both effectiveness and efficiency in instruction following.

Main Results. We compare our method with three major categories of VLN baselines: (1) Sensor-rich baselines that utilize panoramic images, odometry, and depth (e.g., HPN+DN, CMA, GridMM, ETPNav); (2) VLN methods that rely on depth and single first-person RGB without leveraging large-scale vision-language models (e.g., CM2, LAW, WS-MGMap). (3) Video-LLMs based VLN models with single RGB inputs (e.g., NaVid, MapNav, NaVILA, UniNaVid). InternVLA-N1 is evaluated under two settings: RGB-only (S2) and RGB+Depth (S1+S2). As shown in Table 2, our RGB-only variant already outperforms all previous RGB-based methods, achieving a Success Rate (SR) of 55.4% and SPL of 52.1% on R2R Val-Unseen, surpassing NaVILA (SR: 54.0%, SPL: 49.0%) and MapNav (SR: 39.7%, SPL: 37.2%).

5.2. System1 Evaluation

Dataset & Evaluation Metrics. To assess the generalization and robustness of System 1, we build a simulation benchmark using IsaacSim, which reflects potential sim-to-real gap for real-robot deployment. We collect a diverse range of scenarios for a comprehensive evaluation. The scenarios consist of two main categories: randomly generated layouts featuring cluttered obstacles, and professionally designed layouts that cover both residential and commercial environments Wang et al. (2024a). An overview of the evaluation scenes are illustrated in Figure 5. We divide all the evaluation environments into four subset which are **ClutterEnv-Easy (10)**, **ClutterEnv-Hard (10)**, **InternScenes-Home (20)**, **InternScenes-Commercial (20)**. The number represents the amount of evaluation scene assets. Three types of local navigation tasks are evaluated within the environments on a wheeled robot. For no-goal exploration task, we measure the metric **Episode Time** and **Explore Area** to access the collision avoidance and exploration skills. For point-goal navigation and image-goal navigation task, we evaluate both **Success Rate (SR)** and **Success weighted by Path Length (SPL)**. The episode is defined as success if the agent arrives at the goal point within 1.0m. For each task, the robot is randomly initialized and evaluated for 100 episodes within each scene.

Method	Observation				R2R Val-Unseen				RxR Val-Unseen			
	Pano.	Odo.	Depth	S.RGB	NE↓	OS↑	SR↑	SPL↑	NE↓	SR↑	SPL↑	nDTW↑
HPN+DN* Krantz et al. (2021)	✓	✓	✓		6.31	40.0	36.0	34.0	-	-	-	-
CMA* Hong et al. (2022)	✓	✓	✓		6.20	52.0	41.0	36.0	8.76	26.5	22.1	47.0
Sim2Sim* Krantz and Lee (2022)	✓	✓	✓		6.07	52.0	43.0	36.0	-	-	-	-
GridMM* Wang et al. (2023b)	✓	✓	✓		5.11	61.0	49.0	41.0	-	-	-	-
ETPNav* An et al. (2023)	✓	✓	✓		4.71	65.0	57.0	49.0	5.64	54.7	44.8	61.9
ScaleVLN* Wang et al. (2023a)	✓	✓	✓		4.80	–	55.0	51.0	-	-	-	-
InstructNav Long et al. (2024c)	✓	✓	✓	✓	6.89	–	31.0	24.0	-	-	-	-
AG-CMTP Chen et al. (2021)	✓	✓	✓		7.90	39.2	23.1	19.1	-	-	-	-
R2R-CMTP Chen et al. (2021)	✓	✓	✓		7.90	38.0	26.4	22.7	-	-	-	-
LAW Raychaudhuri et al. (2021)		✓	✓	✓	6.83	44.0	35.0	31.0	10.90	8.0	8.0	38.0
CM2 Georgakis et al. (2022)		✓	✓	✓	7.02	41.5	34.3	27.6	-	-	-	-
WS-MGMap Chen et al. (2022)		✓	✓	✓	6.28	47.6	38.9	34.3	-	-	-	-
ETPNav + FF Wang et al. (2024b)		✓	✓	✓	5.95	55.8	44.9	30.4	8.79	25.5	18.1	-
Seq2Seq Krantz et al. (2020b)			✓	✓	7.77	37.0	25.0	22.0	12.10	13.9	11.9	30.8
CMA Krantz et al. (2020b)			✓	✓	7.37	40.0	32.0	30.0	-	-	-	-
NaVid Zhang et al. (2024)				✓	5.47	49.1	37.4	35.9	-	-	-	-
MapNav Zhang et al. (2025c)				✓	4.93	53.0	39.7	37.2	-	-	-	-
NaVILA Cheng et al. (2025)				✓	5.37	57.6	49.7	45.5	-	-	-	-
NaVILA† Cheng et al. (2025)				✓	5.22	62.5	54.0	49.0	6.77	49.3	44.0	58.8
UniNaVid† Zhang et al. (2025a)				✓	5.58	53.3	47.0	42.7	6.24	48.7	40.9	-
InternVLA-N1 (S2)†				✓	4.89	60.6	55.4	52.1	6.41	49.5	41.8	62.6
InternVLA-N1 (S1+S2)†			✓	✓	**4.83**	**63.3**	**58.2**	**54.0**	**5.91**	**53.5**	**46.1**	**65.3**

Table 2 | Comparison with state-of-the-art methods on VLN-CE R2R and RxR Val-Unseen split. ∗ indicates methods using the waypoint predictor from Hong et al. (2022). † denotes methods using additional training data beyond the R2R-CE and RxR-CE benchmarks.

Figure 5 | An overview of ClutteredEnv and InternScenes scenarios for System 1 evaluation.Top rows are from ClutterEnv, bottom row are from InternScenes-Home.

Main Results. We compare our System 1 model with a diverse range of baseline methods. The baselines include GNM Shah et al. (2023a), ViNT Shah et al. (2023b), and NoMad Sridhar et al. (2024) for image-goal and no-goal tasks, as well as DD-PPO Wijmans et al. (2019), iPlanner Yang

et al. (2023), and ViPlanner Roth et al. (2024) for the point-goal navigation task. The main results are presented in Figure 6, Figure 7, and Figure 8. We find that our System 1 possesses several distinctive capabilities that enable it to outperform the baseline methods by a large margin. (1) Robust collision avoidance behavior in out-of-distribution scenarios: Although the majority of training data for is collected from indoor scenes, it achieves **2.7x** better performance than NoMad in the no-goal exploration task within ClutterEnv scenarios. (2) Efficient and consistent path-planning ability: In InternScenes scenarios with complex indoor layouts, our System 1 model excels at inferring connectivity among different areas and achieves **10.9%** higher success rate than previous methods. (3) Image-driven exploration: Most prior local navigation approaches fail at image-goal navigation when the goal image is located far away. However, our model can adaptively balance exploration and exploitation, resulting in **27.1%** better performance than previous methods.

