AgentVLN: Towards Agentic Vision-and-Language Navigation

Zihao Xin[1], Wentong Li[1]*(✉), Yixuan Jiang[1], Ziyuan Huang[1], Bin Wang[2], Piji Li[1], Jianke Zhu[3], Jie Qin[1], and Shengjun Huang[1](✉)

[1] Nanjing University of Aeronautics and Astronautics, China
[2] Shandong University, China
[3] Zhejiang University, China
wentong_li@nuaa.edu.cn, huangsj@nuaa.edu.cn
*Project lead ✉Corresponding authors

Abstract. Vision-and-Language Navigation (VLN) requires an embodied agent to ground complex natural-language instructions into long-horizon navigation in unseen environments. While Vision-Language Models (VLMs) offer strong 2D semantic understanding, current VLN systems remain constrained by limited spatial perception, 2D–3D representation mismatch, and monocular scale ambiguity. In this paper, we propose AgentVLN, a novel and efficient embodied navigation framework that can be deployed on edge computing platforms. We formulate VLN as a Partially Observable Semi-Markov Decision Process (POSMDP) and introduce a **VLM-as-Brain** paradigm that decouples high-level semantic reasoning from perception and planning via a plug-and-play skill library. To resolve multi-level representation inconsistency, we design a cross-space representation mapping that projects perception-layer 3D topological waypoints into the image plane, yielding pixel-aligned visual prompts for the VLM. Building on this bridge, we integrate a context-aware self-correction and active exploration strategy to recover from occlusions and suppress error accumulation over long trajectories. To further address the spatial ambiguity of instructions in unstructured environments, we propose a Query-Driven Perceptual Chain-of-Thought (QD-PCoT) scheme, enabling the agent to actively query and acquire geometric depth information according to task demands. Finally, we construct AgentVLN-Instruct, a large-scale instruction-tuning dataset with dynamic stage routing conditioned on target visibility. Extensive experiments show that AgentVLN consistently outperforms prior state-of-the-art (SOTA) methods on long-horizon VLN benchmarks, offering a practical paradigm for lightweight deployment of next-generation embodied navigation models. Code: https://github.com/Allenxinn/AgentVLN.

Keywords: Vision-and-Language Navigation · Vision-Language Models · Embodied Intelligence

1 Introduction

Vision-and-Language Navigation (VLN) requires embodied agents to make sequential decisions in complex physical environments by following natural-language

instructions based on dynamically acquired egocentric observations. Despite the strong semantic understanding exhibited by Vision-Language Models (VLMs) on 2D images [4,12,26,31,38,43,45], a significant gap remains in effectively deploying them in the highly unstructured and uncertain 3D physical world [6,15,34,52,53].

Current VLN systems can be primarily categorized into two paradigms: single-system and dual-system architectures. Despite their progress, both still face critical challenges. Single-system approaches [48, 50, 57, 62] seek to implicitly map visual and linguistic features to action chunks using massive amounts of video-trajectory data. This black-box mapping struggles to learn the underlying geometric structure of navigation, limiting its cross-environment generalization. Dual-system methods [40, 47] typically introduce diffusion-based modules as trajectory generators while employing VLMs as low-frequency global planners. However, a pronounced cross-space disconnect frequently emerges between the generated 3D trajectories and the 2D visual representations of the VLMs. This stems from the fact that pre-trained VLMs inherently lack the capacity to reason about 3D geometry and metric scale. Moreover, both paradigms struggle with spatial scale ambiguity. Monocular RGB observations lack reliable depth cues, making it hard for VLMs to localize targets involving spatial prepositions. While some methods [40,57,62] inject spatial information through additional depth-estimation modules [28, 42], this typically increases inference latency and context length, hindering efficient deployment.

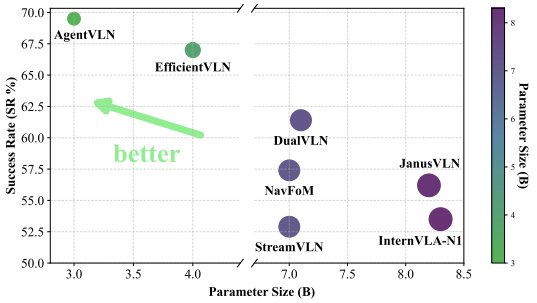

Fig. 1: Performance comparison on the Val-Unseen split of RxR-CE dataset [22]. AgentVLN outperforms existing state-of-the-art models while utilizing a substantially smaller parameter footprint. Furthermore, owing to its lightweight architecture, AgentVLN supports real-time local inference on Jetson embedded edge boards, eliminating the reliance on remote cloud deployment.

To address the aforementioned issues, we propose **AgentVLN**, a novel and highly efficient embodied navigation framework. We reformulate the VLN system under a "VLM-as-Brain" paradigm that decouples high-level cognitive reasoning from low-level planning execution, and model the task as a Partially Observable Semi-Markov Decision Process (POSMDP). In AgentVLN, the VLM acts as the central controller and makes decisions based solely on the current multimodal context, performing global path navigation and local target localization. By alternately invoking perception-level skills and cross-timestep planning-level skills, the system seamlessly scales from short-horizon exploration and error correction to long-horizon navigation, enabling robust generalization.

To tackle the multi-level representation inconsistency in dual-system VLN, we propose a cross-space representation mapping mechanism. In contrast to prior approaches that forcefully align multimodal features in an implicit high-dimensional latent space, we explicitly project 3D waypoints calculated by the perception layer onto the 2D image plane via perspective projection, producing pixel-aligned visual prompts that are directly consumable by the VLM. This design bridges the semantic gap between 2D visual representations and 3D physical structure, significantly improving generalization in long-horizon navigation. We further equip AgentVLN with a context-conditioned self-correction mechanism. When the system encounters occlusions, blind spots, or trajectory deviations, it generates fine-grained exploratory actions to actively recover, mitigating error accumulation and improving robustness in complex, unseen environments.

Furthermore, to mitigate spatial scale ambiguity during local target localization, we introduce a Query-Driven Perceptual Chain-of-Thought (QD-PCoT) scheme to enable a VLN system to actively query for geometric cues on demand. When confronted with spatial ambiguity, the agent actively formulates natural-language queries and invokes perception-layer skills. By integrating multi-round interactive feedback into an explicit chain-of-thought deduction, the agent ultimately infers instruction-consistent target pixel coordinates, which are then passed to planning-level skills to execute navigation. Finally, we build AgentVLN-Instruct, a large-scale instruction-tuning dataset that tightly aligns high-level instructions with low-level skill invocations. It features a dynamic stage-routing mechanism conditioned on target visibility, explicitly mirroring the human wayfinding process: navigate coarsely first, localize precisely second.

As shown in Fig. 1, AgentVLN consistently outperforms prior state-of-the-art methods on the VLN-CE benchmark while using a substantially smaller parameter budget, delivering a superior accuracy–efficiency trade-off. Benefiting from its lightweight design, AgentVLN enables real-time, on-device inference on Jetson embedded edge platforms, removing the need for remote cloud deployment and the associated communication and decoding latency. Moreover, real-world evaluations further validate its robustness in long-horizon planning and efficient execution across diverse scenarios.

