

Rethinking the Embodied Gap in Vision-and-Language Navigation: A Holistic Study of Physical and Visual Disparities

Liyi Wang^{1,2*}, Xinyuan Xia^{2,3*}, Hui Zhao², Hanqing Wang^{2†}, Tai Wang², Yilun Chen², Chengju Liu^{1,4}, Qijun Chen^{1,4†}, Jiangmiao Pang²

¹Tongji University, ²Shanghai AI Laboratory, ³Shanghai Jiao Tong University,

⁴State Key Laboratory of Autonomous Intelligent Unmanned Systems

Abstract

Recent Vision-and-Language Navigation (VLN) advancements are promising, but their idealized assumptions about robot movement and control fail to reflect physically embodied deployment challenges. To bridge this gap, we introduce VLN-PE, a physically realistic VLN platform supporting humanoid, quadruped, and wheeled robots. For the first time, we systematically evaluate several ego-centric VLN methods in physical robotic settings across different technical pipelines, including classification models for single-step discrete action prediction, a diffusion model for dense waypoint prediction, and a train-free, map-based large language model (LLM) integrated with path planning. Our results reveal significant performance degradation due to limited robot observation space, environmental lighting variations, and physical challenges like collisions and falls. This also exposes locomotion constraints for legged robots in complex environments. VLN-PE is highly extensible, allowing seamless integration of new scenes beyond MP3D, thereby enabling more comprehensive VLN evaluation. Despite the weak generalization of current models in physical deployment, VLN-PE provides a new pathway for improving cross-embodiment's overall adaptability. We hope our findings and tools inspire the community to rethink VLN limitations and advance robust, practical VLN models. The code is available at https://crystalsixone.github.io/vln_pe.github.io.

1. Introduction

Vision-and-Language Navigation (VLN) [2] has emerged as a critical task in embodied AI, where agents are required to follow natural language instructions to navigate complex environments. Initially, VLN relied on the MP3D simulator [2], which only supported oracle-based

*Equal contribution

†Corresponding author

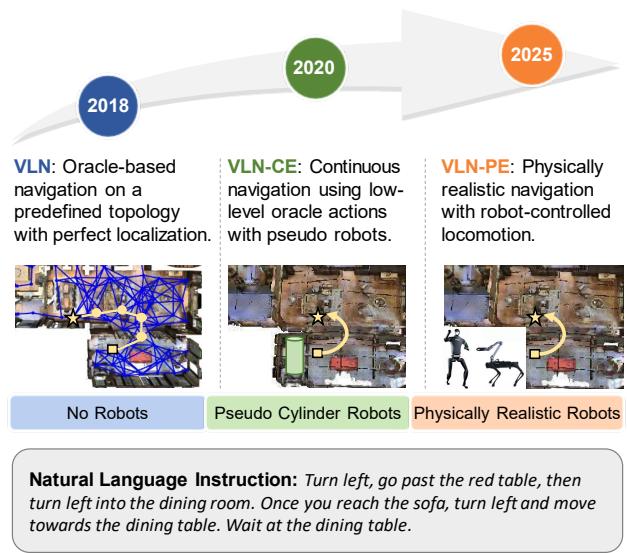


Figure 1. The evolution of vision-and-language navigation (VLN) task. As the techniques and algorithms develop, the settings become more and more practical and challenging.

navigation by jumping between predefined graph nodes. Later, VLN-CE [25] introduced continuous navigation using Habitat [46] (Fig. 1). A variety of advanced methods [1, 36, 56, 64] have highlighted the increasing potential of VLN models in advancing the future of embodied AI.

However, a major gap remains between simulation-based models and their physical deployment, especially when applied to diverse robot types, such as wheeled, humanoid, and quadruped robots. Most current VLN-related benchmarks [13, 26, 41, 47, 52] are designed for ideal wheeled or point-based agents, ignoring the physical embodiment of the robots themselves. Moreover, the testing conditions in current VLN platforms are overly idealized, often neglecting critical physical issues like viewpoint shifts, falling, deadlocks, and motion errors. Meanwhile, the rapid advancement of locomotion algorithms, particularly reinforcement learning (RL) methods for humanoid and

quadruped robots [5, 34, 39], has created a growing demand for an integrated VLN benchmarking platform that supports cross-embodiment data collection, training, and evaluation. However, existing research lacks a full-stack platform integrating realistic robotic dynamics, precise locomotion control, and scalable training. Most studies still rely on simplified environments and idealized navigation policies, often based on navigation meshes, which fail to reflect physical complexities. This raises a crucial question: *To what extent do physical embodiment constraints and visual environmental variations impact the performance of existing VLN methods?*

To date, no study has systematically analyzed the applicability of existing VLN methods to different kinds of physical agents. The performance loss remains largely unknown when transferring models from ideal simulated to physical settings. To address these gaps, there is a pressing need for a more realistic and adaptable VLN benchmark—one that evaluates language-guided navigation while accounting for the unique locomotion and execution challenges.

In this paper, we introduce VLN-PE (Fig. 2), a physically realistic VLN platform and benchmark that provides a comprehensive environment for cross-embodiment (humanoid, quadruped, and wheeled) data collection and systematic evaluation of policies across various robot embodiments and environmental conditions. Built on GRUTopia [54], VLN-PE can seamlessly integrate additional environments, including high-quality synthetic scenes and 3D Gaussian Splatting (3DGS) rendering scenes [22, 60], making it highly extensible and user-friendly. Beyond the widely used MP3D scenes [7], we introduce several high-quality 3D synthetic household scenes and 3DGS-scanned environments for expanding the scope for VLN research and evaluation. Specifically, we reveal the influence of the following physical and visual factors on the VLN models: (1) *Cross-Embodiment Perception*: The effects of robots perceiving and interpreting environments through their unique sensory systems. (2) *Controller Engagement*: The impact of whether or not to add a physical controller to the data acquisition and validation on model performance. (3) *Environmental Conditions*: The impact of diverse environmental conditions, such as light intensity and classification.

To align with the typical robot perception setup—usually consisting of a single ego-centric camera—we assess the following ego-centric VLN methods: (1) Single-step end-to-end methods: Two smaller models, each with approximately 36M parameters—Seq2Seq [25] and CMA [25]—along with a fine-tuned, large video-based model, NaVid [64], which has 7B parameters. (2) Multi-step end-to-end method: Diffusion Policy [6, 10] has shown promise in manipulation tasks but remains underexplored in VLN. We introduce RDP, the first diffusion-based attempt in VLN, as a new baseline method, capable of generat-

ing dense trajectory waypoints. (3) Map-based zero-shot large language model (LLM): VLMaps [21], a zero-shot approach using the LLM to ground the target on a semantic map and navigate using path planning.

Our experiments on VLN-PE reveal several critical insights that highlight limitations in current approaches and suggest promising directions for improvement:

- *SoTA Models Struggle in Physical Environments*: Existing VLN-CE models exhibit a 34% SR relative drop when transferred to physical settings, revealing a gap between pseudo-motion training and physical deployment.
- *Cross-embodiment Sensitivity*: Model performance varies across different robots, primarily due to viewpoint height differences, highlighting the need for height-adaptive or perspective-invariant representations.
- *Multi-Modal Robustness*: RGB-only models degrade significantly in low-light conditions, whereas RGB + depth models perform more reliably, underscoring the value of multi-modal fusion to improve the model’s robustness.
- *Limited Generalization of Standard Datasets*: MP3D-style datasets cannot fully capture environment shifts. A simple baseline with 6M trainable parameters, fine-tuned on our small-scale dataset of the newly introduced scenes, outperforms previous SoTA method in zero-shot settings, suggesting the importance of more diverse training distributions and comprehensive evaluation system.
- *Towards Cross-Embodiment VLN*: In our experiments, co-training across different robots enables a single baseline to generalize across embodiments and achieve the SoTA result, showing an important foundation for the future unified cross-embodiment VLN model.

2. Related Work

Vision-and-Language Navigation. In discrete settings, agents navigate by teleporting between predefined nodes on a navigation graph, which emphasizes cross-modal alignment [15, 29, 56], semantic understanding [11, 38, 55], and historical dependency [1, 42, 57]. In contrast, continuous settings (VLN-CE) [25], implemented through simulators like Habitat [44], allow free movement without navigation graphs. While panorama-based methods have shown strong performance in VLN-CE benchmarks [1, 3, 19, 24], they heavily rely on panoramic RGB and depth sensors. Since most real-world robots are equipped with egocentric RGB-D pinhole cameras, we focus on solutions that align with this hardware configuration. Traditional approaches [12, 16, 25, 58] focused on training end-to-end specific models, but often struggled with generalization due to simulator-specific constraints. Recent advances have introduced several promising directions: (1) Multimodal Large Language Models (MLLM) [9, 23, 33, 64, 65] have shown remarkable potential in cross-modal understanding and generalization; (2) diffusion policies [10, 31, 48] have emerged as an effec-