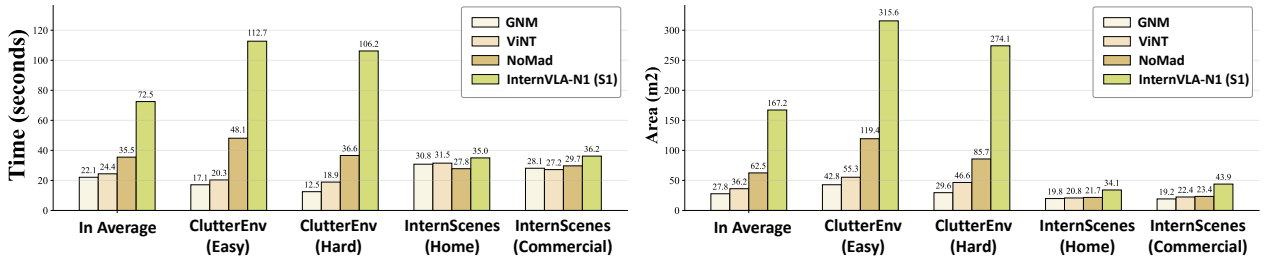

Figure 6 | System 1 evaluation metrics on the no-goal exploration task. Our InternVLA-N1(S1) model achieves more than **2x** performance score than the baselines.

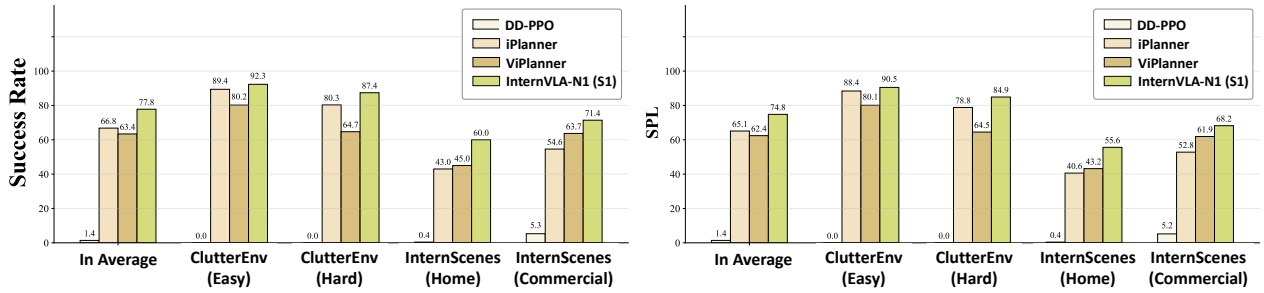

Figure 7 | System 1 evaluation metrics on point-goal navigation task. Our InternVLA-N1(S1) consistently outperforms the previous SOTA approach in all environments.

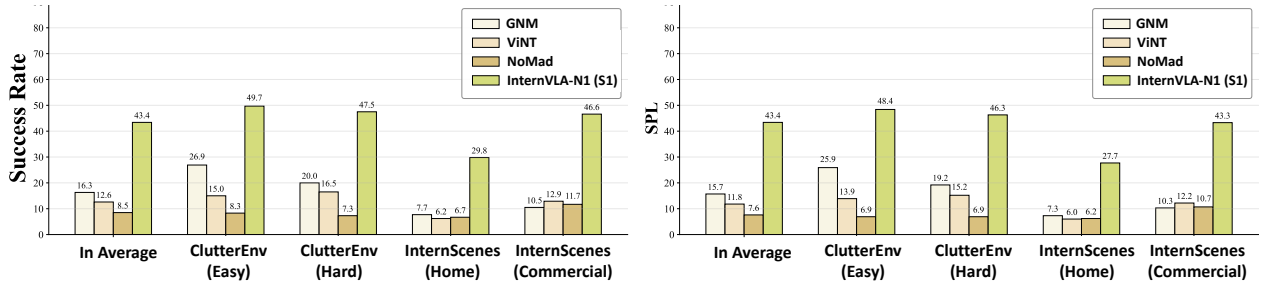

Figure 8 | System 1 evaluation metrics on image-goal navigation task. Our InternVLA-N1(S1) model achieves **27.1%** performance in average better than the baseline methods.

5.3. Dual-System Evaluation

Dataset & Evaluation Metrics. We first evaluate our dual-system on the same VLN-CE benchmark as the System 2 evaluation, by replacing the default point-goal navigation policy in Habitat-Sim with our System 1. We further evaluate our dual-system on VLN-PE Wang et al. (2025b), a physically

Method	R2R Validation Seen							R2R Validation Unseen						
	TL↓	NE↓	FR↓	StR↓	OS↑	SR↑	SPL↑	TL↓	NE↓	FR↓	StR↓	OS↑	SR↑	SPL↑
Random	0.14	8.24	0.30	0	0.30	0.30	0.30	0.11	7.78	0.74	0	3.34	3.04	2.30
`Train on VLN-PE`														
Seq2Seq	10.61	7.53	27.36	4.26	32.67	19.75	14.68	10.85	7.88	26.80	5.57	28.13	15.14	10.77
Seq2Seq+	10.22	7.75	33.43	3.19	30.09	16.86	12.54	9.88	7.85	26.27	6.52	28.79	16.56	12.7
CMA	11.13	7.59	23.71	3.19	34.94	21.58	16.10	11.16	7.98	22.64	3.27	33.11	19.15	14.05
CMA+	8.86	7.14	23.56	3.50	36.17	25.84	21.75	8.70	7.26	21.75	3.27	31.40	22.12	18.65
RDP	13.26	6.76	27.51	1.82	38.60	25.08	17.07	12.70	6.72	24.57	3.11	36.9	25.24	17.73
`Zero-shot Transfer Evaluation from VLN-CE`														
Seq2Seq†	7.80	7.62	20.21	3.04	19.30	15.20	12.79	7.73	7.18	18.04	3.04	22.42	16.48	14.11
CMA†	**6.62**	7.37	20.06	3.95	18.54	16.11	14.64	**6.58**	7.09	17.07	3.79	20.86	16.93	15.24
NaVid	7.54	6.20	**11.25**	**0.46**	24.32	21.58	17.45	7.12	5.94	**8.61**	**0.45**	27.32	22.42	18.58
InternVLA-N1	10.46	**4.71**	15.50	1.06	**57.60**	**53.34**	**44.53**	9.74	**4.73**	12.55	3.04	**56.72**	**50.63**	**43.31**

Table 3 | Evaluation Metrics on VLN-PE benchmark with physical locomotion controller. +: model is first trained on Habitat and fine-tuned on VLN-PE. †: model is trained with data augmentation.