In summary, our key contributions are threefold:

- We propose AgentVLN under a "VLM-as-Brain" paradigm, which decomposing long-horizon navigation into high-level semantic reasoning and skill scheduling over a plug-and-play library, yielding strong generalization in unseen, complex environments.
- We introduce a cross-space representation mapping mechanism that converts the environment's implicit 3D geometric topology into explicit, pixel-aligned visual prompts, and couple it with context-aware fine-grained self-correction and active exploration to enhance robustness.
- We present a novel Query-Driven Perceptual Chain-of-Thought (QD-PCoT) scheme, enabling query-guided target localization by actively querying for missing spatial cues during instruction grounding.

2 Related Work

2.1 Vision-and-Language Navigation

VLN requires an embodied agent to follow natural-language instructions and precisely reach a target destination using only egocentric visual observations. This complex task relies heavily on cross-modal semantic alignment and long-horizon spatial reasoning. Early approaches primarily utilize Seq2Seq-style architectures [11, 16, 19] for direct state-to-action sequence predictions. Subsequent works, such as NavGPT [63] and InstructNav [32], integrate Large Language Models (LLMs) to handle high-level instruction parsing and global path planning. However, these non-end-to-end pipelines inevitably lose fine-grained geometric and spatial cues during modality conversion. To address this gap, NaVid [60] introduces an end-to-end action prediction paradigm based on VLMs, while it remains constrained by context-length limits in long-horizon tasks. More recent methods, including StreamVLN [48] and JanusVLN [57], alleviate this token explosion issue via sliding-window processing and slow-fast memory mechanisms. InternVLA-N1 [40] and DualVLN [47] draw inspiration from human cognition and explore dual-system designs that decouple high-level semantic reasoning from low-level action execution, improving trajectory quality and navigation efficiency. Despite these advances, practical deployment remains challenging. Current VLN systems [13, 40, 48, 57, 59] typically operate at a large parameter scale (often $\geq$7B). Furthermore, auxiliary depth modules [57, 62] introduce substantial computational and storage overhead. Consequently, these methods are often deployed on remote cloud servers, where high-bandwidth data transmission and remote autoregressive generation make it difficult to satisfy the real-time control demands of physical agents in complex, unstructured environments.

2.2 Vision-Language Model

In recent years, VLMs have achieved great advancements in architecture designs and multimodal representation capabilities [7, 26, 27, 31, 44, 54–56]. As a pioneering representative in this domain, the LLaVA series [24, 29–31] establish a widely adopted recipe for extending LLMs to visual perception tasks via visual instruction tuning. Subsequent variants further improve fine-grained visual understanding [24, 30], long-context video comprehension, and precise spatial grounding by introducing dynamic high-resolution processing alongside more comprehensive training corpora. More recently, VLMs are undergoing a paradigm progressively evolving from passive perception and understanding tools into active, decision-making agents [5, 43, 51]. For instance, Qwen3-VL [4, 51] and Qwen3.5 [35] integrate "thinking with images" and function-calling mechanisms, facilitating tool-augmented self-reflection and sophisticated multi-step reasoning. Nevertheless, a glaring limitation persists: current multimodal agent methodologies are almost exclusively constrained to highly structured software environments [25, 36, 49] and inherently fail to generalize to the dynamic, unstructured complexities of real-world physical scenarios.

3 AgentVLN

3.1 Overview

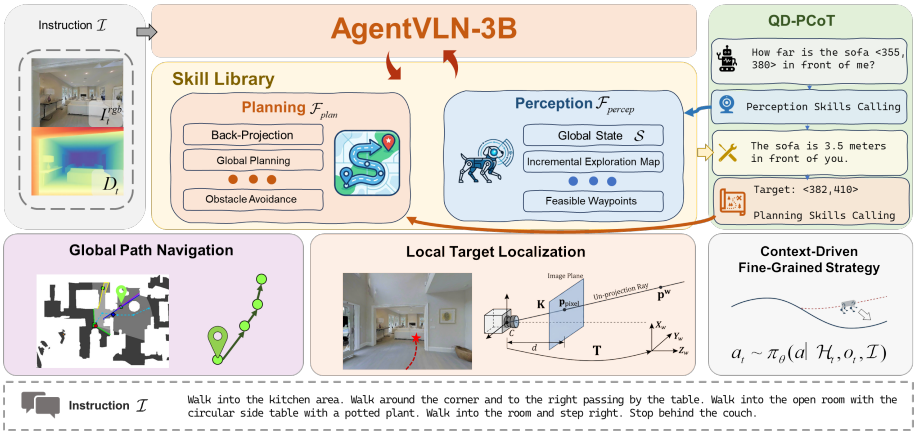

Fig. 2: Overview of the AgentVLN framework. AgentVLN employs a VLM-as-Brain paradigm, decomposing long-horizon navigation into modular skill executions. Additionally, a context-driven fine-grained strategy and QD-PCoT mitigate localization errors and scale ambiguities, ensuring precise 3D target grounding.

In this section, we present AgentVLN, a novel and efficient embodied navigation framework. Unlike prevailing end-to-end approaches that require massive video datasets for pre-training to implicitly learn spatial geometry into policies, AgentVLN adopts a **VLM-as-Brain** embodied paradigm. This paradigm explicitly decouples high-level cognitive reasoning from low-level skill execution. Specifically, we leverage a VLM as the central controller to interpret complex natural-language instructions and decompose long-horizon navigation into a structured sequence of skill invocations. Moreover, AgentVLN abstracts the underlying motion mechanics into a highly modular and plug-and-play skill library. As a result, adapting to novel environments or tasks can be achieved by updating or expanding the skill library, without retraining the multi-billion-parameter VLM.

3.2 Problem Definition

We formalize AgentVLN's decision-making process as a Partially Observable Semi-Markov Decision Process (POSMDP) jointly conditioned on natural language and temporal memory, defined as $\mathcal{M} = \langle \mathcal{S}, \mathcal{O}, \mathcal{F}, \mathcal{T}, \mathcal{I}, \mathcal{H} \rangle$. The state $s \in \mathcal{S}$ encodes the agent's global position and pose, while the observation $o_t \in \mathcal{O}$ comprises egocentric data streams acquired from onboard sensors at time step t. The agent's action space is no longer defined by low-level continuous velocity commands or discrete directional movements (*e.g.*, forward/left/right), but rather by a set of skill functions $\mathcal{F} = \{f_1, f_2, \ldots, f_K\}$. The multi-step transition dynamics

$\mathcal{T}$ describe the state evolution laws over a variable time steps τ induced by executing a skill, denoted as $P(s_{t+\tau}, \tau \mid s_t, f)$. Concretely, at each decision step t, the agent generates a high-level control policy or a structured calling instruction to activate a skill based on the current history context $\mathcal{H}_t$, visual observations o_t, and user instructions $\mathcal{I}$:

$$c_k \sim \pi_\theta(f \mid \mathcal{H}_{t_k}, o_{t_k}, \mathcal{I}), \quad f \in \mathcal{F}, \tag{1}$$

where π_θ is the policy model parameterized by VLM.