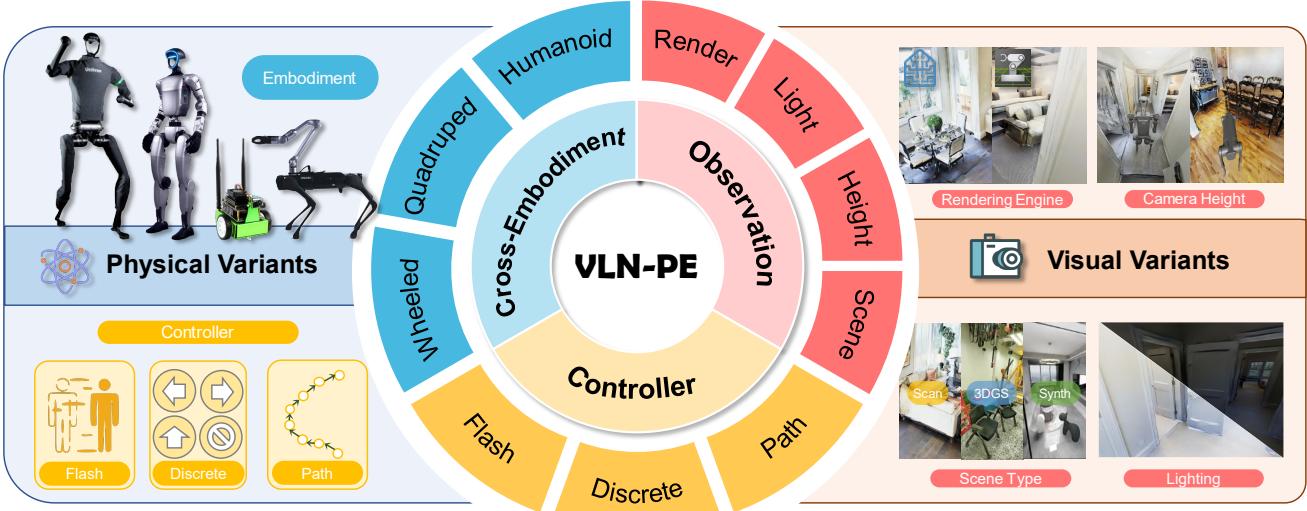


Figure 2. Overview of VLN-PE. This simulator enables researchers to seamlessly integrate various robots and environments, facilitating the exploration of diverse solutions for more physically realistic VLN-related works.

tive approach for continuous waypoint prediction in robot manipulation; and (3) map-based methods [20, 36, 62] allows zero-shot LLM to incorporate semantic understanding and action planning. Despite these advances, the challenge of physically realistic cross-embodiment adaptation has received limited attention until recently. Our work reveals this gap through the VLN-PE benchmark, which provides the first comprehensive analysis of how different physical factors impact these methods, offering valuable insights to guide future VLN research directions.

Language-driven Cross-embodiment Navigation Benchmarks. GRUTopia [54], built on NVIDIA Isaac Sim, introduces a social navigation task that enables interactions between agents and human-like NPCs. Behavior-100 [49] and Behavior-1K [28] focus on various everyday mobile manipulation tasks. Yang *et al.* [61] present a cross-embodiment learning for both manipulation and navigation. ARIOD [59] standardizes datasets with a unified format, rich sensory modalities, and a mix of real and simulated data, enabling large-scale navigation and manipulation across diverse robots. PARTNR [8] provides a benchmark for planning tasks in human-robot collaboration, focusing on studying human-robot coordination in household environments. MO-VLN [32] leverages a 3D simulator based on Unreal Engine 5 to evaluate zero-shot multi-task VLN capabilities with wheeled robots. While these benchmarks contribute valuable datasets, none systematically evaluate VLN methods across diverse robot embodiments, leveraging the full capabilities of these simulators. In this work, we introduce VLN-PE, a novel simulator that supports multiple robot types for data collection, training, and evaluation. By systematically

benchmarking existing ego-centric VLN-CE methods, our study highlights the critical need for more complex and practical VLN settings in future research.

3. VLN-PE Platform and Benchmark

Simulations. Unlike prior studies that rely on animations and predefined positions for pseudo-action execution, VLN-PE is built on a physically realistic simulator, GRUtopia [54], designed to support various robots. It provides RL-based controllers as APIs, enabling the operation of humanoid robots (Unitree H1, G1), quadrupeds (Unitree Aliengo), and wheeled robots (Jetbot). Additionally, thanks to the powerful interactive visualization and rendering capabilities of Isaac Sim, users can easily observe the robot’s movement from custom perspectives.

Scenes. We converted 90 Matterport3D scenes into USD format. During testing, we identified floor gaps caused by reconstruction errors, which could impact robot movement, particularly for legged robots, whose legs might fall into or get stuck in these holes. To address this, we manually fixed all the holes with the help of volunteers. Next, we aligned the coordinates from the VLN-CE (Habitat) benchmark [45] to our platform, ensuring that the initial position and rotation matched the original annotations. We employed disk lighting to enhance lighting conditions, allowing us to adjust light intensity for broader research purposes.

Additionally, since MP3D scenes are reconstructed with limited visual diversity, we introduced two additional scene types to further enhance the experimental environment. First, we incorporated 10 high-quality synthetic home scenes from GRSences [54], which offer significantly im-

proved visual fidelity and physical realism. Second, we included re-built scenes generated using advanced 3D Gaussian techniques [37]. We use the scanned scene of one laboratory environment, which can be observed via online rendering, providing a highly realistic perceptual experience. In theory, many other existing open-sourced scenes can be easily imported into this platform for experimentation.

Datasets. The R2R dataset [2] is the most widely used benchmark for VLN tasks. This dataset is divided into training, validation-seen (same buildings as in training, but with different instructions), and validation-unseen (different buildings from training) splits. We focus on identifying performance gaps using this dataset. Currently, our RL-based locomotion controller does not reliably handle stair navigation in complex environments, so we filter out episodes that include stairs. After stair filtering, the *training/val-seen/val-unseen* splits of R2R remain 8,679/658/1,347 episodes, respectively. For the newly introduced scenes, we sampled trajectories and generated VLN-style instructions using a modular LLM method [17]. After rigorous manual filtering and validation, we created two VLN evaluation datasets: *GRU-VLN10*: 3 scenes for training, 7 for unseen tests, with 441/111/1,287 episodes for training, val-seen, and val-unseen, respectively. *3DGS-Lab-VLN*: A supplementary dataset with 160 training episodes and 640 for evaluation in a 3DGS online rendering environment.

Metrics. Following standard VLN evaluation protocols [2, 25], we use five primary metrics: *Trajectory Length (TL)*, measured in meters; *Navigation Error (NE)*, which quantifies the distance between the predicted and actual stop locations; *Success Rate (SR)*, indicating how often the predicted stop location falls within a predefined distance of the true location; *Oracle Success Rate (OS)*, which assesses the frequency with which any point along the predicted path is within a certain distance of the goal; and *Success Rate weighted by Inverse Path Length (SPL)*, which balances success rate with path efficiency. As physical realism is a key focus of this work, we introduce two more metrics: *Fall Rate (FR)*, which measures the frequency of unintended falls, and *Stuck Rate (StR)*, which quantifies instances where the agent becomes immobilized.

Locomotion Policy. Our used controller APIs are built upon state-of-the-art RL-based locomotion policies [34, 35, 39], specifically designed and optimized for various robot embodiments. Importantly, the same locomotion models used in simulation can be directly applied in the real-robot experiments, ensuring consistency between virtual and physical deployments. This feature enables more practical and scalable experimentation, significantly reducing both time and cost. Additionally, as our codebase is open-source, researchers are encouraged to explore and integrate more advanced locomotion policies within our framework.

4. Baselines

Given that most current robots are equipped with ego-centric RGB-D pinhole cameras, the methods we replicate are designed to align with this constraint. We selected three distinct pipelines for replication and performance comparison: (1) end-to-end training methods, including single-step prediction models (CMA [25], Seq2Seq [25], and NaVid [64]); (2) multi-step prediction models, with our newly proposed baseline—RDP, based on diffusion policy [10]; and (3) a map-based large model retrieval approach (VLMaps) [21] that does not require training.

4.1. End-to-end Train-based Method.

4.1.1. Single-step Discrete Action Classification Methods.

Sequence-to-Sequence (Seq2Seq). The Seq2Seq baseline [25] is a straightforward implementation composed of three components: the instruction encoder, which uses an LSTM to encode GLoVE [40] token embeddings $I = \{w_i\}_{i=1}^L$; the observation encoder, which employs ResNet50 [14] pre-trained for RGB embedding V_t and pre-trained on point-goal navigation for depth D_t , respectively; and a recurrent GRU unit takes the combination of the linear maps of these three inputs and predict the next action a_t . The model can be expressed as follows:

$$h_t = \text{GRU}([V_t, D_t, I], h_{t-1}), \quad (1)$$

$$a_t = \arg \min_a \text{softmax}(W_a h_t + b_a). \quad (2)$$

Cross-modal Attention (CMA). Built based on Seq2Seq, this method incorporates two recurrent networks - one tracking visual observations as Seq2Seq, and the other making decisions based on attended instructions and visual features. Specifically, the first GRU unit works as $h_t^{1st} = \text{GRU}([V_t, D_t, a_{t-1}], h_{t-1}^{1st})$, where $a_{t-1} \in \mathbb{R}^{1 \times 32}$ presents the previous action. Then the instruction features \hat{I}_t attended by the first GRU features are then used to attend to visual (\hat{V}_t) and depth (\hat{D}_t) features using a scaled dot-product attention mechanism:

$$\hat{I}_t = \text{Attn}(I, h_t^{1st}), \quad (3)$$

$$\hat{V}_t = \text{Attn}(V_t, h_t^{1st}), \quad \hat{D}_T = \text{Attn}(D_t, h_t^{1st}). \quad (4)$$

Finally, the second GRU is taken for using the concatenation of these features as inputs and predicting the action:

$$h_t^{2nd} = \text{GRU}([\hat{I}_t, \hat{V}_t, \hat{D}_t, a_{t-1}, h_t^{1st}], h_{t-1}^{2nd}), \quad (5)$$

$$a_t = \arg \min_a \text{softmax}(W_a h_t + b_a). \quad (6)$$

Video-based Large Vision-language Navigation Model (NaVid). Compared with the previous two small specific models, NaVid [64] proposes the first video-based

MLLM for VLN in continuous environments, achieving RGB-only navigation akin to human navigation behavior. Specifically, NaVid is built based on a general-purpose video-based MLLM named LLaMa-VID [30], consisting of a vision encoder, a query generator, an LLM, and two cross-modality projectors. Given the observations up to time t , *i.e.*, a video sequence comprising t frames, NaVid encodes this video to a sequence of tokens via the vision encoder [50] and projects them to a space aligned with language tokens. The special tokens <HIS>, <OBS>, and <NAV> are adopted to demarcate historical, current observations, and the beginning of the LLM to output the actions in linguistic form.

