SEDUALVLN: A SPATIALLY-ENHANCED DUAL-SYSTEM FRAMEWORK FOR VISION-LANGUAGE NAVIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language Navigation (VLN) approaches have currently followed two primary paradigms: the end-to-end Vision-Language Model (VLM) policy fine-tuned on navigation trajectories to directly predict actions, and the zero-shot modular pipeline integrating pre-trained Multimodal Large Language Model (MLLM) for training-free generalization to unseen environments. However, end-to-end methods struggle with long-horizon navigation and lack dynamic reasoning ability, whereas zero-shot methods are constrained by limited spatial grounding for reliable planning and require substantial reasoning time. To bridge this gap, we introduce SEDualVLN, a spatially-enhanced dual-system VLN framework. System 1 is a VLM model enhanced with both global and local spatial awareness, tailored for high-frequency action generation. System 2 integrates a general MLLM with a mapping module, wherein the MLLM performs low-frequency waypoint planning by leveraging top-down views of the real-time 3D map alongside streams of rendered path images. Both systems exploit complementary spatial representations to improve the agent's spatial understanding in VLN tasks. Ultimately, they cooperate to complete the navigation task through a fast-slow coordinated approach. SEDualVLN achieves state-of-the-art performance on VLN-CE benchmarks, and further ablation studies demonstrate the effectiveness of each system and module.

1 INTRODUCTION

Vision-Language Navigation (VLN) Gu et al. (2022); Zhang et al. (2024d) is a challenging and crucial task in embodied AI, which requires agents to accept visual observation inputs and follow language instructions to navigate in an unseen environment. Recent advancements in Multimodal Large Language Model (MLLM) Lin et al. (2024); Zhang et al. (2024c); Yin et al. (2024) and Vision-Language-Action (VLA) Zhang et al. (2024a); Cheng et al. (2024); Zheng et al. (2024) model have significantly increased the appeal of VLN. In particular, advancing explicit spatial reasoning to accurately align language instructions with continuous 3D geometry and environmental topology has emerged as a critical prerequisite for achieving robust, long-horizon navigation in complex settings.

Existing VLN methods are generally divided into two main paradigms: **the end-to-end paradigm** Zhang et al. (2024a); Cheng et al. (2024); Zhang et al. (2024b); Goetting et al. (2024); Song et al. (2025); Wei et al. (2025b); Zeng et al. (2025); Liu et al. (2026b); Li et al. (2026) and **the zero-shot paradigm** Liu et al. (2026a); Yue et al. (2026); Qiao et al. (2025); Chen et al. (2025); Zhang et al. (2026); Jiang & Wang (2025); Lyu et al. (2026); Long et al. (2024); Wang et al. (2025b). The end-to-end VLN paradigm typically employs a unified vision-language policy, fine-tuned on expert navigation trajectories to directly predict sequential actions. While effective in constrained or seen environments, it struggles with long-horizon adaptation and lacks the dynamic, step-wise reasoning required for complex multi-step decision-making. Conversely, zero-shot pipelines leverage pre-trained MLLM to enable training-free generalization across unseen scenes. However, their planning reliability is inherently constrained by the limited spatial grounding and geometric reasoning capabilities of off-the-shelf foundation models. Consequently, existing methodologies face complementary limitations that hinder robust navigation in complex open-world scenarios.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