Method	R2R Validation Seen							R2R Validation Unseen						
	TL↓	NE↓	FR↓	StR↓	OS↑	SR↑	SPL↑	TL↓	NE↓	FR↓	StR↓	OS↑	SR↑	SPL↑
`Train on VLN-PE`														
Seq2Seq	34.59	12.45	0.45	0	48.02	19.0	12.7	19.24	8.27	0.22	0	43.05	15.74	9.70
CMA	151.60	141.50	1.52	0	46.05	19.15	13.71	40.21	31.24	0.22	0	45.06	20.94	14.06
RDP	15.68	7.18	1.06	0	43.30	25.84	18.22	15.12	6.98	0.30	0	42.54	24.94	17.54
`Zero-shot Transfer Evaluation from VLN-CE`														
InternVLA-N1	**11.22**	**3.58**	0	0	**75.23**	**68.84**	**61.33**	**10.11**	**4.13**	0.45	0	**67.63**	**60.36**	**54.93**

Table 4 | Evaluation Metrics on VLN-PE benchmark with flash controller.

realistic VLN platform and benchmark that simulates robot dynamics and control errors encountered in real-world deployment. We consider the R2R dataset Anderson et al. (2018a) with the Humanoid Unitree H1 robot in VLN-PE.

Following standard VLN evaluation protocols Anderson et al. (2018a); Krantz et al. (2020b), five primary metrics are employed: **Trajectory Length (TL)** quantifies the average length of agent's navigation trajectory, measured in meters. **Navigation Error (NE)** records the average distance between the location where the agent finally stops and the designated destination. **Success Rate (SR)** denotes the probability that the agent successfully arrives at the destination. Note that the agent is deemed to have reached the destination if it comes to a stop within a 3-meter radius of the target location. **Oracle Success Rate (OS)** refers to the probability that any point in the agent's navigation trajectory reaches the destination. **SR weighted by Path Length (SPL)** balances SR and TL. As for physical simulation, two additional metrics are applied: **Fall Rate (FR)** measures the frequency of robot falls, **Stuck Rate (StR)** measures the occurrences in which the agent is unable to move.

Main Results. The experiment results shown in Table 2 demonstrate a further 2.8% and 4.1% success rate improvement on R2R and RxR benchmarks over the single System 2. This highlights the importance of the coordination between two systems and the superior performance of System 1 in collision avoidance and path planning. Baseline methods include Seq2Seq Krantz et al. (2020b), CMA Krantz et al. (2020b), RDP Wang et al. (2025b), and NaVid Zhang et al. (2024). Seq2Seq is a sequence-to-sequence model that predicts the next action from RGBD observations with a recurrent policy. CMA utilizes cross-modal attention between RGBD features from Seq2Seq and the instruction to predict the next action. RDP employs a Transformer-based diffusion decoder to predict continuous

relative displacement and yaw angle. NaVid is a video-based large vision language model that aims for better generalization and eliminating Sim2Real gap, while not requiring depth or odometer inputs.

Table 3 demonstrates results on VLN-PE with the physical locomotion controller. Although InternVLA-N1 is not fine-tuned with any H1 robot data from VLN-PE, it still significantly outperforms all models trained on VLN-PE, achieving at least a 17% improvement in SR. Moreover, compared to the single-system model NaVid, which demonstrates favorable zero-shot transfer performance on VLN-PE, InternVLA-N1 exhibits notable superior performance, with improvements of 20.21% in SR and 17.22% SPL, respectively. These results highlight the strong generalization capability and robustness of our dual-system integration design.

Table 4 exhibits results on VLN-PE with the flash controller. As the flash controller directly teleports the agent to the target coordinates, the agent is immune to getting stuck and seldom encounters falling incidents. After excluding challenges involving motion dynamics and stuck situations, the performance of InternVLA-N1 shows further enhancement, achieving at least a 35% improvement in SR compared to the models trained on VLN-PE.

5.4. Extension: World Model Qualitative Results

Our world model is fine-tuned to condition on the latent plan tokens generated by System 2 and to produce corresponding egocentric video sequences that depict navigation toward the expected goal. We evaluate the quality of the generated videos in both simulated and real-world environments. The qualitative results (Figure 9) demonstrate that the world model effectively generates realistic navigation trajectories toward the visualized pixel-level goal, while preserving fine-grained visual details and maintaining consistent geometric structure. In addition, we find that incorporating a video prediction objective can accelerate joint tuning, reducing the number of training epochs required to reach optimal evaluation performance on the R2R-CE benchmark from 4 to just 2 epochs.

5.5. Real-World Experiments

Experiment Setup. We perform real-world experiments on a wheeled (Turtlebot4), quadruped (Unitree Go2) and humanoid (Unitree G1) robots. All robots are equipped with Intel Realsense D455 cameras, which are mounted at different height but angled downward at 15°. We deploy our whole system on a remote machine with an RTX 4090 GPU. The InternVLA-N1 model takes around 20GB memory of the GPU. We evaluate the zero-shot **instruction following** and **obstacle avoidance** performance in multiple indoor and outdoor scenarios.

Pipeline & Speed. Given a VLN instruction initially, the robot continuously captures real-time aligned images (RGB & Depth) and transmits them to the remote server for inference. The server performs asynchronous inference of the dual-system model in the background and returns the latest trajectory or discrete actions to the robot. The trajectory will be transformed into the world coordinate according to the odometry at the time of the inference images and tracked with an MPC controller. We reuse the KV-cache in the multi-turn dialogue of System 2, accelerating the inference speed of the trajectory tokens from about 1.1s to 0.7s. Optimized by using TensorRT, our System 1 model generates 32 trajectories parallelly in about 0.03s. Thanks to the asynchronous pipeline and the inference optimization, the robot can get newer trajectory after the last one is fully tracked to the end, leading to a smoother motion. The System 2 will output STOP flag when reaching the goal of the language instruction. A real-world experiment is considered as successful if **(1)** the robot remains collision-free with all static and dynamic obstacles, **(2)** the robot passes all desired landmarks and stops at the desired goal.