NaVid’s output consists of two variables: an action type, selected from a discrete set, and quantitative arguments for each action. For *Forward*, it predicts the move distance, while for *turn left* and *turn right*, it estimates the rotation degrees. A regular expression parser is used to extract action types and arguments for evaluation and deployment.

4.1.2. Multistep Continuous Prediction Method.

Recurrent Diffusion Policy for VLN (RDP). Recently, diffusion policy [10] has demonstrated impressive trajectory smoothness and stability for manipulation tasks. Later, these strategies were applied to point-goal or image-goal navigation tasks [4, 6, 48], which inspired our exploration of their potential in the VLN domain. This paper explores RDP (Fig. 3) as a new baseline method, using the diffusion generative head to support multistep continuous prediction.

The RDP model takes ego-centric RGB-D input, where RGB and instructions are encoded separately using the LongCLIP [63], and depth is encoded using a pre-trained ResNet50 model, as in CMA. Information across vision and language is exchanged and aligned using two multi-head, multi-layer cross-modal attention modules [53]. The model uses a Transformer structure as the diffusion decoder, taking the fused features c_t as the condition. For the action space, RDP employs continuous relative displacement and yaw angle, represented as a set of continuous ground-truth actions $\{\Delta x_t, \Delta y_t, \Delta \text{yaw}_t\}_{t=1}^T$ for the next T steps. The

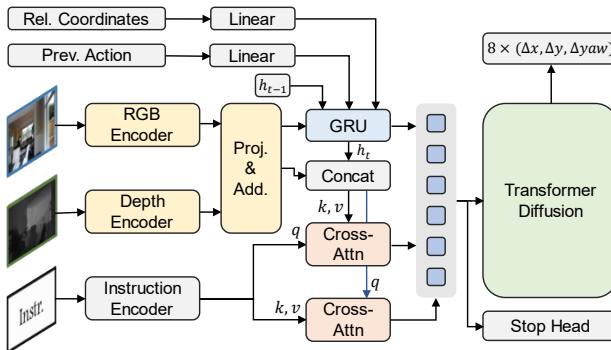


Figure 3. Framework of the recurrent diffusion policy (RDP).

overall training and sampling processes follow DDPM [18]. The iterative denoising process works as follows:

$$a_t^{k-1} = \alpha \cdot (a_t^k - \gamma \epsilon_\theta(c_t, a_t^k, k) + \mathcal{N}(0, \mu^2 I)), \quad (7)$$

where k denotes the number of denoising steps, ϵ_θ is a noise prediction network parameterized by θ , and α , γ , and μ are functions of the noise schedule.

Unlike previous applications of diffusion models in robotic manipulation tasks or local navigation tasks, VLN requires more advanced cross-modal understanding, long-term memory, and reasoning abilities. One of the key challenges in designing this model was how to represent historical information. In this paper, we employ a recurrent GRU structure to maintain and update historical observations, enhancing the model’s ability to capture long-range dependencies. Additionally, we observe that the diffusion model struggle to determine when to stop based on language inputs. To mitigate this, we introduce an additional MLP prediction head for stop progression, which is continuously updated from 0 to 1 during navigation. This stop branch assists the diffusion model in making more precise stopping decisions. The overall loss function is defined as:

$$\mathcal{L}_{RDP} = MSE(\epsilon^k, \epsilon_\theta(c_t, a_t^0 + \epsilon^k, k)) \quad (8)$$

$$+ \lambda \cdot MSE(\mathcal{S}_{stop}(c_t), \hat{p}_{stop}), \quad (9)$$

where \hat{p}_{stop} represents the ground-truth stop progress of each step λ is set to 10 through our experiments. For more details, please refer to our Appendix.

4.2. Modular Map-based Train-free Method.

Improved VLMaps. Train-free, map-based models have demonstrated significant potential, driven by advancements in vision-and-language models with open-set vocabulary capabilities and the versatility of LLMs. A representative work, VLMaps [21] leverages LLMs to parse complex natural language commands into a sequence of code-like subgoals that include object and positional relations (*e.g.*, “robot.move_in_between(‘sofa’, ‘chair’”)). These subgoals are then localized on a semantic map. VLMaps employs LSeg [27] to ground 3D point cloud voxels into semantic embeddings, aligning them with text embeddings of subgoals via contrastive learning. The LLM generates executable Python code, enabling the robot to define functions, logical structures, and parameterized API calls. In our experiments, we observed that some landmarks referenced in VLN instructions may not be visible initially. To address this, we incorporate an exploration policy VLFM [62] for frontier detection. Specifically, when the robot fails to detect the target object from the current pixel embeddings, it performs a turn-around maneuver, moves to each frontier, and evaluates whether the viewpoint is likely to contain the next landmark using the image and text encoders from LSeg. Please refer to our appendix for more details.

Idx	Method	Val Seen								Val Unseen							
		TL↓	NE↓	FR↓	Str↓	OS↑	SR↑	SPL↑	TL↓	NE↓	FR↓	Str↓	OS↑	SR↑	SPL↑		
1	Random	0.14	8.24	0.30	0.00	0.30	0.30	0.30	0.11	7.78	0.74	0.00	3.34	3.04	2.30	<i>Zero-shot transfer evaluation from VLN-CE</i>	
2	Seq2Seq-Full [25]	8.29	7.59	19.15	3.04	17.63	13.83	11.17	8.28	6.99	13.88	3.79	21.83	15.00	11.99		
3	CMA-Full [25]	6.51	7.28	17.93	6.23	18.09	15.50	14.00	6.55	7.02	15.07	4.31	19.82	16.04	14.63		
4	CMA	4.44	7.59	21.12	1.98	13.33	11.32	10.20	4.77	7.39	15.21	3.41	15.23	10.31	10.04		
5	CMA-Height1.8	4.87	7.71	17.17	2.28	14.06	12.87	11.59	5.14	7.41	16.11	2.08	17.02	13.58	12.21		
6	NaVid [64]	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		
<i>Train on VLN-PE</i>																	
7	Seq2Seq	9.89	7.73	22.19	3.04	30.55	19.60	15.67	9.51	7.91	19.67	3.71	27.62	15.89	12.58		
8	CMA	10.23	7.32	23.40	2.43	31.04	21.12	16.15	10.09	7.43	18.63	3.12	31.33	18.78	14.56		
9	RDP	12.00	7.10	22.95	3.95	33.43	23.86	17.35	11.76	6.97	18.75	4.58	34.06	21.98	16.44		
10	Seq2Seq+	9.92	7.54	26.11	5.59	31.93	19.58	15.13	10.21	7.64	21.82	5.12	30.47	18.13	14.06		
11	CMA+	9.09	6.68	20.21	2.74	37.99	28.72	24.24	8.61	7.11	18.63	4.83	31.55	23.31	19.66		
<i>Train-free Map-based Exploration and Navigation</i>																	
12	VLMaps* [21]	-	-	-	-	-	-	-	15.73	6.98	23.00	0.00	20.00	20.00	12.70		

Table 1. Evaluation of existing ego-centric VLN-CE solutions in VLN-PE using the humanoid robot (Unitree H1) with the RL-based controller for implementation on the R2R dataset [2]. The units for TL and NE are meters (m), while all other metrics are expressed as percentages (%). * means that we randomly choose 200 episodes for train-free map-based LLM evaluation.

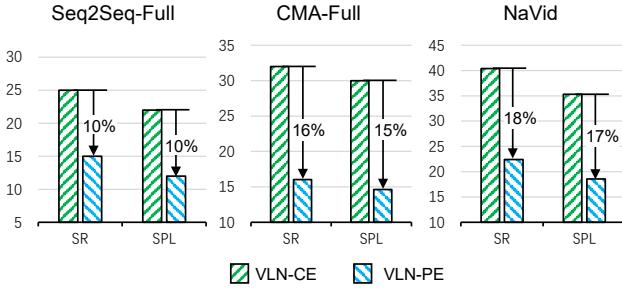


Figure 4. Performance declines in zero-shot testing on the Humanoid H1 robot in VLN-PE on R2R val-unseen.

5. Experiments

In this section, we aim to answer the following questions:

- (1) How well do VLN-CE models transfer to VLN-PE?
- (2) How do physical controllers influence performance?
- (3) How does cross-embodiment data impact adaptation?
- (4) How do observation conditions affect accuracy?

5.1. Evaluation of VLN-CE Solutions in VLN-PE

In Tab. 1, we evaluate ego-centric VLN-CE solutions using the Humanoid Unitree H1 with an RL-based locomotion controller on the R2R dataset [2] under three different conditions. Please note that unless otherwise specified, our experiments are conducted on the R2R dataset.

Zero-shot Performance. As shown in Tab. 1 and Fig. 4, models transferred directly from VLN-CE to VLN-PE experience significant performance drops. Specifically, Seq2Seq-Full (#2), CMA-Full (#3), and NaVid (#6) see SR declines of 10%, 16%, and 18%, respectively. Generally, NaVid exhibits better generalization compared to smaller models. As its training code is unavailable, we

assess its zero-shot performance as a 7B-parameter model. Despite some performance drop, NaVid achieves the highest zero-shot navigation results, highlighting the potential of MLLMs for VLN. Additionally, as a zero-shot, map-based LLM solution, VLMaps (#12) demonstrates reasonable results in the VLN-PE setting. This suggests that, beyond end-to-end action prediction, map-based intelligent navigation could also be a viable approach for VLN.