Motivated by the human cognitive pattern of "perceive before act", we further decompose the skill set into perception-level skills $\mathcal{F}_{percep}(\tau = 0)$ and planning-level skills $\mathcal{F}_{plan}(\tau > 0)$, with $\mathcal{F} = \mathcal{F}_{percep} \cup \mathcal{F}_{plan}$. Perception-level skills extract geometric or semantic features from the environment and return augmented observations Δo, which are integrated into the context to reduce local ambiguity and inform subsequent decisions. In contrast, planning-level skills perform closed-loop physical execution: given a skill policy π_f, they produce a sequence of low-level actions a_t over τ time steps, following the multi-step transition distribution:

$$P(s_{t_k+\tau}, \tau \mid s_{t_k}, f_{c_k}) = \sum_{a_1 \dots a_\tau} \left(\prod_{i=0}^{\tau-1} P(s_{t_k+i+1} \mid s_{t_k+i}, a_{t_k+i}) \pi_{f_{c_k}}(a_{t_k+i}) \right) \beta_{f_{c_k}}(s_{t_k+\tau}). \tag{2}$$

Here, $\pi_{f_{c_k}}(a_{t_k+i})$ is the control signal produced by the navigation skill at step i, $P(s_{t_k+i+1} \mid s_{t_k+i}, a_{t_k+i})$ denotes the local state transition under action a_{t_k+i}, and $\beta_{f_{c_k}}(s_{t_k+\tau}) \in [0, 1]$ is the skill termination condition that determines whether the VLM needs to initiate a new decision cycle. If $\beta = 0$, the current planning skill continues; otherwise, the VLM updates the state context and proceeds to the next semi-Markov decision-making cycle.

In VLN, the VLM is optimized to precisely translate high-dimensional, complex semantic intentions with spatiotemporal motion trajectories in the physical space. In the initial stages of an episode, the VLM tends to alternately invoke perception and planning skills. First, it calls upon $\mathcal{F}_{percep}$ given $\mathcal{O}$ to construct a local geometric map. Subsequently, incorporating the spatial prepositions within $\mathcal{I}$, the VLM triggers a planning skill to execute a τ-step movement. When the visual features in the observation o_t highly match the destination description in the navigation instruction $\mathcal{I}_{nav}$, the decision-making process automatically switches from global path navigation to local target localization, guiding the robot to accomplish precise docking.

3.3 Global Path Navigation

To address 3D trajectory failures arising from multi-level representation inconsistencies, typically of popular dual-system architectures, AgentVLN introduces a novel cross-space representation mapping mechanism. During navigation, the VLM first invokes the perception-level skill $\mathcal{F}_{percep}$ to update the agent's global state $\mathcal{S}$ and construct an incremental exploration map. At each time step t, the agent receives multimodal observations $o_t = \{I_t^{rgb}, D_t\}$ together with its pose in the world coordinate system $T_t = [R_t \mid \mathbf{t}_t] \in SE(3)$, where $R_t \in SO(3)$ is

the rotation matrix and $\mathbf{t}_t = [x_t, y_t, h_c]^T$ is the translation vector with camera height h_c. Given the camera intrinsic matrix $K \in \mathbb{R}^{3\times3}$, for any valid pixel $\mathbf{p}_{img} = [u, v, 1]^T$ with depth d in D_t, we back-project it into a 3D point $\mathbf{P}^w$ in the world coordinate system via

$$\mathbf{P}^w = R_t(d \cdot K^{-1}\mathbf{p}_{img}) + \mathbf{t}_t. \tag{3}$$

As the agent moves, these back-projected points are incrementally updated into a global occupancy grid map, enabling planning-level skills to generate a series of 3D waypoints. To enable the VLM to make navigation decisions directly within the visual space, let a globally feasible path point in the world coordinate system be denoted as $\mathbf{P}^w_{path} = [X_{path}, Y_{path}, 0]^T$. Through the inverse transformation of the camera pose T_t and perspective projection, we calculate the pixel coordinates $\mathbf{p}^{img}_{path} = [u_{path}, v_{path}, 1]^T$ of this 3D path point on the current observation:

$$s \cdot \mathbf{p}^{img}_{path} = KR_t^{-1}(\mathbf{P}^w_{path} - \mathbf{t}_t), \tag{4}$$

where s is the depth scale factor of the mapped point in the camera coordinate system. By converting the environment's implicit 3D topology into explicit, pixel-aligned visual prompts, the VLM only needs to select the matching path in the pixel space based on the semantics of $\mathcal{I}$. It then invokes the planning-level skills $\mathcal{F}_{plan}$ to restore this selection to a 3D waypoint for navigation control. This mechanism not only resolves the VLM's inherent deficit in spatial perception but also significantly improves generalization in long-horizon navigation.

Furthermore, to mitigate the issues, like view occlusions, instruction ambiguity, and accumulated localization errors in complex environments, we present a context-driven, fine-grained closed-loop self-correction and active exploration mechanism. When the current observation o_t contains no globally feasible path projection that is semantically relevant to $\mathcal{I}$, the agent performs commonsense reasoning and local geometric inference conditioned on $\mathcal{H}_t$ and $\mathcal{I}$, and outputs fine-grained actions:

$$a_t \sim \pi_\theta(a \mid \mathcal{H}_t, o_t, \mathcal{I}), \quad a \in \{\texttt{Forward, Left, Right}\}. \tag{5}$$

Importantly, in blind spots without feasible global path guidance, the agent is not forced to execute long-distance blind displacements. Instead, this context-based fine-tuning enables the agent to autonomously look around and to promptly correct its course when deviating from the intended route. Once a valid global path mapping is successfully captured, AgentVLN resumes alternating between perception- and planning-level skills for efficient long-horizon execution. This dynamic switching between macro-skill invocation and context-driven corrective fine-tuning endows AgentVLN with both high efficiency and strong self-recovery, improving robustness under trajectory deviations.

3.4 Local Target Localization

As navigation proceeds, once the agent's observation o_t contains multimodal evidence that strongly matches the destination description in the instruction $\mathcal{I}$,

AgentVLN smoothly transitions from exploration-based global path planning to local target localization. In this stage, the VLM locks onto the destination target coordinates in the egocentric 2D view, and achieves precise docking through spatial geometric remapping.

However, in real-world unstructured environments, instructions frequently exhibit spatial ambiguity. Under such ambiguity, relying solely on monocular RGB makes it difficult for the VLM to accurately predict the target's pixel coordinates, frequently leading to failures due to scale uncertainty and depth-perspective ambiguity. To address this key limitation, we propose a novel Query-Driven Perceptual Chain-of-Thought (QD-PCoT) scheme, endowing the agent with the query-guided ability to actively query missing information.