Main Results. We select several representative real-world scenarios such as office, canteen, street

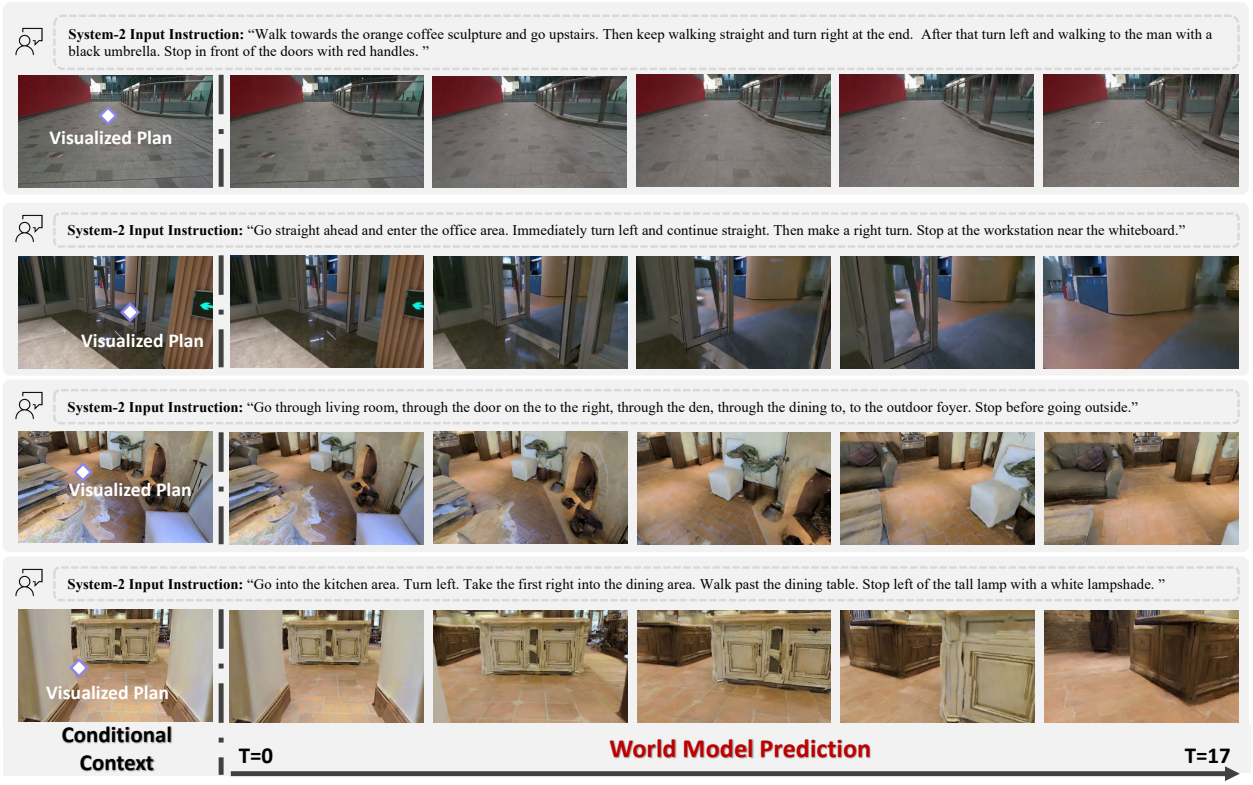

Figure 9 | Qualitative results of the InternVLA-N1 world model. The top two rows show prediction results in real-world scenarios, while the bottom two show the results in unseen simulation scene.

and convenience store for evaluation. Note that all real-world experiments are zero-shot without collecting any scene-specific data for finetuning our model. Qualitative results are presented in Figure 10. Thanks to the dual-system pattern, InternVLA-N1 can perceive high-frequency images of a dynamic environment and plan reactive collision-free trajectories to finish the long-horizon vision language navigation task (Panel 1). Evaluation in canteen (Panel 2) shows that our whole system selects correct pixel goals and generates safe trajectories in cluttered environments. For long-horizon instruction following and semantic understanding, our whole system plans smooth trajectories to pass all desired landmarks and stops at the final goal in office and street (Panel 3 & 4). Our whole system demonstrates robustness for dynamic pedestrians, staircases and varying light during the whole process. Furthermore, we conduct experiments (Panel 5 & 6) to test our system's ability on human-like short instructions instead of step-by-step instructions in our training sets. Results show that the robot can also understand the instructions and finish the tasks in some cases. Our InternVLA-N1 model is robust across different real-world platforms. While different robots exhibit variation in camera height, vibration and tracking performance, InternVLA-N1 still finishes the VLN task well. For detailed demonstrations of our experiments, please refer to our homepage.

Baselines & Metrics. To quantitatively evaluate the robustness and generalization of InternVLA-N1 in real-world scenarios, we compared the performance of our model with other baseline methods across hallway (easy VLN instruction), bedroom (medium VLN instruction in a single room), and office (hard VLN instruction of room-to-room) scenarios. Baselines include traditional learning-based method CMA Krantz et al. (2020a), VLM-based methods NaVid Zhang et al. (2024), NaVILA Cheng et al. (2025) and our previous work StreamVLN Wei et al. (2025) that outputs discrete actions. We conducted 20 experiments in each scenario for each model, aiming to observe the Success Rate (SR) and Navigation Errors (NE) of performing the VLN task. Quantitative and qualitative results are shown