In-domain Training and Fine-tuning. Notably, Seq2Seq-Full and CMA-Full were trained on Habitat with extensive data augmentation, including dagger-based training and EnvDrop-augmented datasets [51], totaling 175K additional samples. Despite this, they underperform compared to models trained from scratch on in-domain data collected within VLN-PE—without any augmentation (#7 and #8). Interestingly, fine-tuning Seq2Seq and CMA (denoted as Seq2Seq+ and CMA+) with their SoTA weights using the training dataset from VLN-PE significantly improves performance. In particular, CMA+ (#11) surpasses NaVid’s zero shot performance, achieving SR 28.72 vs. 21.58 and SPL 24.24 vs. 17.45 on val-seen. These results indicate that existing VLN models tend to overfit to specific simulation platforms, resulting in poor generalization when directly transferred to new environments or settings. However, incorporating diverse domain data can further enhance overall navigation performance. Without additional augmentation, #5 explores the impact of observation height, aligning the agent’s camera height with H1, which improves transferability compared to #4 with the default height of 1.2m. Comparing #9 and #7, #8, we find that RDP outperforms CMA and Seq2Seq when trained from scratch. This marks the first application of continuous dense low-level offset prediction in VLN, highlighting diffusion policy-based ap-

Idx	Method	Val Seen							Val Unseen						
		TL↓	NE↓	FR↓	StR↓	OS↑	SR↑	SPL↑	TL↓	NE↓	FR↓	StR↓	OS↑	SR↑	SPL↑
1	NaVid (ZS)	4.88	5.05	27.02	0.00	30.63	25.23	22.24	5.86	4.75	10.87	0.60	23.73	18.64	13.99
2	CMA-CLIP w/o FT	5.41	5.91	68.47	6.31	18.02	10.81	9.94	6.47	5.51	55.40	12.28	22.30	15.31	13.12
3	CMA-CLIP	8.95	4.50	23.42	0.90	57.66	31.53	27.52	9.41	5.41	27.20	5.75	46.15	22.46	17.93
4	RDP w/o FT	11.24	5.83	57.66	10.81	36.04	18.02	12.73	13.02	5.31	34.45	9.44	51.81	26.19	18.70
5	RDP	10.51	4.96	22.52	5.41	48.65	32.43	27.74	9.70	4.93	13.99	5.75	45.30	28.52	22.53

Table 2. Evaluation of end-to-end methods on the GRU-VLN10 dataset using the Humanoid robot to assess model generalization in an out-of-MP3D-domain setting. ZS means the zero-shot implementation. “w/o FT” means without fine-tuning in the specific scenes.

Idx	Method	3DGS-Lab-VLN							Val Unseen						
		TL↓	NE↓	FR↓	StR↓	OS↑	SR↑	SPL↑	TL↓	NE↓	FR↓	StR↓	OS↑	SR↑	SPL↑
1	NaVid (ZS)	1.44	5.70	4.65	0.00	6.20	5.81	1.00							
2	CMA-CLIP w/o FT	12.06	6.40	42.75	2.54	48.78	16.72	10.66							
3	CMA-CLIP	8.46	5.38	11.74	0.63	30.67	24.88	17.43							
4	RDP w/o FT	18.36	5.63	40.39	1.08	53.78	26.78	12.17							
5	RDP	9.85	4.73	17.97	0.63	34.84	30.63	22.69							

Table 3. Performance on the 3DGS-Lab-VLN val-unseen dataset.

proaches as a promising research direction.

Collision Avoidance. Another notable observation is that the MLLM-based method, NaVid, exhibits significantly lower StR and FR compared to all other methods. This is particularly interesting as it highlights the potential of MLLMs in addressing robotic challenges, such as autonomous obstacle avoidance and deadlock recovery. This advantage can be attributed to the world knowledge capabilities of large models, which allow NaVid to make context-aware decisions about its environment, helping it avoid falls and deadlocks. However, NaVid also has its limitations. We observed that in 70% of episodes, it tends to rotate continuously near the goal for over 25 steps before issuing a stop signal, rather than proactively stopping when reaching the target. This suggests that large models may still struggle with precise goal recognition, indicating an area for future improvement in MLLM-based navigation.

Out-of-MP3D-domain Study. Since most VLN methods have been developed and evaluated within MP3D environments, we introduce 10 high-quality indoor scenes from GRUScenes [54] to expand the evaluation scope. To address the limited vocabulary size in the original CMA model, we replace its instruction encoder with Long-CLIP [63], denoted as CMA-CLIP. As shown in Tab. 2, fine-tuning small models (#3 and #5) using just 441 episodes from 3 scene leads to a significant performance boost compared to NaVid’s zero-shot results (e.g., RDP improves SR by 18.86% and SPL by 17.66% on average). Surprisingly, on the 3DGS-Lab-VLN dataset (Tab. 3), NaVid completely fails, continuously rotating and achieving only 5.81 SR. We suspect this is due to rendering noise introduced by 3DGS, which, while imperceptible to humans, may significantly disrupt the large model’s RGB perception. It is important to note that these results do not suggest that small models are sufficient, but rather highlight the substantial room for improvement in current MLLM-based navigation models.

5.2. Impact of Physical Controller

One key advantage of our platform is its support for realistic physical simulation across multiple robot embodiments, particularly legged robots such as humanoid and quadruped robots. This raises an important research question: *Can training a VLN model on data collected with physical controller engagement—where robots experience natural walking dynamics and perceptual feedback from motion disturbances—enhance adaptation and performance on legged robots?* To explore this, we conducted comparative experiments using the Humanoid robot in VLN-PE.

Collect w/ loco	Eval w/ loco	Val Seen							Val Unseen						
		FR↓	StR↓	OS↑	SR↑	SPL↑	FR↓	StR↓	OS↑	SR↑	SPL↑				
✗	✗	0.00	0.00	39.51	20.21	15.07	0.00	0.00	38.60	21.31	15.89				
✓	✗	0.00	0.00	50.30	18.54	12.91	0.00	0.00	43.76	15.90	10.54				
✗	✓	31.76	4.56	18.39	12.92	11.26	25.69	4.75	19.67	13.51	11.54				
✓	✓	23.40	2.43	31.04	21.12	16.15	18.63	3.12	31.33	18.78	14.56				

Table 4. Comparison of the controller engagement for the data collecting and evaluation in training CMA from scratch.

As shown in Tab. 4, the results demonstrate that the model performance is highest when the physical locomotion controller used during data collection and evaluation remains consistent. Conversely, when the controllers differ between training and evaluation, performance drops significantly. Additionally, training with data collected using a controller helps mitigate robot falls and stuck to some degrees, improving overall stability. This demonstrates that for legged robots, the controller-based data collection pipeline provided by our simulator proves to be crucial for achieving more reliable performance.

5.3. Impact of Cross-embodiment Data

Fig. 5 evaluates NaVid’s zero-shot performance on different robot types, revealing varying degrees of performance degradation. A key advantage of VLN-PE is its ability to seamlessly support diverse robot models for navigation and data collection. This raises our interest: *Can cross-robot data improve model training, enabling a unified model that generalizes across different robot embodiments (“One-for-All”)?* To explore this, we collected R2R data from humanoid, quadrupedal, and wheeled robots within VLN-PE. Tab. 5 presents key findings from this study.

(1) VLN models are highly sensitive to motion dynamics and camera height. When directly transferring mod-

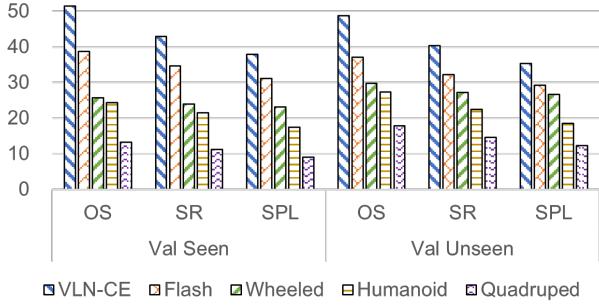


Figure 5. Performance of NaVid across different robot platforms.

els to our platform, the best performance is observed on the Humanoid’s flash setting, whose smooth movement is most similar to that in the original Habitat platform. The quadruped robot (about 0.5 m), with a significantly lower camera height, causes the model to nearly fail completely.

(2) Robot-specific viewpoint data enhances model adaptation. Fine-tuning on platform-collected data significantly improves OSR, increasing from 19.82 to 31.33 for humanoid robots and 7.84 to 32.37 for quadrupeds on val-unseen. This underscores the importance of specialized training data for adapting VLN models to different embodiments. Additionally, with limited data, smaller models struggle with accurate stopping—a key challenge in VLN.

(3) Cross-embodiment training significantly enhances overall performance and enables a *One-for-All* model. As indicated by the bolded results, models trained using a combination of data from all three robot types consistently achieve the best performance. This improvement can be attributed both to increased data volume and to the benefits of multi-view learning, which enhances the model’s understanding of both the environment and its own embodiment.

Robot	Training Data				Val Seen			Val Unseen		
	VLN-CE	Humanoid	Quadruped	Wheeled	OS	SR	SPL	OS	SR	SPL
Humanoid	✓				18.09	15.50	14.00	19.82	16.04	14.63
		✓			31.04	21.12	16.15	31.33	18.78	14.56
	✓	✓	✓	✓	35.26	22.64	18.07	31.70	19.30	16.97
Quadruped	✓				32.37	26.44	22.76	34.08	26.87	23.54
		✓			4.96	2.07	1.69	7.84	4.73	3.80
	✓	✓	✓	✓	31.61	17.93	14.34	32.37	17.00	13.40
Wheeled	✓				32.11	21.88	18.84	31.72	17.15	14.68
	✓	✓	✓	✓	30.40	24.47	20.93	29.62	23.83	20.75
		✓			15.32	12.61	12.61	13.75	11.02	10.80
Flash		✓			16.98	15.09	14.58	14.63	12.01	11.78
	✓	✓	✓	✓	21.51	19.35	18.71	21.44	15.83	14.87
		✓			23.66	20.50	19.53	22.71	20.02	19.38

Table 5. Comparison of the CMA model using different combination of cross-embodiment training data.

5.4. Impact of Lighting Conditions

Omniverse Kit-based applications provide versatile lighting options to simulate various environments. We use three settings: Disc Light (DL) – simulating panel lighting, with



Figure 6. Environments under different light conditions.