Specifically, when the model detects spatial ambiguity, it does not blindly output coordinates. Instead, it generates an intermediate natural-language query *(e.g., "How many meters away is the chair in front of me?")* and triggers a specific perception-level skill $\mathcal{F}_{percep}$. The returned geometric/semantic information is then fed back into the context as an incremental textual prompt. Guided by this multi-step reasoning process, the model then outputs precise destination pixel coordinates $\mathbf{p}_{target}^{img} = [u_{target}, v_{target}, 1]^T$ that strictly align with the instructional intent. To project the pixel coordinates into 3D coordinate system, we extract its depth $d_{target} = D_t(u_{target}, v_{target})$ from the depth map D_t and back-project it into the local camera coordinate system:

$$\mathbf{P}_{target}^{c} = d_{target} \cdot K^{-1}\mathbf{p}_{target}^{img} = \begin{bmatrix} \frac{(u_{target}-c_x)\cdot d_{target}}{f_x} \\ \frac{(v_{target}-c_y)\cdot d_{target}}{f_y} \\ d_{target} \end{bmatrix}, \tag{6}$$

where f_x, f_y are focal lengths, and c_x, c_y represent the principal point. Given the current pose T_t, we transform the 3D point to the global coordinate system via a rigid-body transformation:

$$\mathbf{P}_{target}^{w} = R_t\mathbf{P}_{target}^{c} + \mathbf{t}_t. \tag{7}$$

In this way, the VLM's high-level semantic selection in the 2D visual plane is seamlessly translated into 3D target coordinates $\mathbf{P}_{target}^{w}$ in physical space, after which planning-level skills are invoked to execute accurate goal-reaching and docking.

3.5 AgentVLN-Instruct Dataset

To bridge the gap in existing datasets regarding the alignment between high-level language instructions and low-level skills, we construct AgentVLN-Instruct, a large-scale navigation dataset in the Habitat simulator based on R2R and RxR datasets, covering 61 indoor scenes. Unlike traditional methods that employ a singular control policy throughout the entire navigation process, AgentVLN-Instruct introduces a dynamic stage-routing mechanism conditioned on target visibility. Concretely, at each decision step, the process is routed to either global

path navigation or local target localization based on the target's visibility in I_t^{rgb}. This design mirrors the human wayfinding process, *i.e.*, navigate coarsely first, localize precisely second.

In the **global path planning** phase, the model is required to trigger perception-level skills $\mathcal{F}_{percep}$ and we select the mapped waypoint closest to the expert trajectory as the navigation label. This guides the VLM to learn how to select the most instruction-consistent direction among multiple visual cues. Additionally, random noise is injected to force deviations from the optimal path. When no valid waypoints exist in I_t^{rgb}, the expected output format is switched to fine-grained action sequences, encouraging the model to acquire context-driven self-localization and closed-loop error correction capabilities. In the **local target localization** phase, we design a multi-round, interactive perceptual Chain-of-Thought finetuning format that simulates the agent's active reasoning and perceptual querying behavior, ultimately producing the target pixel coordinates.

Through explicit Chain-of-Thought prompting and lightweight, plug-and-play skill invocation, we directly activate the robust commonsense reasoning and logical discrimination already present in the pre-trained VLM. By bridging 2D visual evidence and 3D physical structure through natural-language-guided perceptual reasoning, our formulation not only improves localization success in complex and ambiguous scenarios but also endows the model's decision-making process more interpretable, providing a novel paradigm for lightweight general-purpose embodied agents.

4 Experiment

4.1 Datasets and Evaluation Metrics

To evaluate the model under varying levels of navigation difficulty and semantic complexity, we conduct extensive experiments on the benchmark datasets R2R-CE [3] and RxR-CE [22]. However, existing VLN datasets only provide coarse instruction–waypoint mappings, which are insufficient to support the finetuning requirements of high-level skill invocation and QD-PCoT reasoning in the VLM-as-Brain architecture. To address this limitation, we construct a large-scale instruction tuning dataset, AgentVLN-Instruct, based on the Habitat simulator [39]. Unlike traditional trajectory collection approaches, this dataset incorporates four key components: target-visibility-driven dynamic routing, generalizable skill invocation, localization reasoning, and active question–answer interactions. To preserve the model's general multimodal capabilities, we additionally include the LLaVA-Video-178K dataset [61] during training. For evaluation, we follow standard VLN benchmarking protocols and report commonly used navigation metrics, including Success Rate (SR), Oracle Success Rate (OS), Success weighted by Path Length (SPL), and Navigation Error (NE).

4.2 Implementation Details

We adopt Qwen2.5-VL-3B [5] as the brain of AgentVLN. During the instruction-tuning phase, we freeze the model's visual encoder and multimodal projection

Table 1: Comparison results with SOTA methods on the Val-Unseen dataset for R2R-CE. Our method outperforms other approaches on the same benchmarks, even having fewer parameters.

Method	Observation				R2R-CE			
	S.RGB	Pano.	Depth	Odo.	NE ↓	OS ↑	SR ↑	SPL ↑
HPN+DN [20]		✓	✓	✓	6.31	40.0	36.0	34.0
VLN BERT [17]		✓	✓	✓	5.74	53.0	44.0	39.0
CMA [17]		✓	✓	✓	6.20	52.0	41.0	36.0
Reborn [2]		✓	✓	✓	5.40	57.0	50.0	46.0
Ego^2-Map [18]		✓	✓	✓	5.54	56.0	47.0	41.0
DreamWalker [41]		✓	✓	✓	5.53	59.0	49.0	44.0
ETPNav [1]		✓	✓	✓	4.71	65.0	57.0	49.0
Seq2Seq [21]	✓		✓		7.77	37.0	25.0	22.0
RGB-CMA [21]	✓				9.55	10.0	5.0	4.0
AG-CMTP [9]		✓	✓		7.90	39.0	23.0	19.0
R2R-CMTP [9]		✓	✓		7.90	38.0	26.0	22.0
LAW [37]	✓		✓	✓	6.83	44.0	35.0	31.0
CM2 [14]	✓		✓		7.02	41.0	34.0	27.0
WS-MGMap [10]	✓		✓		6.28	47.0	38.0	34.0
ETPNav+FF [46]	✓		✓		5.95	55.8	44.9	30.4
AO-Planner [8]		✓	✓		5.55	59.0	47.0	33.0
NaVid-7B [60]	✓				5.47	49.0	37.0	35.0
Uni-NaVid-7B [59]	✓				5.58	53.3	47.0	42.7
NaVILA-7B [13]	✓				5.22	62.5	54.0	49.0
StreamVLN-7B [48]	✓				5.10	64.0	55.7	50.9
NavFoM-7B [58]	✓				5.01	64.9	56.2	51.2
DecoVLN-7B [50]	✓				5.01	63.5	56.3	50.5
JanusVLN-8.2B [57]	✓				4.78	65.2	60.5	56.8
EfficientVLN-4B [62]	✓				4.18	**73.7**	64.2	55.9
DualVLN-7.1B [47]	✓				4.05	70.7	64.3	58.5
StreamVLN-7B [48]	✓		✓		4.98	64.2	56.9	51.9
InternVLA-N1-8.3B [40]	✓		✓		4.83	63.3	58.2	54.0
AgentVLN-3B	✓		✓		**3.88**	73.5	**67.2**	**64.7**

layer. The network is optimized using the AdamW [33] with a batch size of 128. For the learning rate schedule, we employ a cosine annealing strategy with a peak learning rate of 2×10^{-5}, incorporating a warmup ratio of 0.03 at the initial stage of training. The egocentric observations are captured at a resolution of 640×480 with a 110° field of view, and the temporal window for the multimodal historical context is set to 8 frames. All experiments were conducted on a computing cluster equipped with 32 NVIDIA A100 GPUs. Please refer to the *supplemental material* for more details.