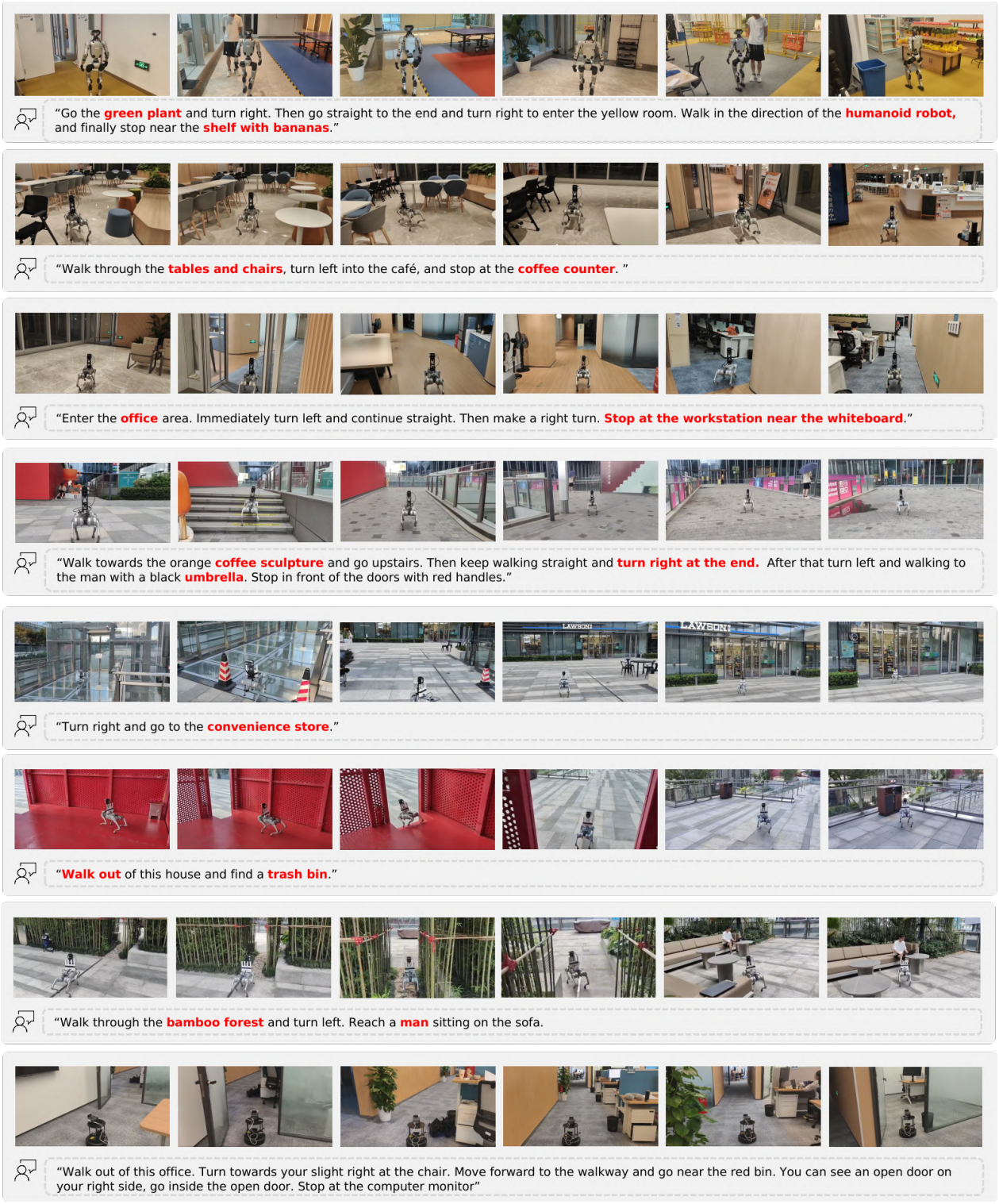

Figure 10 | Real-world experiment visualization of InternVLA-N1 in diverse scenarios.

in Figure 11 and 12. Lightweight model CMA can only complete VLN tasks under simple scenarios and instructions, with an SR significantly lower than that of VLM-based methods. This advantage might be attributed to their strong visual-language understanding capabilities inherited from their VLM foundation models. Among all VLM-based methods, NaVid tends to rotate in place after travelling some distance and fails under complex instructions. NaVILA demonstrates long-horizon instruction

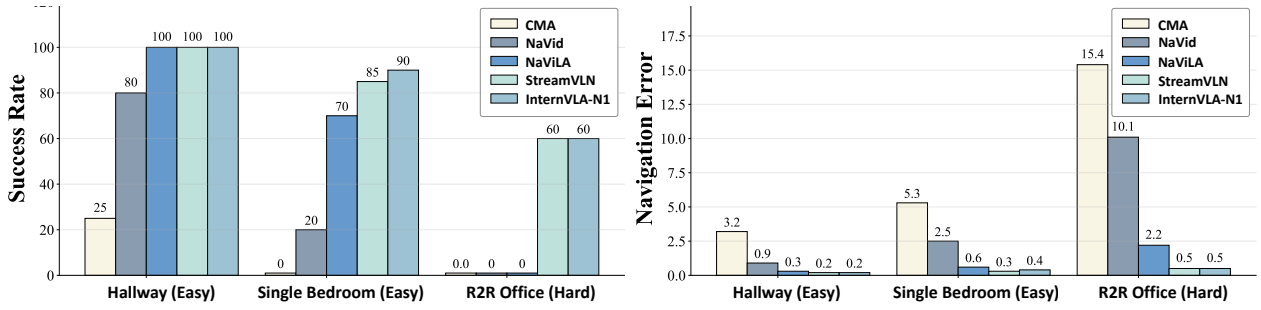

Figure 11 | Quantitative metrics across different VLN approaches.

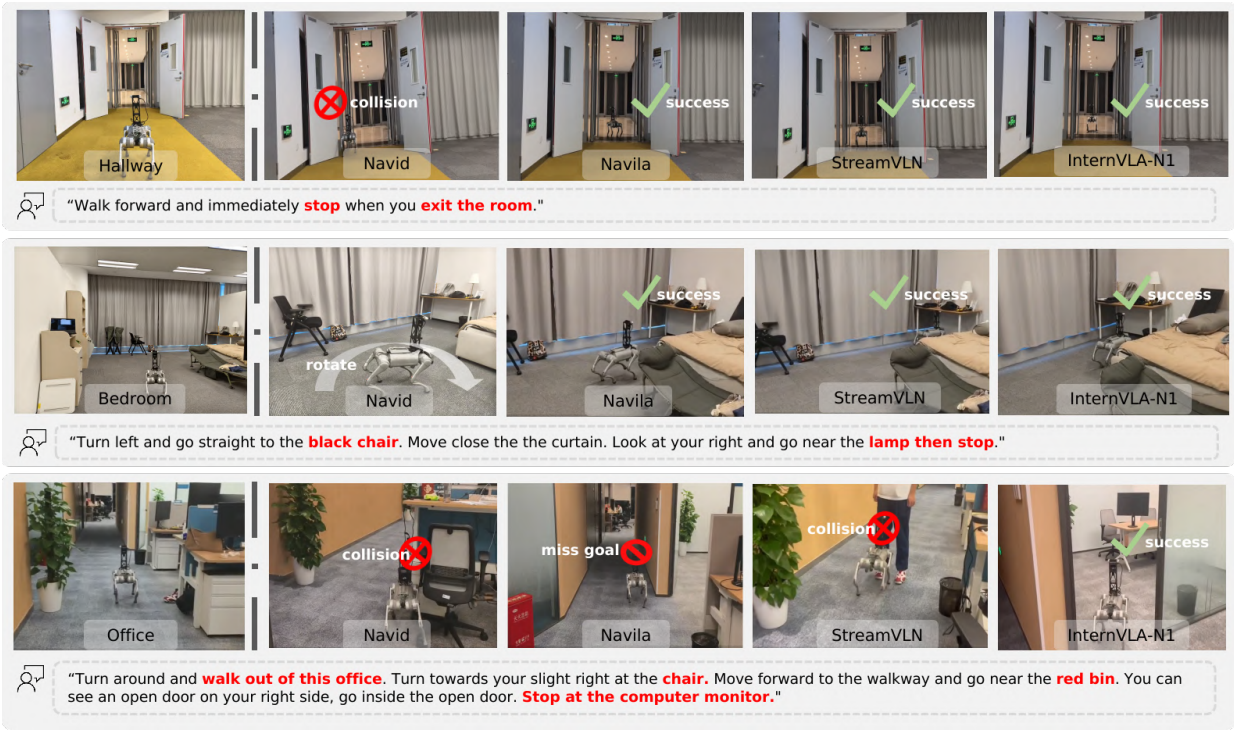

Figure 12 | Behavior comparison and failure analysis among different methods in real-world.

following ability, but it fails to navigate to the final goal based on the language instruction in complex office scenario. In contrast, StreamVLN show superior performance in long-horizon static scenarios, but it has a low SR when dynamic obstacles such as pedestrians block its path. Our InternVLA-N1 dual-system model achieves favourable SR and NE in both static and dynamic scenarios. It can avoid dynamic pedestrians present in the environment and finish the final VLN task successfully.