Method	Light	Val Seen			Val Unseen		
		OS↑	SR↑	SPL↑	OS↑	SR↑	SPL↑
NaVid (ZS)	DL5000	24.32	21.58	17.45	27.32	22.42	18.58
	DL300	17.02	14.59	12.19	12.47	9.95	9.01
	CL	21.23	18.38	16.10	14.27	11.17	9.34
CMA	DL5000	31.04	21.12	16.15	31.33	18.78	14.56
	DL300	30.06	19.02	15.55	30.49	17.37	15.34
	CL	28.23	19.13	14.45	29.02	17.04	15.28
RDP	DL5000	33.43	23.86	17.35	34.06	21.98	16.44
	DL300	29.94	20.52	15.71	29.40	22.27	17.15
	CL	29.87	21.06	16.89	30.87	22.78	16.23

Table 6. Comparison of using different light conditions.

intensity set to 5000 for daytime and 300 for nighttime; and Camera Light (CL) – simulating light from the robot. This may cause reflections on the front wall, leaving the sides darker. As shown in Tab. 6, NaVid suffers the largest performance drop when lighting deviates from ideal conditions, with SR decreasing by 12.47% under DL300 and 11.25% under CL on val-unseen. In contrast, CMA and RDP are less affected. The key reason behind this discrepancy lies in modality reliance: NaVid relies solely on RGB input, making it highly susceptible to lighting variations, whereas CMA and RDP utilize both RGB and depth, allowing them to maintain stability when RGB quality degrades. This underscores the need to integrate multimodal environmental information, such as depth or radar, to improve model generalization and robustness.

6. Conclusion

We introduce VLN-PE, a realistic VLN platform and benchmark designed to enhance physical deployment across diverse robot embodiments. It enables cross-embodiment data collection, evaluation, and optimization under realistic locomotion and environmental conditions. Through systematic experiments on VLN methods based on the egocentric pinhole cameras, we expose critical physical and visual disparities that challenge existing approaches and benchmarks. VLN-PE offers a grounded framework to foster more generalizable VLN models for future physical embodied AI development.

Acknowledgments

We sincerely thank all our collaborators for their valuable support and contributions. This paper is supported by the National Natural Science Foundation of China under Grants (624B2105, 62473295, 62233013, 62333017). This work is funded in part by the Na-

tional Key R&D Program of China (2022ZD0160201), and Shanghai Artificial Intelligence Laboratory.

References

- [1] 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. [1](#), [2](#), [15](#)
- [2] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683, 2018. [1](#), [4](#), [6](#), [13](#)
- [3] Peter Anderson, Ayush Shrivastava, Joanne Truong, Arjun Majumdar, Devi Parikh, Dhruv Batra, and Stefan Lee. Sim-to-real transfer for vision-and-language navigation. In *Conference on Robot Learning*, pages 671–681. PMLR, 2021. [2](#), [15](#)
- [4] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. *arXiv preprint arXiv:2412.03572*, 2024. [5](#)
- [5] Qingwei Ben, Feiyu Jia, Jia Zeng, Junting Dong, Dahua Lin, and Jiangmiao Pang. Homie: Humanoid locomanipulation with isomorphic exoskeleton cockpit. *arXiv preprint arXiv:2502.13013*, 2025. [2](#)
- [6] Wenzhe Cai, Jiaqi Peng, Yuqiang Yang, Yujian Zhang, Meng Wei, Hanqing Wang, Yilun Chen, Tai Wang, and Jiangmiao Pang. Navdp: Learning sim-to-real navigation diffusion policy with privileged information guidance. *arXiv preprint arXiv:2505.08712*, 2025. [2](#), [5](#)
- [7] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. [2](#)
- [8] Matthew Chang, Gunjan Chhablani, Alexander Clegg, Mikael Dallaire Cote, Ruta Desai, Michal Hlavac, Vladimir Karashchuk, Jacob Krantz, Roozbeh Mottaghi, Priyam Parashar, et al. Partnr: A benchmark for planning and reasoning in embodied multi-agent tasks. *arXiv preprint arXiv:2411.00081*, 2024. [3](#)
- [9] An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Xueyan Zou, Jan Kautz, Erdem Biyik, Hongxu Yin, Sifei Liu, and Xiaolong Wang. Navila: Legged robot vision-language-action model for navigation. *arXiv preprint arXiv:2412.04453*, 2024. [2](#)
- [10] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023. [2](#), [4](#), [5](#)
- [11] Yibo Cui, Liang Xie, Yakun Zhang, Meishan Zhang, Ye Yan, and Erwei Yin. Grounded entity-landmark adaptive pre-training for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12043–12053, 2023. [2](#)
- [12] Ronghao Dang, Lu Chen, Liuyi Wang, He Zongtao, Chengju Liu, and Qijun Chen. Multiple thinking achieving metaability decoupling for object navigation. In *International Conference on Machine Learning (ICML)*, 2023. [2](#)
- [13] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihns, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation. *Advances in Neural Information Processing Systems*, 35:5982–5994, 2022. [1](#)
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. [4](#)
- [15] Zongtao He, Liuyi Wang, Ronghao Dang, Shu Li, Qingqing Yan, Chengju Liu, and Qijun Chen. Learning depth representation from rgb-d videos by time-aware contrastive pre-training. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2023. [2](#)
- [16] Zongtao He, Liuyi Wang, Lu Chen, Shu Li, Qingqing Yan, Chengju Liu, and Qijun Chen. Multimodal evolutionary encoder for continuous vision-language navigation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1443–1450. IEEE, 2024. [2](#)
- [17] Zongtao He, Liuyi Wang, Lu Chen, Chengju Liu, and Qijun Chen. Navcomposer: Composing language instructions for navigation trajectories through action-scene-object modularization. *arXiv preprint arXiv:2507.10894*, 2025. [4](#), [13](#)
- [18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. [5](#), [12](#)
- [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. [2](#)
- [20] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10608–10615. IEEE, 2023. [3](#)
- [21] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, London, UK, 2023. [2](#), [4](#), [5](#), [6](#), [12](#)
- [22] 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. [2](#)
- [23] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. [2](#)
- [24] 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. 2, 15
- [25] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *Computer Vision – ECCV 2020*, pages 104–120. Springer International Publishing. 1, 2, 4, 6, 12, 13
- [26] 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. 1
- [27] Boyi Li, Kilian Q. Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation, 2022. 5
- [28] Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Wensi Ai, Benjamin Martinez, et al. Behavior-1k: A human-centered, embodied ai benchmark with 1,000 everyday activities and realistic simulation. *arXiv preprint arXiv:2403.09227*, 2024. 3
- [29] Jialu Li and Mohit Bansal. Improving vision-and-language navigation by generating future-view image semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10803–10812, 2023. 2
- [30] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024. 5
- [31] Jing Liang, Amirreza Payandeh, Daeun Song, Xuesu Xiao, and Dinesh Manocha. Dtg: Diffusion-based trajectory generation for mapless global navigation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5340–5347. IEEE, 2024. 2
- [32] Xiwen Liang, Liang Ma, Shanshan Guo, Jianhua Han, Hang Xu, Shikui Ma, and Xiaodan Liang. Mo-vln: A multi-task benchmark for open-set zero-shot vision-and-language navigation. *Computing Research Repository arXiv Preprints*, 2023. 3
- [33] Xiwen Liang, Liang Ma, Shanshan Guo, Jianhua Han, Hang Xu, Shikui Ma, and Xiaodan Liang. Cornav: Autonomous agent with self-corrected planning for zero-shot vision-and-language navigation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12538–12559, 2024. 2
- [34] Junfeng Long, Junli Ren, Moji Shi, Zirui Wang, Tao Huang, Ping Luo, and Jiangmiao Pang. Learning humanoid locomotion with perceptive internal model. *arXiv preprint arXiv:2411.14386*, 2024. 2, 4
- [35] Junfeng Long, Zirui Wang, Quanyi Li, Liu Cao, Jiawei Gao, and Jiangmiao Pang. Hybrid internal model: Learning agile legged locomotion with simulated robot response. In *ICLR*, 2024. 4
- [36] 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. 1, 3
- [37] Tao Lu, Mulin Yu, Lining Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20654–20664, 2024. 4
- [38] Abhinav Moudgil, Arjun Majumdar, Harsh Agrawal, Stefan Lee, and Dhruv Batra. Soat: A scene-and object-aware transformer for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 34:7357–7367, 2021. 2
- [39] Guoping Pan, Qingwei Ben, Zhecheng Yuan, Guangqi Jiang, Yandong Ji, Shoujie Li, Jiangmiao Pang, Houde Liu, and Huazhe Xu. Roboduet: Learning a cooperative policy for whole-body legged loco-manipulation. *IEEE Robotics and Automation Letters*, 2025. 2, 4
- [40] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 4
- [41] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9982–9991, 2020. 1
- [42] Yanyuan Qiao, Yuankai Qi, Yicong Hong, Zheng Yu, Peng Wang, and Qi Wu. Hop+: History-enhanced and order-aware pre-training for vision-and-language navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2
- [43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 12
- [44] 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. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. Curran, 2021. 2
- [45] Sonia Raychaudhuri, Saim Wani, Shivansh Patel, Unnat Jain, and Angel X Chang. Language-aligned waypoint (law) supervision for vision-and-language navigation in continuous environments. *arXiv preprint arXiv:2109.15207*, 2021. 3
- [46] 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. 1
- [47] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting

- grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749, 2020. 1
- [48] 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. 2, 5, 12
- [49] Sanjana Srivastava, Chengshu Li, Michael Lingelbach, Roberto Martín-Martín, Fei Xia, Kent Elliott Vainio, Zheng Lian, Cem Gokmen, Shyamal Buch, Karen Liu, et al. Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. In *Conference on robot learning*, pages 477–490. PMLR, 2022. 3
- [50] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 5
- [51] Hao Tan, Licheng Yu, and Mohit Bansal. Learning to navigate unseen environments: Back translation with environmental dropout. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2610–2621, 2019. 6
- [52] Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. Vision-and-dialog navigation. In *Conference on Robot Learning*, pages 394–406. PMLR, 2020. 1
- [53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 5
- [54] Hanqing Wang, Jiahe Chen, Wensi Huang, Qingwei Ben, Tai Wang, Boyu Mi, Tao Huang, Siheng Zhao, Yilun Chen, Sizhe Yang, et al. Grutopia: Dream general robots in a city at scale. *arXiv preprint arXiv:2407.10943*, 2024. 2, 3, 7
- [55] Liuyi Wang, Zongtao He, jiagui Tang, Ronghao Dang, naijia Wang, Chengju Liu, and Qijun Chen. A dual semantic-aware recurrent global-adaptive network for vision-and-language navigation. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2023. 2
- [56] Liuyi Wang, Jiagui Tang, Zongtao He, Ronghao Dang, Chengju Liu, and Qijun Chen. Vision-and-language navigation via causal learning. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2
- [57] 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. 2
- [58] Zihan Wang, Xiangyang Li, Jiahao Yang, Yeqi Liu, and Shuqiang Jiang. Sim-to-real transfer via 3d feature fields for vision-and-language navigation. *arXiv preprint arXiv:2406.09798*, 2024. 2
- [59] Zhiqiang Wang, Hao Zheng, Yunshuang Nie, Wenjun Xu, Qingwei Wang, Hua Ye, Zhe Li, Kaidong Zhang, Xuewen Cheng, Wanxi Dong, et al. All robots in one: A new standard and unified dataset for versatile, general-purpose embodied agents. *arXiv preprint arXiv:2408.10899*, 2024. 3
- [60] Lining Xu, Yuanbo Xiangli, Sida Peng, Xingang Pan, Nanxuan Zhao, Christian Theobalt, Bo Dai, and Dahua Lin. Grid-guided neural radiance fields for large urban scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8296–8306, 2023. 2
- [61] Jonathan Yang, Catherine Glossop, Arjun Bhorkar, Dhruv Shah, Quan Vuong, Chelsea Finn, Dorsa Sadigh, and Sergey Levine. Pushing the limits of cross-embodiment learning for manipulation and navigation. *arXiv preprint arXiv:2402.19432*, 2024. 3
- [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. 3, 5, 13
- [63] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision*, pages 310–325. Springer, 2024. 5, 7, 12
- [64] 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. *Robotics: Science and Systems*, 2024. 1, 2, 4, 6, 12, 15
- [65] Gengze Zhou, Yicong Hong, and Qi 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. 2

Appendix

A. Implementation Details

Since the most commonly used mobile robots are equipped with ego-centric RGB-D pinhole cameras currently, we primarily evaluate VLN methods without panoramic views in this work. For classic end-to-end single-step discrete action prediction models (Seq2Seq [25], CMA [25], and NaVid [64]), we directly use their publicly available code and pre-trained weights. For the other two model types—the end-to-end continuous multi-step prediction model (RDP) and the map-based LLM model (VLMaps [21])—we introduce several modifications, which are detailed in this appendix.

A.1. Recurrent Diffusion Policy for VLN

The RDP model takes ego-centric RGB-D observations and language instructions as inputs. Since some instructions exceed the 77-word limit in standard CLIP [43], LongCLIP [63] is used as both the RGB and instruction encoders. The depth encoder follows CMA, using ResNet50 pre-trained on point-goal tasks. Each RGB image is represented by five tokens: The first token encodes the global feature, while the remaining four tokens capture semantic information via grid pooling [64]. The flattened depth features are added to the first RGB token, resulting in a fused visual feature dimension of $\mathbb{R}^{5 \times h_d}$, where $h_d = 512$. To improve progress awareness, we incorporate previous 4-step actions (PA) and relative coordinates (RC) from the starting point, both represented in $(\Delta x, \Delta y, \Delta \text{yaw})$. The key difference is that PA encodes the last four steps relative to the current position, while RC represents the current position relative to the starting point.

For historical observation encoding, initially, we experimented with a video-based format, similar to NaVid, where stacked images provided long-term sequence information. However, this approach led to rapid convergence of the diffusion process to small losses, causing severe overfitting. Through our experimentation, we found that employing a recurrent GRU structure to maintain and update historical observations improved generalization:

$$h_t = \text{GRU}([V_c, RC, PA], h_{t-1}). \quad (10)$$

Then, we apply two cross-attention mechanisms to align attended vision ($q = \text{Concat}(h_t, V_c)$) and language features $I = \{w_i\}_{i=1}^L$, where each modality serves as the key and value for the other:

$$g_1 = \text{CrossAttn}(q, I, I), \quad g_2 = \text{CrossAttn}(I, q, q). \quad (11)$$

Finally, the condition feature for the diffusion model is formed by concatenating all extracted features:

$$c_t = \text{Concat}(g_1, g_2, h_t, RC, PA). \quad (12)$$

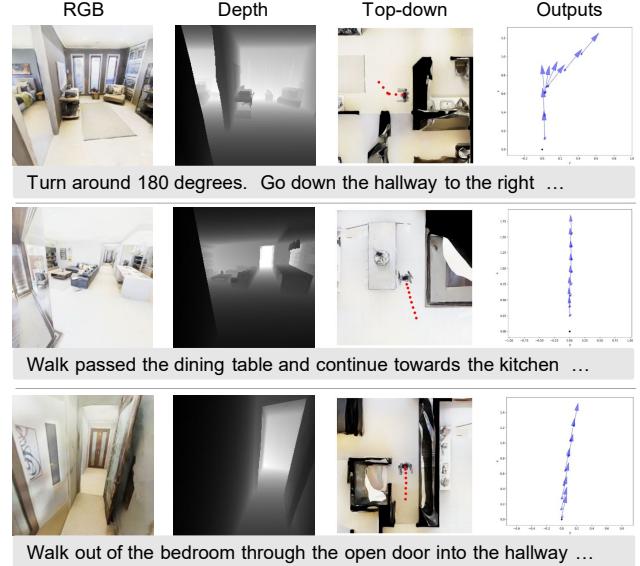


Figure 7. Examples of the robot observations and RDP outputs.

We employ a transformer-based diffusion module [48] with one encoder layer and three decoder layers. During training, the ground-truth trajectory coordinates $T \times (\Delta x, \Delta y, \Delta \text{yaw})$ are perturbed with random noise, and the network is trained to predict and remove this noise. The iterative denoising process follows DDPM [18]. Additionally, we introduce a self-attention-based stop prediction head to determine the current stop progress (from 0 to 1). The stop signal is triggered if: All predicted actions from the diffusion head are below the threshold 0.1, or the stop progress output exceeds 0.8. The output of the RDP is shown in Fig. 7. During navigation, RDP predicts 8 future trajectory waypoints and executes 4 steps per iteration.

In our experiments, RDP demonstrated improvements over the previous baseline models (Seq2Seq and CMA) when trained from scratch. However, there remains significant potential for further enhancement. As this paper primarily focuses on the new physical VLN platform (VLN-PE), we introduce RDP as a baseline method for predicting trajectory waypoints, which can be further integrated with control-theoretic approaches like the Model Predictive Control (MPC) framework to enhance motion smoothness, addressing the jerky transitions seen in discrete action-based methods. We hope this work can inspire and support some future research in this direction.

A.2. Improved VLMaps

VLMaps differs from traditional end-to-end models by utilizing a spatial semantic map representation that directly integrates pre-trained vision-language features of the physical world. This approach enables natural language-based map indexing without requiring additional labeled data. Therefore, we chose this method as one of the technical pipelines

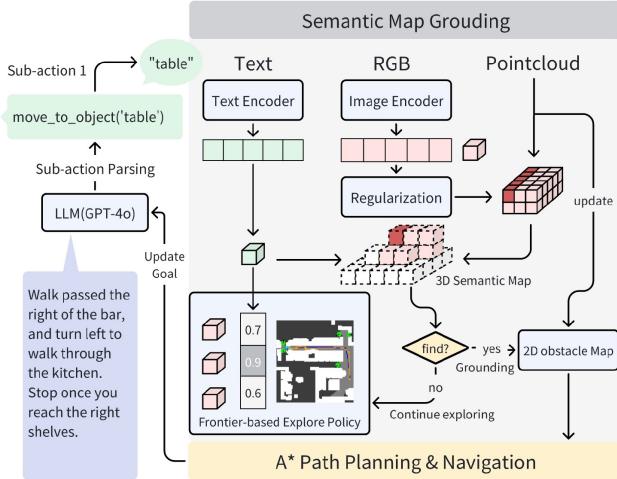


Figure 8. Framework of the improved VLMaps.

for evaluation. However, the original VLMaps lacks a direct exploration policy and struggles with room-level descriptions (*e.g.*, “enter the living room”), which require an agent-oriented perspective rather than reliance on a global semantic map. To address these limitations, we improved the VLMaps framework (as shown in Fig. 8) with two key enhancements.

Exploration Policy: Inspired by VLFM [62], we implement a frontier detection strategy, where a frontier is defined as the boundary between explored and unexplored areas. When the robot fails to detect the target object from the current pixel embeddings, it performs a turn-around maneuver, moves to each frontier, and evaluates whether the viewpoint is likely to contain the next landmark using the image and text encoders from LSeg. For instance, we observed that “table” exhibits a higher similarity score with scenes of “dining room” compared to “toilet,” validating the policy’s ability to guide the robot toward plausible directions.

Room-Level Descriptions: Similarly, we leverage the CLIP module from LSeg as a classifier to assess whether the viewpoint aligns with the intended room context. Specifically, we use a predefined set of room names (“living room,” “dining room,” “bedroom,” “kitchen,” “toilet,” “others”) as text inputs to index the current RGB image. Upon successful room detection, we naturally incorporate actions such as `self.move_to_room('room_name')`.