4.3 Navigation Experiments

Tab. 1 and Tab. 2 present the performance of AgentVLN compared against SOTA methods on the Val-Unseen splits of the R2R-CE [3] and RxR-CE [22]

Table 2: Comparison with SOTA methods on RxR-CE Val-Unseen split.

Method	S.RGB	Pano.	Depth	Odo.	NE ↓	SR ↑	SPL ↑	nDTW ↑
VLN BERT [17]		✓	✓	✓	8.98	27.0	22.6	46.7
CMA [17]		✓	✓	✓	8.76	26.5	22.1	47.0
Reborn [2]		✓	✓	✓	5.98	48.6	42.0	63.3
ETPNav [1]		✓	✓	✓	5.64	54.7	44.8	61.9
Seq2Seq [21]	✓		✓		12.10	13.9	11.9	30.8
LAW [37]	✓		✓	✓	10.90	8.0	8.0	38.0
ETPNav+FF [46]	✓		✓	✓	8.79	25.5	18.1	-
AO-Planner [8]		✓	✓		7.06	43.3	30.5	50.1
Uni-NaVid-7B [59]	✓				6.24	48.7	40.9	-
NaVILA-7B [13]	✓				6.77	49.3	44.0	58.8
NavFoM-7B [58]	✓				5.51	57.4	49.4	60.2
JanusVLN-8.2B [57]	✓				6.06	56.2	47.5	62.1
StreamVLN-7B [48]	✓				6.16	51.8	45.0	61.9
InternVLA-N1-7B [40]	✓				6.41	49.5	41.8	62.6
DecoVLN-7B [50]	✓				5.73	54.2	46.3	63.5
EfficientVLN-4B [62]	✓				**3.88**	67.0	54.3	68.4
DualVLN-7.1B [47]	✓				4.58	61.4	51.8	70.0
StreamVLN-7B [48]	✓		✓		6.22	52.9	46.0	61.9
InternVLA-N1-8.3B [40]	✓		✓		5.91	53.5	46.1	65.3
AgentVLN-3B	✓		✓		3.92	**69.5**	**61.3**	**74.6**

benchmarks. The evaluated baselines encompass waypoint prediction models relying on multi-sensor inputs, as well as large vision-language navigation models. Operating with a remarkably compact parameter scale of merely 3B, AgentVLN outperforms the SOTA method of the same class, InternVLA, by substantial margins of 9.0% SR and 10.7% SPL on the R2R dataset, all while maintaining a significantly smaller parameter footprint. To compensate for the VLM's inherent deficit in 3D scale perception within the physical world, prior methods such as EfficientVLN introduce monocular depth estimation modules to compel the VLM to implicitly learn 3D geometric relationships. However, this approach not only incurs severe computational overhead but also heavily exacerbates the model's memory burden. In contrast, on the R2R and RxR datasets, AgentVLN achieves absolute improvements of 3.0 and 2.5 in SR over EfficientVLN, respectively. More notably, on the SPL metric which strictly reflects the optimality of the navigation trajectory and overall decision-making efficiency. Our method yields impressive improvements of 8.8% and 7.0%. These results explicitly validate that the VLM-as-Brain paradigm, when integrated with context-driven fine-grained self-correction, effectively elevates navigation efficiency. We display the visualization results in Fig. 3.

In conclusion, the experimental results substantiate that AgentVLN significantly bolsters navigation performance with absolutely zero additional parameter overhead, thereby establishing a novel and highly effective paradigm for next-generation, lightweight, and high-precision embodied navigation frameworks.

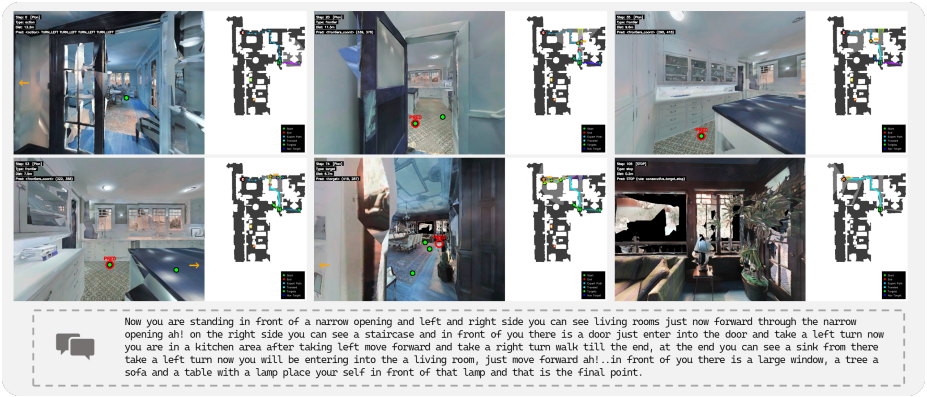

Fig. 3: Visualization of AgentVLN's navigation. Green points represent the visual prompts generated by the perception-level skills, whereas red circles denote the navigation waypoints selected or predicted by the model. Notably, when traversing narrow passages or confronting severe visual occlusions, the model seamlessly outputs fine-grained atomic actions for trajectory fine-tuning. This demonstrates its robust capability to achieve highly accurate, collision-free navigation relying exclusively on egocentric observations.

4.4 Real-World Deployment

Existing VLN models, burdened by massive parameter counts and a reliance on auxiliary 3D vision encoders, typically necessitate offloading data streams to cloud servers for asynchronous inference. The compounding of autoregressive generation latency from large models and network communication delays renders these models inadequate for the high-frequency control demands of real-world scenarios. In stark contrast, AgentVLN adopts Qwen2.5-VL-3B [5] as its base model and eschews any 3D feature fusion modules, thereby enabling fully localized execution on Jetson edge computing platforms. As shown in Fig. 4, experimental results demonstrate that AgentVLN exhibits outstanding navigation performance in both indoor and outdoor environments.