6. Conclusion

In this report, we introduce InternVLA-N1, the first dual-system vision-language navigation foundation model. Our asynchronous framework integrates multi-modal reasoning, instruction-following, long-horizon planning, and real-time dynamic obstacle avoidance within a unified architecture. These capabilities demonstrate strong zero-shot generalization in open-world settings and can be directly deployed across diverse robotic platforms. A persistent challenge in the field has been the limited scale of available navigation datasets. To address this, we also release InternData-N1, a large-scale, high-quality dataset designed to support complex navigation tasks. We hope that all our open-sourced resources will benefit the broader research community and foster continued advancements in embodied AI and robot navigation.

References

D. An, Y. Qi, Y. Li, Y. Huang, L. Wang, T. Tan, and J. Shao. Bevbert: Multimodal map pre-training for language-guided navigation. *arXiv preprint arXiv:2212.04385*, 2022.

D. An, H. Wang, W. Wang, Z. Wang, Y. Huang, K. He, and L. Wang. Etpnav: Evolving topological planning for vision-language navigation in continuous environments. *arXiv preprint arXiv:2304.03047*, 2023.

P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683, 2018a.

P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683, 2018b.

S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. Fan, Y. Fang, D. Fox, F. Hu, S. Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Q. Bu, H. Li, L. Chen, J. Cai, J. Zeng, H. Cui, M. Yao, and Y. Qiao. Towards synergistic, generalized, and efficient dual-system for robotic manipulation. *arXiv preprint arXiv:2410.08001*, 2024.

Q. Bu, Y. Yang, J. Cai, S. Gao, G. Ren, M. Yao, P. Luo, and H. Li. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025.

W. Cai, J. Peng, Y. Yang, Y. Zhang, M. Wei, H. Wang, Y. Chen, T. Wang, and J. Pang. Navdp: Learning sim-to-real navigation diffusion policy with privileged information guidance. *arXiv preprint arXiv:2505.08712*, 2025.

A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

M. Chang, A. Gupta, and S. Gupta. Semantic visual navigation by watching youtube videos. *Advances in Neural Information Processing Systems*, 33:4283–4294, 2020.

J. Chen, B. Lin, R. Xu, Z. Chai, X. Liang, and K.-Y. K. Wong. Mapgpt: Map-guided prompting with adaptive path planning for vision-and-language navigation. *arXiv preprint arXiv:2401.07314*, 2024.

J. Chen, B. Lin, X. Liu, L. Ma, X. Liang, and K.-Y. K. Wong. Affordances-oriented planning using foundation models for continuous vision-language navigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 23568–23576, 2025.

K. Chen, J. K. Chen, J. Chuang, M. Vázquez, and S. Savarese. Topological planning with transformers for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

P. Chen, D. Ji, K. Lin, R. Zeng, T. H. Li, M. Tan, and C. Gan. Weakly-supervised multi-granularity map learning for vision-and-language navigation. *arXiv preprint arXiv:2210.07506*, 2022.

A.-C. Cheng, Y. Ji, Z. Yang, Z. Gongye, X. Zou, J. Kautz, E. Bıyık, H. Yin, S. Liu, and X. Wang. Navila: Legged robot vision-language-action model for navigation. *Robotics: Science and Systems*, 2025.

M. Denninger, M. Sundermeyer, D. Winkelbauer, D. Olefir, T. Hodan, Y. Zidan, M. Elbadrawy, M. Knauer, H. Katam, and A. Lodhi. Blenderproc: Reducing the reality gap with photorealistic rendering. In *16th Robotics: Science and Systems, RSS 2020, Workshops*, 2020.

A. Eftekhar, L. Weihs, R. Hendrix, E. Caglar, J. Salvador, A. Herrasti, W. Han, E. VanderBilt, A. Kembhavi, A. Farhadi, et al. The one ring: a robotic indoor navigation generalist. In *The first CVPR workshop on 3D Vision Language Models (VLMs) for Robotics Manipulation: Opportunities and Challenges*, 2024.

FigureAI. Helix: A vision-language-action model for generalist humanoid control. Technical report, FigureAI, 02 2025. URL https://www.figure.ai/news/helix.

D. Fox, W. Burgard, and S. Thrun. The dynamic window approach to collision avoidance. *IEEE Robotics & Automation Magazine*, 4(1):23–33, 1997.

H. Fu, B. Cai, L. Gao, L.-X. Zhang, J. Wang, C. Li, Q. Zeng, C. Sun, R. Jia, B. Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10933–10942, 2021.

C. Gao, L. Jin, X. Peng, J. Zhang, Y. Deng, A. Li, H. Wang, and S. Liu. Octonav: Towards generalist embodied navigation. *arXiv preprint arXiv:2506.09839*, 2025.

G. Georgakis, K. Schmeckpeper, K. Wanchoo, S. Dan, E. Miltsakaki, D. Roth, and K. Daniilidis. Cross-modal map learning for vision and language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

H. He, Y. Ma, W. Wu, and B. Zhou. From seeing to experiencing: Scaling navigation foundation models with reinforcement learning. *arXiv preprint arXiv:2507.22028*, 2025.

N. Hirose, A. Sadeghian, M. Vázquez, P. Goebel, and S. Savarese. Gonet: A semi-supervised deep learning approach for traversability estimation. In *2018 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 3044–3051. IEEE, 2018.

N. Hirose, D. Shah, A. Sridhar, and S. Levine. Sacson: Scalable autonomous control for social navigation. *IEEE Robotics and Automation Letters*, 9(1):49–56, 2023.

Y. Hong, Z. Wang, Q. Wu, and S. 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*, 2022.

M. Z. Irshad, N. C. Mithun, Z. Seymour, H.-P. Chiu, S. Samarasekera, and R. Kumar. Semantically-aware spatio-temporal reasoning agent for vision-and-language navigation in continuous environments. In *2022 26th International conference on pattern recognition (ICPR)*, pages 4065–4071. IEEE, 2022.

D. Kahneman. *Thinking, fast and slow*. macmillan, 2011.

S. Karaman and E. Frazzoli. Sampling-based algorithms for optimal motion planning. *The International Journal of Robotics Research*, 30(7):846–894, 2011.

H. Karnan, A. Nair, X. Xiao, G. Warnell, S. Pirk, A. Toshev, J. Hart, J. Biswas, and P. 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.

M. Khanna, Y. Mao, H. Jiang, S. Haresh, B. Shacklett, D. Batra, A. Clegg, E. Undersander, A. X. Chang, and M. Savva. Habitat synthetic scenes dataset (hssd-200): An analysis of 3d scene scale and realism tradeoffs for objectgoal navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16384–16393, 2024.