In the VLN-PE, we apply additional techniques to reduce the fall rate and stuck rate. For example, we implement an A* algorithm as the local planner, assigning higher costs to dilated and unexplored areas. When executing commands like `self.move_forward(1)`, the robot may collide with obstacles if not properly oriented. To address, we define a cost function to identify the optimal node n^* from the robot’s perspective: $n^* = \arg \min_n (\|dist(n, x_0) - dist(x_g, x_0)\| + \alpha\gamma)$, where x_g is the goal position, x_0 is the

current position, $\alpha = 0.25$ is a weight parameter, and γ is the angle required to face the target. This ensures the robot slightly reorients itself before moving forward, minimizing collision risks. An example of the improved VLMaps is shown in Fig. 9. Compared to end-to-end methods, map-based modular approaches offer more explainable and reliable results. However, their performance heavily depends on mapping and localization accuracy, which could limit the practical deployment.

A.3. Experimental Details

All training experiments are conducted using NVIDIA RTX 4090 GPUs. The CMA and Seq2Seq models are trained on a single GPU with a batch size of 2, requiring approximately one day to converge. The RDP model is trained on 4 GPUs using PyTorch’s DataParallel module, with a total batch size of 8, and completes training in around two days. All models are optimized using the AdamW optimizer with a learning rate of 1×10^{-4} . The maximum trajectory length is set to 200. For evaluation, the CMA model requires approximately 4 hours to complete a full evaluation on the R2R-CE benchmark when run in parallel on 8 GPUs.

A.4. Datasets

The trajectories sampling strategy for the newly introduced datasets (GRU-VLN10 and 3DGS-Lab-VLN) is as follows: (a) generate a freemap, (b) randomly sample start-goal pairs, and (c) filter out invalid paths (overly short, long, or similar ones). Instructions are generated via a modular pipeline [17] with action and environment recognition, GPT-4 in-context description, and human refinement. Comparisons of datasets are presented in Fig. 10.

A.5. Metrics

Metrics. Following standard VLN evaluation protocols [2, 25], we use five primary metrics: *Trajectory Length (TL)*, measured in meters; *Navigation Error (NE)*, which quantifies the distance between the predicted and actual stop locations; *Success Rate (SR)*, indicating how often the predicted stop location falls within a predefined distance of the true location; *Oracle Success Rate (OS)*, which assesses the frequency with which any point along the predicted path is within a certain distance of the goal; and *Success Rate weighted by Inverse Path Length (SPL)*, which balances success rate with path efficiency. As physical realism is a key focus of this work, we introduce two more metrics: *Fall Rate (FR)*, which measures the frequency of unintended falls, and *Stuck Rate (StR)*, which quantifies instances where the agent becomes immobilized. Specifically, “Fall” is the robot having a roll $> 15^\circ$ or pitch $> 35^\circ$, or a center-of-mass-to-foot height below a robot-specific threshold. “Stuck” is defined as both position and heading change $< 0.2\text{m}$ and 15° for 50 steps.

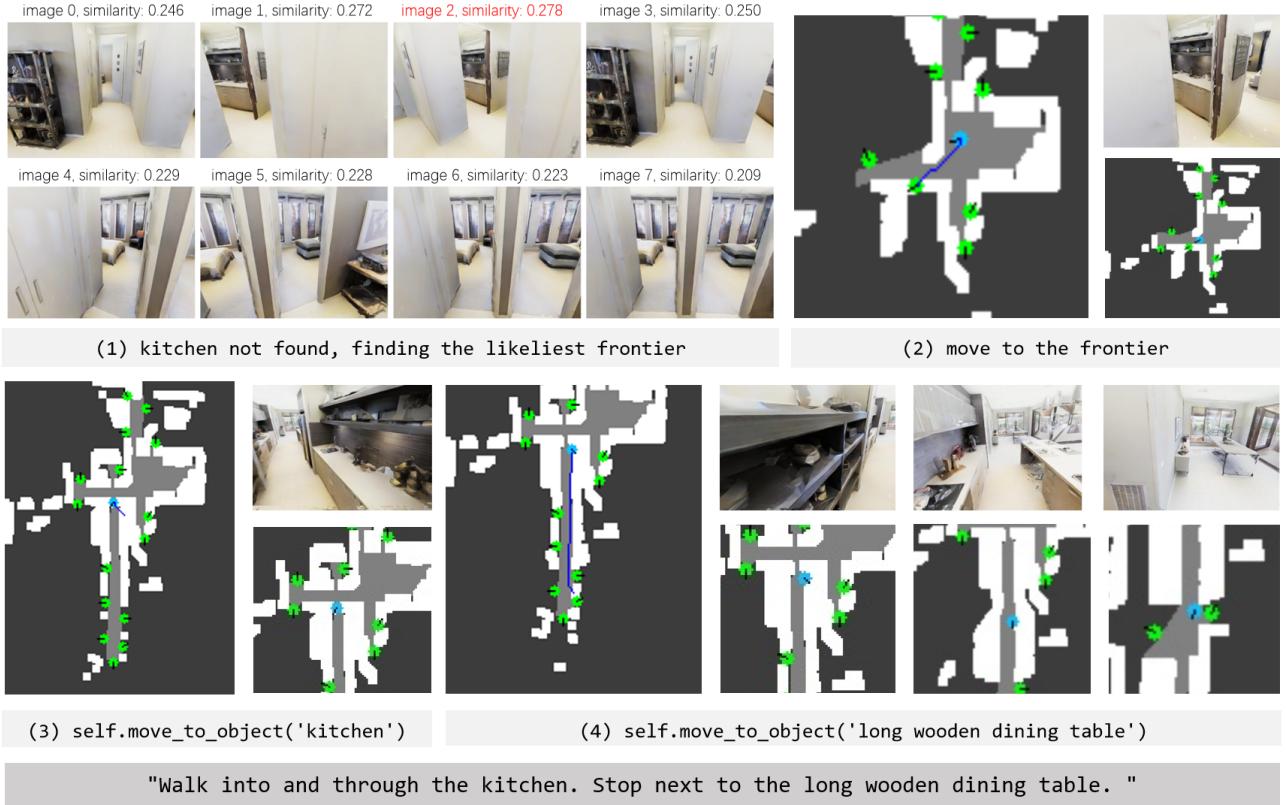


Figure 9. Example of improved VLMaps. Blue dot: current position (black line: orientation). Green dot: frontiers (black line: exploration orientation). White: dilated obstacles. Light gray: explored area. Dark gray: unexplored area. Blue line: local planner trajectory.

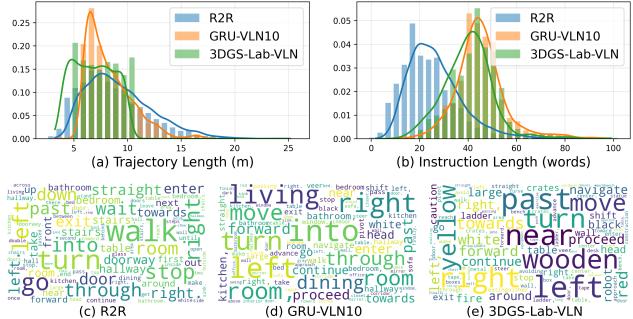


Figure 10. Comparison of distributions across datasets.

A.6. Controllers

Thanks to NVIDIA Isaac Sim’s advanced physical simulation capabilities, we can seamlessly apply various control theories, making the low-level control policy more diverse and aligned with real-world robotic applications. In this work, we utilize three types of controllers for experimentation: flash control, move-by-speed control, and move-along-path control.

- Flash Control: This mechanism mimics platforms that lack physical cross-embodiment support, allowing the

agent to instantly reach the target position without considering physical motion constraints.

- Move-by-Speed Control: This method simulates realistic motion dynamics by controlling the agent’s velocity using linear and angular speed commands. For legged humanoid and quadruped robots, we employ the RL-based policies to regulate movement, ensuring the robot follows the required forward and rotational speeds. For wheeled robots, we use a differential drive controller to manage navigation. For end-to-end models, we implement discrete actions using this controller.
- Move-along-Path Control: This approach enables the agent to follow a predefined trajectory, replicating path-following behaviors in robotic navigation. For the Map-based method (VLMaps), we apply the A* path planning algorithm and use this controller with a PID system to ensure smooth trajectory following.

A.7. Fine-tune on the specific datasets

To better evaluate out-of-MP3D-style domain generalization, we collect additional VLN datasets using GRUScenes and 3DGS-rendered environments. Since these datasets are primarily used for evaluation, only a small portion of the

data is allocated for training, while the majority is reserved for testing. For CMA and RDP, all training experiments use a learning rate of 1e-4 with a cosine learning schedule. In Tab. 2 and Tab. 3, “w/o FT” refers to direct zero-shot transfer using VLN-PE-R2R-trained weights for evaluation without further in-domain fine-tuning in these new scenes. On the GRU-VLN10 dataset, small models significantly improved after 10 epochs of fine-tuning, whereas the SoTA large model NaVid showed limited zero-shot performance. This highlights the limited diversity of existing VLN benchmarks, which could not fully assess model generalization.

B. Impact of Sim-to-Real Transfer

Compared to traditional VLN simulators and platforms, VLN-PE introduces a significant advancement by supporting physical VLN across diverse robot types, enabling data collection, training, and closed-loop evaluation in physical settings. We begin by identifying the limitations of existing VLN algorithms when deployed in physical environments, as initially verified within a physics-enabled simulator. After fine-tuning on data collected through VLN-PE, we observe consistent performance improvements within the simulated physical setup. Encouraged by these results, we further evaluate our approach in real-world settings. Specifically, we conduct experiments using a Unitree Go2 robot equipped with an Intel RealSense D455 RGB-D camera across 14 indoor episodes (see Table 7 and Fig. 11). The model fine-tuned with VLN-PE demonstrates improved adaptation and generalization, confirming the practical effectiveness of our platform.

In particular, we observe that the CMA baseline model, after VLN-PE fine-tuning, exhibits more confident forward movement and better semantic grounding during navigation. In contrast, the CMA Full baseline trained solely on VLN-CE struggles in real-world conditions, frequently resulting in aimless rotation and poor generalization. One notable remaining challenge is the handling of the stop action. CMA often fails to robustly predict when to stop. To mit-

Method	Fine-tuned on VLN-PE	OS↑	SR↑
CMA	✗	14.29 57.14	7.14 28.57

Table 7. Impact of VLN-PE on real-world performance.