We employ the Unitree Go2 quadruped robot as our physical platform, equipped with an Intel RealSense D455 camera to acquire the egocentric observation $\mathbf{O}$. The perception-layer skill library incorporates the RTAB-Map [23] algorithm alongside a suite of camera processing utilities. Upon the VLM triggering a perception skill, RTAB-Map executes real-time local SLAM and incrementally updates the 3D occupancy grid map. Concurrently, inverse perspective projection is utilized to supply the VLM with pixel-level visual prompts. Once the VLM outputs the target pixel coordinates, these coordinates are instantaneously remapped into 3D target waypoints within the global coordinate system. Subsequently, the planning-level skills encompassing path generation algorithms and robot control software packages are invoked to generate smooth, obstacle-avoiding trajectories, which are then translated into low-level control signals. Notably, this highly modular methodology is readily generalizable to other sensor

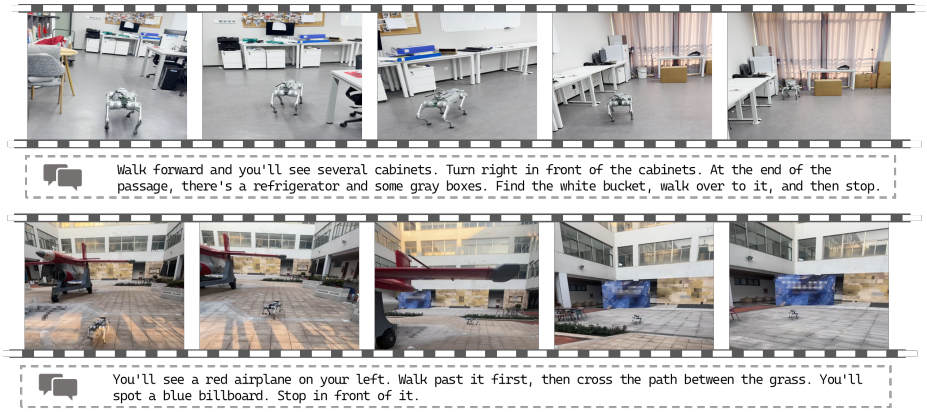

Fig. 4: Navigation results in real-world indoor and outdoor environments. Experimental results demonstrate that, regardless of whether the agent is navigating through complex, confined indoor spaces or outdoor scenarios with challenging illumination conditions, the proposed model consistently and accurately comprehends natural language instructions, enabling it to rapidly plan and execute precise navigation trajectories.

modalities such as LiDAR, adapting to new sensors merely requires substituting the relevant perception-layer skills within the library, entirely circumventing the need to retrain the multi-billion-parameter foundation model. AgentVLN achieves real-time inference directly on the Jetson Orin NX edge computing module, pioneering a novel paradigm for autonomous navigation characterized by low latency, high precision, and exceptional environmental adaptability. Specifically, by leveraging TensorRT and the vLLM framework to implement operator fusion and efficient context management, a single forward pass and autoregressive generation require approximately 3 seconds. This latency is further compressed to a mere 1 second following W4A16 quantization. Please refer to the *supplemental material* for detailed hardware specifications and configurations.

4.5 Ablation Studies

Incremental Ablation Study. To comprehensively dissect the individual contributions of each module within the AgentVLN architecture to the overall system navigation performance, we conducted systematic ablation studies on the R2R validation set. As presented in the Tab. 3, under an end-to-end paradigm devoid of explicit geometric guidance, lightweight VLMs struggle profoundly to implicitly learn complex 3D spatial geometric relationships and physical reachability solely from the appearance features of 2D images. Consequently, the baseline model achieves 38.6% SR merely, exposing severe deficiencies in the VLM's spatial perception capabilities. Building upon the baseline, the introduction of the cross-space representation mapping mechanism and the establishment of the VLM-as-Brain high-level scheduling paradigm yield substantial improvements: SR and OS surge by absolute margins of 21.1% and 16.6%, respectively, while NE decreases significantly to 4.67m. This demonstrates that alleviating

Table 3: Ablation study of different modules, where CDFG denotes context-driven fine-grained strategy.

VLM-as-Brain	CDFG	QD-PCoT	NE ↓	OS↑	SR↑	SPL↑
			6.53	48.50	38.60	35.10
✓			4.67	65.10	59.70	55.60
✓	✓		3.90	72.10	65.60	63.40
✓	✓	✓	**3.88**	**73.50**	**67.20**	**64.70**

the VLM's cognitive load associated with processing low-level action spaces and implicit spatial reasoning, thereby allowing it to focus exclusively on high-level semantic-visual matching and skill scheduling can significantly enhance model performance. Furthermore, the incorporation of the context-driven fine-grained strategy propels the SR steadily to 65.6%, while the NE is further compressed to 3.90m. The fundamental driver behind this performance leap lies in its effective mitigation of the irreversible error accumulation problem typical of long-horizon tasks. Finally, by integrating the QD-PCoT mechanism during the local target localization stage, the SR and SPL are further elevated to 67.2% and 64.7%, respectively. This indicates that in complex real-world scenarios, directly regressing target coordinates from 2D images frequently suffers from severe deviations induced by scale hallucinations. The QD-PCoT mechanism ensures that when confronted with complex environments, the VLM no longer resorts to blind, black-box predictions. Instead, it actively invokes perception-layer skills to acquire informational feedback, thereby elegantly eliminating the depth ambiguity of 2D vision without incurring any additional parameter overhead.

Temporal Context Length. To rigorously investigate the comprehensive impact of context length on overall navigation performance, we conducted an ablation study evaluating temporal context windows of varying lengths on the R2R validation set. As shown in Fig. 5, when the model is deprived of an effective temporal receptive field, the system inevitably succumbs to shortsightedness, yielding a mere 56.7% SR. In complex indoor environments, an overly sparse history of frames fundamentally fails to provide the model with sufficient prior contextual information. The VLM struggles to build a coherent awareness of its motion trajectory, making it highly susceptible to disorientation at local dead-ends or instruction inflection points. Conversely,

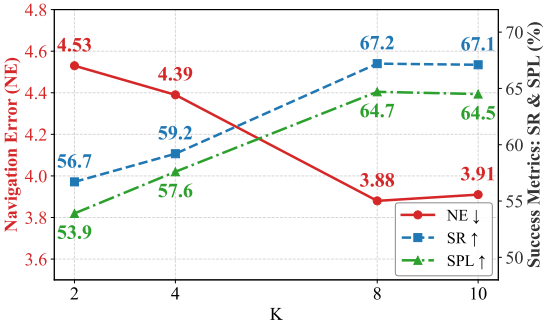

Fig. 5: Ablation on the impact of different temporal context length.

when the context window is optimally expanded to 8 frames, the model successfully achieves precise alignment between high-level navigational reasoning and dynamic visual transformations, peaking at an SR of 67.2% and SPL of 64.7%. However, further extending the sequence length actively exacerbates the issue of attention dilution. An excessively long context window inadvertently introduces a substantial volume of stale visual features that are entirely irrelevant to the current decision-making step. This accumulated visual noise inevitably disperses the VLM's attention weights away from critical, real-time visual prompts, thereby inducing the model to generate severe spatial hallucinations and degrading overall trajectory planning.