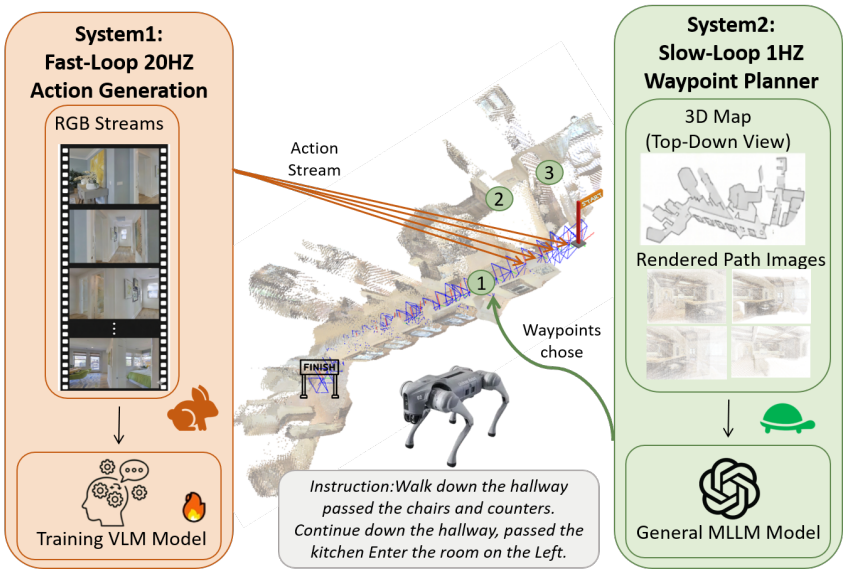

Figure 1: Overview of SEDualVLN. System 1 (orange) generates fast, low-level actions from RGB streams via a VLM, while System 2 (green) plans waypoints using top-down 3D map and rendering path images through a general MLLM. The systems cooperate in a fast-slow coordinated loop.

Although recent studies Wei et al. (2025a); Zhong et al. (2026) have proposed dual-system approaches to compensate for the limitations of individual-systems, these methods merely provide a coarse integration of the two fundamental paradigms, without fundamentally enhancing the capabilities of the underlying systems. Specifically, compared to semantic capabilities, spatial capabilities are the most abstract and easily overlooked in VLN tasks, yet spatial awareness can directly improve the agent's robustness and success rate in long-horizon navigation.

First, to address the limitations of the single system in VLN, we introduce a dual-system with fast–slow cooperation to jointly complete navigation tasks. Next, to enhance the spatial awareness of the underlying systems in spatial navigation, we equip System 1 with Global and Local Spatial Enhancement Strategies, enabling the VLM model to comprehend the environmental structure from coarse to fine. In System 2, we adopt a three-stage "Mapping–Rendering–Reasoning" pipeline, which provides the MLLM with reasoning cues from both global and dynamic perspectives.

Our contributions are as follows:

1. **Dual-System VLN Framework:** We propose SEDualVLN, a dual-system architecture that combines a fast VLM-based action generator with a slow MLLM-based planner, coordinated by a scheduler, to jointly perform complex navigation tasks.

2. **Spatially-Enhanced VLM (System 1):** We introduce Global and Local Spatial Enhancement strategies for the VLM model, enabling coarse-to-fine comprehension of the environment and improving spatial awareness for robust long-horizon navigation.

3. **Mapping–Rendering–Reasoning MLLM (System 2):** We design a three-stage pipeline that provides the MLLM with both global and dynamic visual cues, enhancing spatial reasoning and waypoint planning in unseen environments.

4. **Comprehensive Evaluation:** We demonstrate that SEDualVLN achieves state-of-the-art performance on VLN-CE benchmarks and conduct ablation studies to validate the effectiveness of each system and module in improving navigation success and robustness.

2 Related Work

2.1 Vision-Language Navigation (VLN)

Vision-Language Navigation (VLN) requires an embodied agent to follow natural language instructions to navigate in unseen photorealistic 3D environments Anderson et al. (2018); Qi et al. (2020); Krantz et al. (2020); Ku et al. (2020). Early studies mainly focused on discrete settings (e.g.,R2R Anderson et al. (2018)), emphasizing cross-modal alignment via attention mechanisms and topological memory for long-term planning Tan et al. (2019); Hong et al. (2021). Recent advances Cheng et al. (2024); Wei et al. (2025b;a) have shifted toward continuous environments (e.g.,VLN-CE Krantz et al. (2020); Ku et al. (2020)), exploring pretraining, data augmentation, and online adaptation techniques to improve generalizability.