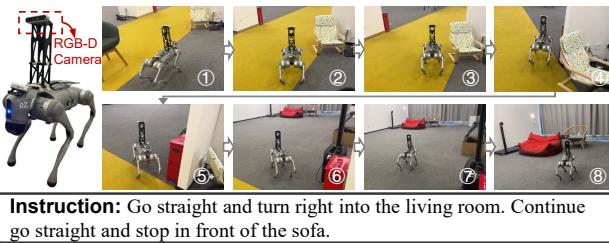


Figure 11. Real-world experiments using a Unitree Go2 robot.

igate this, we let the robot output the stop action when the predicted probability of the stop action exceeds 1×10^{-4} .

C. Analysis of Failure Cases

In Tab. 3, we observe that the SoTA ego-centric model, NaViD [64], shows exceptionally poor zero-shot performance (e.g., 5.8 SR and 1.0 SPL) on our 3DGS-Lab-VLN datasets. The possible reasons for the performance degradation could be summarized as follows: Firstly, the use of 3D Gaussian Splatting (3DGS) for rendering may introduce artifacts and distortions. As shown in Fig. 13, rendering artifacts can cause some blurring in ground areas and distant details, introducing subtle distortions that may go unnoticed by the human eye. In our experiments, the model relying solely on RGB input is highly vulnerable to such pixel-level noise, leading to failure in affected scenes. This underscores the need for research on image perturbations and related safety issues in VLN models. Additionally, we note that the NaViD model frequently rotates in circles to find a better viewpoint for localizing the target, which accounts for 70% of failures (Fig. 12). In summary, our findings on the limitations of current SoTA VLN methods align with the conclusions of this paper. We hope our insights and tools will drive the development of more robust and generalizable VLN models, especially in diverse, non-MP3D-style environments.

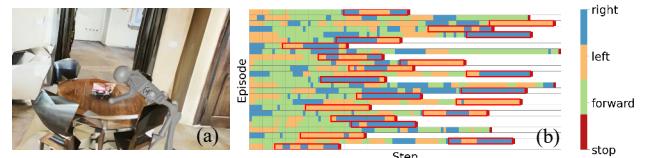


Figure 12. Visualization of the failure cases. (a) shows a typical failure case where the agent collides and falls. (b) highlights NaVid’s repetitive turning before stopping (red box).

D. Limitations and Future Work

While this work evaluates various ego-centric VLN methods, some other state-of-the-art VLN approaches rely on panoramic observations [1, 3, 24]. These methods use panoramic views with depth to generate sparse waypoint connections, integrating them with discrete VLN techniques for path selection—an approach that has demonstrated strong performance in previous non-physical settings. Since our primary goal is to evaluate the existing VLN methods under the physical settings, we adopt an ego-centric view setting to align with current robotic perception systems. However, as robotics evolves—potentially resembling autonomous driving systems with more diverse RGB or radar sensors—future robots may benefit from panoramic perception. Thus, we plan to extend our evaluation to



"Steer right towards the red ladder, then arrive at the white table with black chairs. Turn left near the yellow crane, advance past the robot statue, and move forward towards the desk with chairs. Finally, head left towards the office chair."

Figure 13. Visualization of the failure case for the ego-centric SoTA VLN model, NaViD, in 3DGS online-rendered scenes. The model tends to predict rotation actions, indicating a failure to interpret the intended trajectory.

panoramic VLN methods in future work. Additionally, with multi-robot support and real-time 3DGS scene rendering, our platform has significant potential to facilitate a real-sim-real VLN pipeline, enhancing real-world adaptability for embodied agents in familiar environments. We leave this for future research.

E. Additional Qualitative Examples

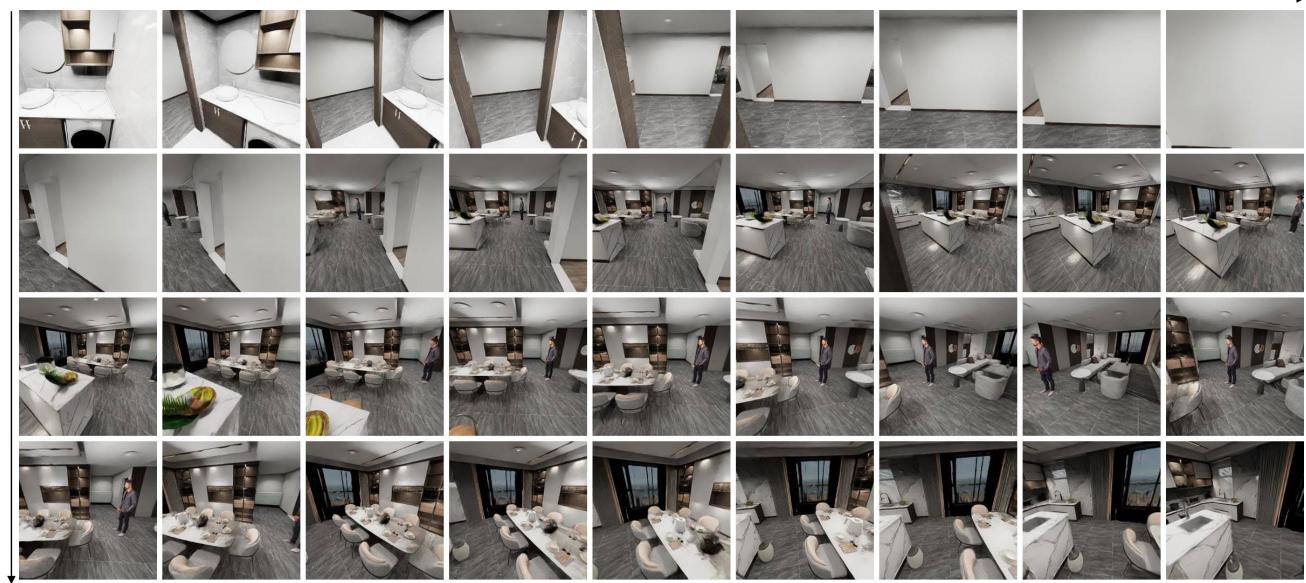
To better illustrate the observations and environments within VLN-PE, we provide supplementary videos showcasing the significant shaking and instability experienced by physical agents during navigation. Additionally, Fig. 14 presents different viewpoints—ego-centric, third-person, and top-down—using various robot types in VLN-PE.

Fig. 15 displays trajectories and instructions from our newly introduced 10 high-quality synthetic scenes (GRU-VLN10) and a 3DGS online-rendered scene (3DGS-Lab-VLN), supporting out-of-MP3D-style evaluations.

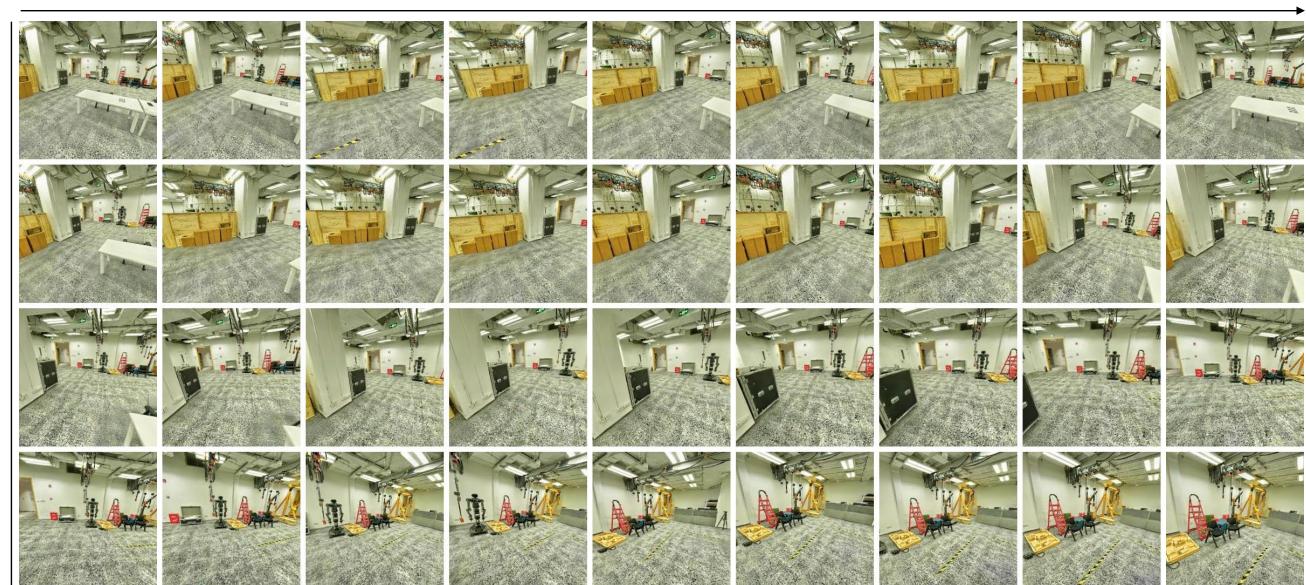


* In the simulation, we raised the Jetbot-wheeled robot's camera height to approximately 1.2 m to compensate for its low position.

Figure 14. Visualization of different robot viewpoints in VLN-PE. Leveraging the powerful interactive capabilities of Isaac Sim, researchers can easily observe robot motion from various perspectives within the environment.



Turn left into the bathroom with a round mirror above the sink. Move through the indoor space, steering left into the empty room. Proceed to the living room, then enter the dining room with a table and chairs near the window. Continue straight to the kitchen, stopping at the sink.



Turn right, move forward past the wooden counter to the long white table, then turn left at the workbench. Continue forward, passing the white bench and wooden crates, then enter near the wall with exposed pipes. Finally, turn left and stop by the ladder.

Figure 15. Examples of trajectories and instructions from our introduced GRU-VLN10 and 3DGS-Lab-VLN datasets.