D. Kramer and C. Stachniss. Timed elastic bands for time-optimal point-to-point navigation in constrained environments. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3316–3322. IEEE, 2012.

J. Krantz and S. Lee. Sim-2-sim transfer for vision-and-language navigation in continuous environments. In *European Conference on Computer Vision*, 2022.

J. Krantz, E. Wijmans, A. Majumdar, D. Batra, and S. Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *European Conference on Computer Vision*, pages 104–120. Springer, 2020a.

J. Krantz, E. Wijmans, A. Majundar, D. Batra, and S. Lee. Beyond the nav-graph: Vision and language navigation in continuous environments. In *European Conference on Computer Vision (ECCV)*, 2020b.

J. Krantz, A. Gokaslan, D. Batra, S. Lee, and O. Maksymets. Waypoint models for instruction-guided navigation in continuous environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

A. Ku, P. Anderson, R. Patel, E. Ie, and J. Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. *arXiv preprint arXiv:2010.07954*, 2020.

B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

B. Lin, Y. Nie, Z. Wei, J. Chen, S. Ma, J. Han, H. Xu, X. Chang, and X. Liang. Navcot: Boosting llm-based vision-and-language navigation via learning disentangled reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025a.

F. Lin, R. Nai, Y. Hu, J. You, J. Zhao, and Y. Gao. Onetwovla: A unified vision-language-action model with adaptive reasoning. *arXiv preprint arXiv:2505.11917*, 2025b.

K. Lin, P. Chen, D. Huang, T. H. Li, M. Tan, and C. Gan. Learning vision-and-language navigation from youtube videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8317–8326, 2023.

W. Liu, H. Zhao, C. Li, J. Biswas, B. Okal, P. Goyal, Y. Chang, and S. Pouya. X-mobility: End-to-end generalizable navigation via world modeling. *arXiv preprint arXiv:2410.17491*, 2024.

X. Liu, J. Li, Y. Jiang, N. Sujay, Z. Yang, J. Zhang, J. Abanes, J. Zhang, and C. Feng. Citywalker: Learning embodied urban navigation from web-scale videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6875–6885, 2025.

J. Long, J. Ren, M. Shi, Z. Wang, T. Huang, P. Luo, and J. Pang. Learning humanoid locomotion with perceptive internal model. *arXiv preprint arXiv:2411.14386*, 2024a.

J. Long, Z. Wang, Q. Li, L. Cao, J. Gao, and J. Pang. Hybrid internal model: Learning agile legged locomotion with simulated robot response. In *ICLR*, 2024b.

Y. Long, W. Cai, H. Wang, G. Zhan, and H. Dong. Instructnav: Zero-shot system for generic instruction navigation in unexplored environment. *arXiv preprint arXiv:2406.04882*, 2024c.

Y. Long, X. Li, W. Cai, and H. Dong. Discuss before moving: Visual language navigation via multi-expert discussions. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 17380–17387. IEEE, 2024d.

Z. Luo, W. Y. a. J. Gong, M. Wang, Z. Zhang, X. Wang, Y. Xie, and X. Tan. Navimaster: Learning a unified policy for gui and embodied navigation tasks. *arXiv preprint arXiv:2508.02046*, 2025.

X. Meng, X. Yang, S. Jung, F. Ramos, S. S. Jujjavarapu, S. Paul, and D. Fox. Aim my robot: Precision local navigation to any object. *IEEE Robotics and Automation Letters*, 2025.

G. Pan, Q. Ben, Z. Yuan, G. Jiang, Y. Ji, S. Li, J. Pang, H. Liu, and H. Xu. Roboduet: Learning a cooperative policy for whole-body legged loco-manipulation. *IEEE Robotics and Automation Letters*, 2025.

Y. Qi, Q. Wu, P. Anderson, X. Wang, W. Y. Wang, C. Shen, and A. v. d. Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9982–9991, 2020.

Y. Qiao, W. Lyu, H. Wang, Z. Wang, Z. Li, Y. Zhang, M. Tan, and Q. Wu. Open-nav: Exploring zero-shot vision-and-language navigation in continuous environment with open-source llms. *arXiv preprint arXiv:2409.18794*, 2024.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

S. K. Ramakrishnan, A. Gokaslan, E. Wijmans, O. Maksymets, A. Clegg, J. Turner, E. Undersander, W. Galuba, A. Westbury, A. X. Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*, 2021.

R. Ramrakhya, E. Undersander, D. Batra, and A. Das. Habitat-web: Learning embodied object-search strategies from human demonstrations at scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5173–5183, 2022.

S. Raychaudhuri, S. Wani, S. Patel, U. Jain, and A. X. Chang. Language-aligned waypoint (law) supervision for vision-and-language navigation in continuous environments. *arXiv preprint arXiv:2109.15207*, 2021.

P. Roth, J. Nubert, F. Yang, M. Mittal, and M. Hutter. Viplanner: Visual semantic imperative learning for local navigation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5243–5249. IEEE, 2024.

P. Roth, J. Frey, C. Cadena, and M. Hutter. Learned perceptive forward dynamics model for safe and platform-aware robotic navigation. *arXiv preprint arXiv:2504.19322*, 2025.

M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. 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.

D. Shah, B. Eysenbach, G. Kahn, N. Rhinehart, and S. Levine. Rapid exploration for open-world navigation with latent goal models. *arXiv preprint arXiv:2104.05859*, 2021.

D. Shah, A. Sridhar, A. Bhorkar, N. Hirose, and S. Levine. Gnm: A general navigation model to drive any robot. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7226–7233. IEEE, 2023a.

D. Shah, A. Sridhar, N. Dashora, K. Stachowicz, K. Black, N. Hirose, and S. Levine. Vint: A foundation model for visual navigation. In *Conference on Robot Learning*, pages 711–733. PMLR, 2023b.

L. X. Shi, M. R. Equi, L. Ke, K. Pertsch, Q. Vuong, J. Tanner, A. Walling, H. Wang, N. Fusai, A. Li-Bell, et al. Hi robot: Open-ended instruction following with hierarchical vision-language-action models. In *Forty-second International Conference on Machine Learning*, 2025.

A. Sridhar, D. Shah, C. Glossop, and S. 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.

J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

A. Szot, A. Clegg, E. Undersander, E. Wijmans, Y. Zhao, J. Turner, N. Maestre, M. Mukadam, D. Chaplot, O. Maksymets, A. Gokaslan, V. Vondrus, S. Dharur, F. Meier, W. Galuba, A. Chang, Z. Kira, V. Koltun, J. Malik, M. Savva, and D. Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

H. Tan, L. Yu, and M. Bansal. Learning to navigate unseen environments: Back translation with environmental dropout, 2019. URL https://arxiv.org/abs/1904.04195.