End-to-end VLN approaches integrate visual perception, language understanding, and action prediction in a unified framework by fine-tuning Video-LLMs Zhang et al. (2024c); Lin et al. (2024) or Vision-Language-Action (VLA) models on large-scale navigation data. These methods Zhang et al. (2024b;a); Cheng et al. (2024); Wei et al. (2025b) leverage video stream encoding, spatio-temporal context modeling, and techniques such as token pruning or DAgger Ross et al. (2011) to generate low-level actions directly from RGB observations Zhang et al. (2024b); Wei et al. (2025b); Liu et al. (2026b). While achieving strong benchmark performance, they often struggle with long-horizon navigation due to accumulating context length and redundant historical token processing, and exhibit limited dynamic reasoning in complex scenes.

Zero-shot VLN approaches employ powerful multimodal LLMs (e.g., GPT-4o OpenAI (2024)) as planners with carefully designed prompts and Chain-of-Thought reasoning, often in a training-free manner. Dual-process frameworks further combine lightweight expert models for routine navigation with LLMs for anomalous cases. These methods Yue et al. (2026); Chen et al. (2025); Zhong et al. (2026) excel at commonsense reasoning and generalization but lack fine-grained spatial and geometric understanding, leading to suboptimal decisions in precise spatial correlation and continuous video streams.

2.2 Spatial Reasoning in VLN

Spatial reasoning is fundamental to VLN, requiring precise alignment between linguistic instructions, environmental topology, and continuous 3D geometry. Recent video-based VLMs and VLA models Zhang et al. (2024b); Wei et al. (2025b); Zeng et al. (2025) enhance spatial retention through temporal context modeling and memory decoupling. However, their spatial representations remain largely implicit, making them susceptible to metric drift and poor interpretability in complex layouts. To mitigate this, explicit spatial reasoning has emerged as a critical direction. Topological planners Chen et al. (2022) construct evolving connectivity graphs with dual-scale transformers for global action planning, while geometry-aware methods Huo et al. (2023) integrate depth/normal maps with slot attention to learn geometry-enhanced visual representations. Dedicated spatial grounding frameworks like Spatial-VLN Yue et al. (2026) and SpatialNav Zhang et al. (2026) explicitly model geometric constraints and spatial relations through structured perception and scene graphs. Concurrently, LLM-driven approaches Zhong et al. (2026); Wei et al. (2025a) decompose navigation into deliberative spatial planning via chain-of-thought or constraint-aware filtering. Despite their logical strength, these methods predominantly operate over abstract textual cues or fragmented 2D proxies, lacking direct access to continuous, physically consistent 3D geometry—frequently inducing spatial hallucination and degrading long-horizon robustness.

To bridge this gap, we propose a spatially-enhanced dual-system architecture. System 1 explicitly models spatial semantics and connectivity to establish a robust topological prior for efficient path grounding. System 2 leverages 3D mapping to render observed trajectories, providing high-fidelity, multi-view-consistent 3D context for deliberative planning. By unifying explicit topological reasoning with physically grounded 3D observation, our framework achieves more accurate, interpretable, and robust spatial reasoning compared to prior implicit or abstraction-dependent methods.