5 Conclusion

In this paper, we present AgentVLN, a novel and efficient embodied navigation framework. Unlike conventional methods that rely on massive datasets for implicit 3D representation learning, we introduce a VLM-as-Brain paradigm that decoupled high-level natural language cognitive reasoning from low-level perceptual control. This paradigm empowers the agent to dynamically adapt to unseen environments via a plug-and-play skill library. Specifically, we design a cross-space representation mapping mechanism that successfully eliminates the representation mismatches inherent in multi-level architectures. Synergized with a context-driven self-correction and active exploration strategy, this mechanism significantly bolsters the model's generalization performance. Furthermore, to tackle the persistent issue of inaccurate local target localization, we present the QD-PCoT, which effectively resolves semantic and spatial ambiguities within complex, unstructured environments while incurring zero parameter overhead. Finally, we construct and open-source AgentVLN-Instruct, a large-scale instruction-tuning dataset engineered with a dynamic stage-routing mechanism. Overall, AgentVLN provides a robust and practical paradigm for the development of next-generation lightweight embodied navigation frameworks.

Acknowledgment

This work was partially supported by the National Natural Science Foundation of China (No. U2441285, No. U25A20533, No. 62506165, No. 62276129), New Generation Artificial Intelligence-National Science and Technology Major Project (No. 2025ZD0122903), the Natural Science Foundation of Jiangsu Province (No. BK20250082), the Fundamental Research Funds for the Central Universities (No. NE2025010, No. NS2025038), the Jiangsu Funding Program for Excellent Postdoctoral Talent (No. 2025ZB306), and the Natural Science Foundation of Shandong Province (No. ZR2025QC1592).

References

1. An, D., Wang, H., Wang, W., Wang, Z., Huang, Y., He, K., Wang, L.: Etpnav: Evolving topological planning for vision-language navigation in continuous envi-

ronments. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024) 10, 11
2. An, D., Wang, Z., Li, Y., Wang, Y., Hong, Y., Huang, Y., Wang, L., Shao, J.: 1st place solutions for rxr-habitat vision-and-language navigation competition (2022) 10, 11
3. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., Van Den Hengel, A.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018) 9, 10
4. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025) 2, 4
5. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025) 4, 9, 12
6. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., Xia, F.: Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14455–14465 (2024) 2
7. Chen, G., Li, Z., Wang, S., Jiang, J., Liu, Y., Lu, L., Huang, D.A., Byeon, W., Le, M., Ehrlich, M., et al.: Eagle 2.5: Boosting long-context post-training for frontier vision-language models. Advances in Neural Information Processing Systems **38**, 91077–91100 (2026) 4
8. Chen, J., Lin, B., Liu, X., Liang, X., Wong, K.Y.K.: Affordances-oriented planning using foundation models for continuous vision-language navigation. arXiv preprint (2024) 10, 11
9. Chen, K., Chen, J.K., Chuang, J., Vázquez, M., Savarese, S.: Topological planning with transformers for vision-and-language navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021) 10
10. Chen, P., Ji, D., Lin, K., Zeng, R., Li, T.H., Tan, M., Gan, C.: Weakly-supervised multi-granularity map learning for vision-and-language navigation. In: Advances in Neural Information Processing Systems. vol. 35, pp. 27443–27456 (2022) 10
11. Chen, P., Sun, X., Zhi, H., Zeng, R., Li, T.H., Liu, G., Tan, M., Gan, C.: A^2nav: Action-aware zero-shot robot navigation by exploiting vision-and-language ability of foundation models. In: Advances in Neural Information Processing Systems. vol. 36 (2023) 4
12. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24185–24198 (2024) 2
13. Cheng, A.C., Ji, Y., Yang, Z., Gongye, Z., Zou, X., Kautz, J., Bıyık, E., Yin, H., Liu, S., Wang, X.: Navila: Legged robot vision-language-action model for navigation. Robotics: Science and Systems (2025) 4, 10, 11
14. Georgakis, G., Schmeckpeper, K., Wanchoo, K., Dan, S., Miltsakaki, E., Roth, D., Daniilidis, K.: Cross-modal map learning for vision and language navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022) 10

15. Gholami, M., Rezaei, A., Weimin, Z., Mao, S., Zhou, S., Zhang, Y., Akbari, M.: Spatial reasoning with vision-language models in ego-centric multi-view scenes. arXiv preprint arXiv:2509.06266 (2025) 2
16. Guhur, P.L., Tapaswi, M., Chen, S., Laptev, I., Schmid, C.: Airbert: In-domain pretraining for vision-and-language navigation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1634–1643 (2021) 4
17. Hong, Y., Wang, Z., Wu, Q., Gould, S.: 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) 10, 11
18. Hong, Y., Zhou, Y., Zhang, R., Dernoncourt, F., Bui, T., Gould, S., Tan, H.: Learning navigational visual representations with semantic map supervision (2023) 10
19. Irshad, M.Z., Mithun, N.C., Seymour, Z., Chiu, H.P., Samarasekera, S., Kumar, R.: Semantically-aware spatio-temporal reasoning agent for vision-and-language navigation in continuous environments. In: 2022 26th International conference on pattern recognition. pp. 4065–4071 (2022) 4
20. Krantz, J., Gokaslan, A., Batra, D., Lee, S., Maksymets, O.: Waypoint models for instruction-guided navigation in continuous environments. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021) 10
21. Krantz, J., Wijmans, E., Majundar, A., Batra, D., Lee, S.: Beyond the nav-graph: Vision and language navigation in continuous environments (2020) 10, 11
22. Ku, A., Anderson, P., Patel, R., Ie, E., Baldridge, J.: 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. pp. 4392–4412 (2020) 2, 9, 10
23. Labbé, M., Michaud, F.: Rtab-map as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation. Journal of field robotics **36**(2), 416–446 (2019) 12
24. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024) 4
25. Li, K., Meng, Z., Lin, H., Luo, Z., Tian, Y., Ma, J., Huang, Z., Chua, T.S.: Screenspot-pro: Gui grounding for professional high-resolution computer use. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 8778–8786 (2025) 4
26. Li, W., Yuan, Y., Liu, J., Tang, D., Wang, S., Qin, J., Zhu, J., Zhang, L.: Tokenpacker: Efficient visual projector for multimodal llm. International Journal of Computer Vision **133**(10), 6794–6812 (2025) 2, 4
27. Li, Z., Chen, G., Liu, S., Wang, S., VS, V., Ji, Y., Lan, S., Zhang, H., Zhao, Y., Radhakrishnan, S., et al.: Eagle 2: Building post-training data strategies from scratch for frontier vision-language models. arXiv preprint arXiv:2501.14818 (2025) 4
28. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025) 2
29. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26296–26306 (2024) 4

30. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024), https://llava-vl.github.io/blog/2024-01-30-llava-next/ 4
31. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023) 2, 4
32. Long, Y., Cai, W., Wang, H., Zhan, G., Dong, H.: Instructnav: Zero-shot system for generic instruction navigation in unexplored environment. arXiv preprint arXiv:2406.04882 (2024) 4
33. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019) 10
34. Qu, D., Song, H., Chen, Q., Yao, Y., Ye, X., Ding, Y., Wang, Z., Gu, J., Zhao, B., Wang, D., et al.: Spatialvla: Exploring spatial representations for visual-language-action model. arXiv preprint arXiv:2501.15830 (2025) 2
35. Qwen Team: Qwen3.5: Towards native multimodal agents (February 2026), https://qwen.ai/blog?id=qwen3.5 4
36. Rawles, C., Clinckemaillie, S., Chang, Y., Waltz, J., Lau, G., Fair, M., Li, A., Bishop, W., Li, W., Campbell-Ajala, F., et al.: Androidworld: A dynamic benchmarking environment for autonomous agents. In: International Conference on Learning Representations (2025) 4
37. Raychaudhuri, S., Wani, S., Patel, S., Jain, U., Chang, A.X.: Language-aligned waypoint (law) supervision for vision-and-language navigation in continuous environments. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 4018–4028 (2021) 10, 11
38. Shi, M., Liu, F., Wang, S., Liao, S., Radhakrishnan, S., Zhao, Y., Huang, D.A., Yin, H., Sapra, K., Yacoob, Y., et al.: Eagle: Exploring the design space for multimodal llms with mixture of encoders. In: International Conference on Learning Representations (2025) 2
39. Szot, A., Clegg, A., Undersander, E., Wijmans, E., Zhao, Y., Turner, J., Maestre, N., Mukadam, M., Chaplot, D., Maksymets, O., Gokaslan, A., Vondrus, V., Dharur, S., Meier, F., Galuba, W., Chang, A., Kira, Z., Koltun, V., Malik, J., Savva, M., Batra, D.: Habitat 2.0: Training home assistants to rearrange their habitat. In: Advances in Neural Information Processing Systems (2021) 9
40. Team, I.: InternVLA-N1: An open dual-system navigation foundation model with learned latent plans (2025) 2, 4, 10, 11
41. Wang, H., Liang, W., Van Gool, L., Wang, W.: Dreamwalker: Mental planning for continuous vision-language navigation (2023) 10
42. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5294–5306 (2025) 2
43. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024) 2, 4
44. Wang, S., Chen, G., Huang, D.a., Li, Z., Li, M., Liu, G., Kautz, J., Alvarez, J.M., Zhang, L., Yu, Z.: Videoitg: Multimodal video understanding with instructed temporal grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24640–24650 (2026) 4
45. Wang, S., Liu, S., Kuang, Y., Wei, X., Liu, Y., Li, Z., Man, Y., Chen, G., Tao, A., Liu, G., et al.: Locateanything: Fast and high-quality vision-language grounding with parallel box decoding. arXiv preprint arXiv:2605.27365 (2026) 2
46. Wang, Z., Li, X., Yang, J., Liu, Y., Jiang, S.: Sim-to-real transfer via 3d feature fields for vision-and-language navigation. In: Conference on Robot Learning. pp. 2982–2995. PMLR (2024) 10, 11

47. Wei, M., Wan, C., Peng, J., Yu, X., Yang, Y., Feng, D., Cai, W., Zhu, C., Wang, T., Pang, J., Liu, X.: Ground slow, move fast: A dual-system foundation model for generalizable vision-and-language navigation. In: International Conference on Learning Representations (2026) 2, 4, 10, 11
48. Wei, M., Wan, C., Yu, X., Wang, T., Yang, Y., Mao, X., Zhu, C., Cai, W., Wang, H., Chen, Y., et al.: Streamvln: Streaming vision-and-language navigation via slowfast context modeling. arXiv preprint arXiv:2507.05240 (2025) 2, 4, 10, 11
49. Xie, T., Zhang, D., Chen, J., Li, X., Zhao, S., Cao, R., Hua, T.J., Cheng, Z., Shin, D., Lei, F., et al.: Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. Advances in Neural Information Processing Systems **37**, 52040–52094 (2024) 4
50. Xin, Z., Li, W., Jiang, Y., Wang, B., Cong, R., Qin, J., Huang, S.: Decovln: Decoupling observation, reasoning, and correction for vision-and-language navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 32410–32420 (2026) 2, 10, 11
51. Xu, J., Guo, Z., Hu, H., Chu, Y., Wang, X., He, J., Wang, Y., Shi, X., He, T., Zhu, X., et al.: Qwen3-omni technical report. arXiv preprint arXiv:2509.17765 (2025) 4
52. Yu, H., Li, W., Wang, S., Chen, J., Zhu, J.: Inst3d-lmm: Instance-aware 3d scene understanding with multi-modal instruction tuning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14147–14157 (2025) 2
53. Yuan, Y., Dang, R., Li, L., Li, W., Jiao, D., Li, X., Zhao, D., Wang, F., Zhang, W., Xiao, J., et al.: Eoc-bench: Can mllms identify, recall, and forecast objects in an egocentric world? arXiv preprint arXiv:2506.05287 (2025) 2
54. Yuan, Y., Li, W., Liu, J., Tang, D., Luo, X., Qin, C., Zhang, L., Zhu, J.: Osprey: Pixel understanding with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28202–28211 (2024) 4
55. Yuan, Y., Zhang, H., Li, W., Cheng, Z., Zhang, B., Li, L., Li, X., Zhao, D., Zhang, W., Zhuang, Y., et al.: Videorefer suite: Advancing spatial-temporal object understanding with video llm. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18970–18980 (2025) 4
56. Yuan, Y., Zhang, W., Li, X., Wang, S., Li, K., Li, W., Xiao, J., Zhang, L., Ooi, B.C.: Pixelrefer: A unified framework for spatio-temporal object referring with arbitrary granularity. arXiv preprint arXiv:2510.23603 (2025) 4
57. Zeng, S., Qi, D., Chang, X., Xiong, F., Xie, S., Wu, X., Liang, S., Xu, M., Wei, X.: Janusvln: Decoupling semantics and spatiality with dual implicit memory for vision-language navigation. arXiv preprint arXiv:2509.22548 (2025) 2, 4, 10, 11
58. Zhang, J., Li, A., Qi, Y., Li, M., Liu, J., Wang, S., Liu, H., Zhou, G., Wu, Y., Li, X., et al.: Embodied navigation foundation model. arXiv preprint arXiv:2509.12129 (2025) 10, 11
59. Zhang, J., Wang, K., Wang, S., Li, M., Liu, H., Wei, S., Wang, Z., Zhang, Z., Wang, H.: Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks. Robotics: Science and Systems (2025) 4, 10, 11
60. Zhang, J., Wang, K., Xu, R., Zhou, G., Hong, Y., Fang, X., Wu, Q., Zhang, Z., Wang, H.: Navid: Video-based vlm plans the next step for vision-and-language navigation. Robotics: Science and Systems (2024) 4, 10
61. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Video instruction tuning with synthetic data (2024), https://arxiv.org/abs/2410.02713 9
62. Zheng, D., Huang, S., Li, Y., Wang, L.: Efficient-vln: A training-efficient vision-language navigation model. arXiv preprint arXiv:2512.10310 (2025) 2, 4, 10, 11

63. Zhou, G., Hong, Y., Wu, Q.: Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 7641–7649 (2024) 4