T. Wan, A. Wang, B. Ai, B. Wen, C. Mao, C.-W. Xie, D. Chen, F. Yu, H. Zhao, J. Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

H. Wang, J. Chen, W. Huang, Q. Ben, T. Wang, B. Mi, T. Huang, S. Zhao, Y. Chen, S. Yang, et al. Grutopia: Dream general robots in a city at scale. *arXiv preprint arXiv:2407.10943*, 2024a.

H. Wang, A. H. Tan, A. Fung, and G. Nejat. X-nav: Learning end-to-end cross-embodiment navigation for mobile robots. *arXiv preprint arXiv:2507.14731*, 2025a.

L. Wang, X. Xia, H. Zhao, H. Wang, T. Wang, Y. Chen, C. Liu, Q. Chen, and J. Pang. Rethinking the embodied gap in vision-and-language navigation: A holistic study of physical and visual disparities. *arXiv preprint arXiv:2507.13019*, 2025b.

S. Wang, Y. Wang, W. Li, Y. Wang, M. Chen, K. Wang, Z. Su, X. Cai, Y. Jin, D. Li, et al. Monodream: Monocular vision-language navigation with panoramic dreaming. *arXiv preprint arXiv:2508.02549*, 2025c.

S. Wang, J. Zhang, M. Li, J. Liu, A. Li, K. Wu, F. Zhong, J. Yu, Z. Zhang, and H. Wang. Trackvla: Embodied visual tracking in the wild. *arXiv preprint arXiv:2505.23189*, 2025d.

Z. Wang, J. Li, Y. Hong, Y. Wang, Q. Wu, M. Bansal, S. Gould, H. Tan, and Y. Qiao. Scaling data generation in vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023a.

Z. Wang, X. Li, J. Yang, Y. Liu, and S. Jiang. Gridmm: Grid memory map for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023b.

Z. Wang, X. Li, J. Yang, Y. Liu, and S. Jiang. Sim-to-real transfer via 3d feature fields for vision-and-language navigation. *arXiv preprint arXiv:2406.09798*, 2024b.

J. Wasserman, K. Yadav, G. Chowdhary, A. Gupta, and U. Jain. Last-mile embodied visual navigation. In *Conference on Robot Learning*, pages 666–678. PMLR, 2023.

M. Wei, C. Wan, X. Yu, T. Wang, Y. Yang, X. Mao, C. Zhu, W. Cai, H. Wang, Y. Chen, et al. Streamvln: Streaming vision-and-language navigation via slowfast context modeling. *arXiv preprint arXiv:2507.05240*, 2025.

E. Wijmans, A. Kadian, A. Morcos, S. Lee, I. Essa, D. Parikh, M. Savva, and D. Batra. Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames. In *International Conference on Learning Representations*, 2019.

G. Williams, A. Aldrich, and E. A. Theodorou. Model predictive path integral control: From theory to parallel computation. In *2015 American Control Conference (ACC)*, pages 6281–6286. IEEE, 2015.

F. Xia, A. R. Zamir, Z.-Y. He, A. Sax, J. Malik, and S. Savarese. Gibson Env: real-world perception for embodied agents. In *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*. IEEE, 2018.

A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

F. Yang, C. Wang, C. Cadena, and M. Hutter. iplanner: Imperative path planning. *Proceedings of Robotics: Science and System XIX*, page 064, 2023.

L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.

L. Yu, B. Shi, R. Pasunuru, B. Muller, O. Golovneva, T. Wang, A. Babu, B. Tang, B. Karrer, S. Sheynin, et al. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*, 2(3), 2023.

W. Yuan, J. Duan, V. Blukis, W. Pumacay, R. Krishna, A. Murali, A. Mousavian, and D. Fox. Robopoint: A vision-language model for spatial affordance prediction in robotics. In *Conference on Robot Learning*, pages 4005–4020. PMLR, 2025.

Y. Zeng, H. Ren, S. Wang, J. Huang, and H. Cheng. Navidiffusor: Cost-guided diffusion model for visual navigation. *arXiv preprint arXiv:2504.10003*, 2025.

J. Zhang, K. Wang, R. Xu, G. Zhou, Y. Hong, X. Fang, Q. Wu, Z. Zhang, and H. Wang. Navid: Video-based vlm plans the next step for vision-and-language navigation. *Robotics: Science and Systems*, 2024.

J. Zhang, K. Wang, S. Wang, M. Li, H. Liu, S. Wei, Z. Wang, Z. Zhang, and H. Wang. Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks. *Robotics: Science and Systems*, 2025a.

L. Zhang, X. Hao, Y. Tang, H. Fu, X. Zheng, P. Wang, Z. Wang, W. Ding, and S. Zhang. *nava*^{3}: Understanding any instruction, navigating anywhere, finding anything. *arXiv preprint arXiv:2508.04598*, 2025b.

L. Zhang, X. Hao, Q. Xu, Q. Zhang, X. Zhang, P. Wang, J. Zhang, Z. Wang, S. Zhang, and R. Xu. Mapnav: A novel memory representation via annotated semantic maps for vision-and-language navigation, 2025c. URL https://arxiv.org/abs/2502.13451.

D. Zheng, S. Huang, L. Zhao, Y. Zhong, and L. Wang. Towards learning a generalist model for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13624–13634, 2024.

G. Zhou, Y. Hong, and Q. Wu. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. *arXiv preprint arXiv:2305.16986*, 2023.

G. Zhou, Y. Hong, and Q. Wu. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7641–7649, 2024.

X. Zhou, Z. Wang, H. Ye, C. Xu, and F. Gao. Ego-planner: An esdf-free gradient-based local planner for quadrotors. *IEEE Robotics and Automation Letters*, 6(2):478–485, 2020.

A. Author contributions

All contributors are listed in alphabetical order by their last names.

A.1. Core Contributors

Wenzhe Cai, Delin Feng, Yu Liu, Jiangmiao Pang, Jiaqi Peng, Chenyang Wan, Hanqing Wang, Liuyi Wang, Tai Wang, Meng Wei, Yuqiang Yang, Xiqian Yu, Chenming Zhu.

A.2. Contributors

Peizhou Cao, Yilun Chen, Zeyu He, Wensi Huang, Hengjie Li, Dahua Lin, Jingli Lin, Xihui Liu, Yilin Long, Xiaohan Mao, Yu Qiao, Jiawei Qiu, Yuan Shen, Yukai Wang, Xueyuan Wei, Chao Wu, Zhenyu Yang, Jia Zeng, Yiming Zeng, Siqi Zhang, Jingjing Zhang, Shenghan Zhang, Shi Zhang, Yuchang Zhang, Hui Zhao, Bowen Zhou, Yuanzhen Zhou, Haoyi Zhu, Shaohao Zhu.