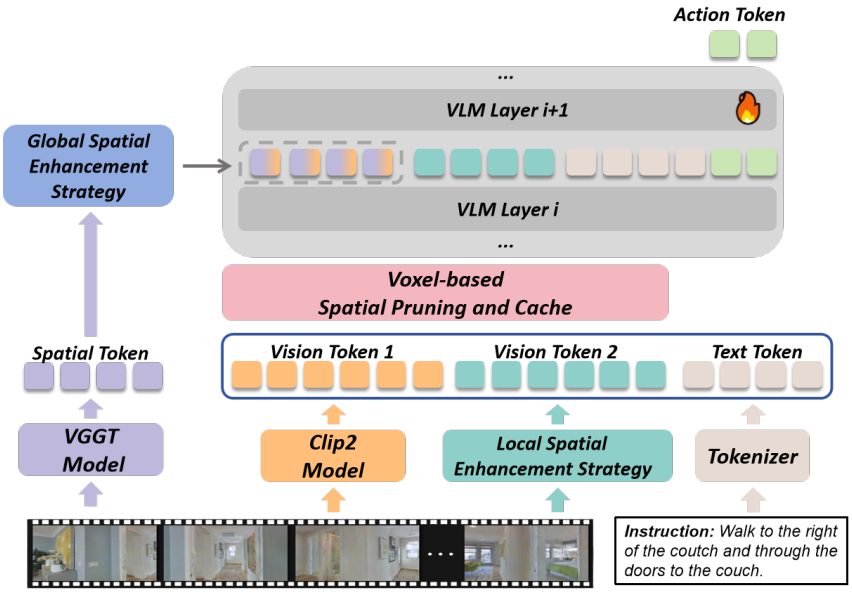

Figure 2: Framework of System 1. Given RGB image streams and language instructions, the model extracts visual, spatial, and text tokens. During training, a frozen VGGT model provides feature-level spatial supervision for the visual tokens. The remaining tokens are processed by voxel-based spatial pruning and caching before being fed into the VLM.

3 METHOD

We formulate continuous VLN task as a sequential decision-making problem under monocular RGB perception. At each step t, the agent processes a language instruction I alongside a trajectory of ego-centric observations $O_t = \{x_0, \ldots, x_t\}$, where x_τ is the visual input at time τ. To navigate successfully, the agent must learn a policy $\pi(O_t, I)$ to dictate the next atomic action $a_{t+1} \in \mathcal{A}$ based on its visual history and the given instruction. The discrete action space is restricted to $\mathcal{A} = \{MoveForward, TurnLeft, TurnRight, Stop\}$. An episode is deemed a success if the agent issues a $Stop$ command within a proximity threshold of 3m from the target place, without exceeding a maximum budget of 500 steps.

As illustrated in Figure 1, SEDualVLN is a dual-system VLN framework. System 1 (Sec. 3.1), is a VLM model that generates actions based on video stream input. System 2 (Sec. 3.2), consists of a general MLLM model and mapping module that determines optimal navigation waypoints using 3D mapping overhead view and frontier-path rendering visualizations. Systems 1 and 2 operate at different frequencies to achieve a fast-slow coordinated approach: while System 2 requires more time for waypoint-level reasoning, System 1 executes multiple low-level actions during this period. Once System 2 completes its planning, the agent proceeds to execute the planned waypoints, thereby mitigating delays from lengthy reasoning while still incorporating global spatial information.

3.1 SYSTEM1: SPATIALLY-ENHANCED VLM MODEL

The overall framework of System 1 is illustrated in Figure 2. Our model is built upon the StreamVLN Wei et al. (2025b) backbone, adopting its multi-turn dialogue mechanism, KV cache, and spatial token pruning. We introduce two major modifications, enhancing its spatial awareness for long-range tasks. First, we introduce a global spatial enhancement strategy that encourages the VLM's visual tokens to encode spatial information without requiring explicit depth-map inputs. Second, we introduce Local Spatial Enhancement Strategy, enabling the VLM to focus more effectively on the connectivity of nearby passages and thereby reducing the risk of taking incorrect branches.

Global Spatial Enhancement Strategy Compared to approaches that provide explicit depth maps or reconstructed 3D geometry as inputs, we enhance spatial awareness through feature-level super-

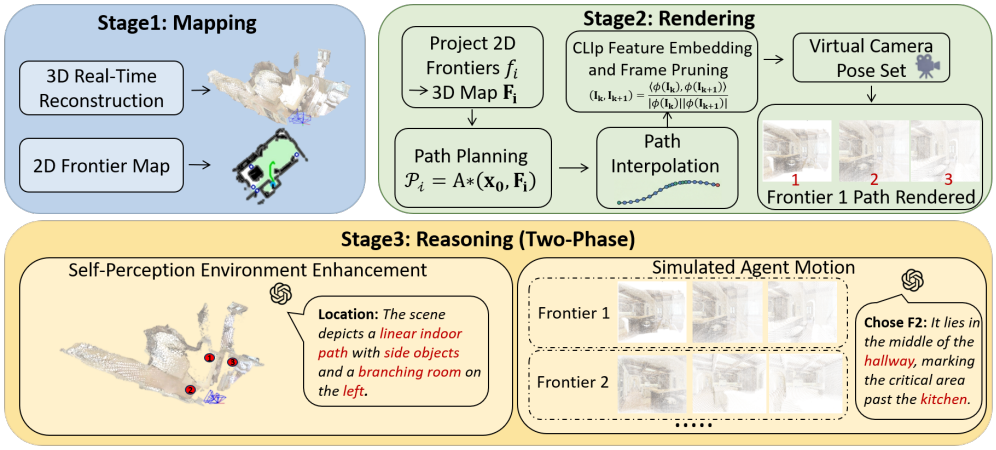

Figure 3: System 2 Workflow. It consists of three stages: Mapping, which builds a 3D map and 2D frontier map; Rendering, which generates virtual camera views along planned paths; and Reasoning, where the agent first interprets the environment and then evaluates frontier paths against the instruction to select the optimal navigation target.

vision. Specifically, during training, we use a frozen pretrained VGGT model to provide spatially aware feature targets for the VLM's visual tokens. Prior work Huang et al. (2024) has shown that supervising relatively deep but not the deepest layers is most effective for improving action performance, likely because the deeper layers lose more low-level visual features, making it more difficult for them to receive supervision targeted at spatial representations. Therefore, we apply supervision at intermediate attention fusion layers of LLaVA-Video Zhang et al. (2024c) (the specific layers are selected based on ablation studies).

For each frame along the agent's navigation sequence, visual token embeddings are projected through a two-layer MLP with batch normalization to align with the spatial representation dimension. During training, these projected embeddings are supervised against spatial features generated by a frozen pretrained VGGT model Wang et al. (2025a), with positional encodings added to preserve spatial ordering. The supervision is implemented by minimizing the cosine distance between the projected visual tokens and the corresponding spatial representations, encouraging the VLM to encode geometric and structural cues in its visual tokens. VGGT is used only to generate the training targets and is not required during inference. Formally, let $\mathbf{V}_t$ denote the projected visual token at frame t, $\mathbf{S}_t$ the spatial representation from the 3D model, and $\mathbf{p}_t$ the positional embedding; the combined training objective is defined as

$$\mathcal{L}_{\text{SE}} = \mathcal{L}_{\text{action}} + \alpha \cdot \frac{1}{N} \sum_{t=1}^{N} \Big[1 - \cos\left(\mathbf{V}_t, \mathbf{S}_t + \mathbf{p}_t\right) \Big], \tag{1}$$

where N is the number of frames and α balances the action prediction loss and the spatial supervision loss.

Local Spatial Enhancement Strategy. In VLN tasks, it is common for an agent to move from one region to another, where passages play a critical role as the connecting structures between regions. Most failures of VLM occur when the agent enters an incorrect passage, making it unable to return to the correct trajectory. Extracting global visual tokens from the current RGB frame using a 3D model is a common approach; however, such visual tokens often struggle to capture passage existence due to interference from numerous irrelevant tokens. To address this issue, we propose a channel connectivity extraction module, designed to enable the VLM to explicitly focus on passage connectivity. This module operates as follows: (1) the presence of passages (e.g., corridors, doors) is detected using an open-vocabulary Grounding DINO Liu et al. (2024) approach; (2) the detected passages are then segmented via SAM Kirillov et al. (2023) to obtain binary masks; (3) all passage masks are set to 1 while non-passage regions are set to 0, completing the image encoding. Finally,

Table 1: Performance comparison of VLN methods on R2R and RxR Val-Unseen splits. * indicates methods using the waypoint predictor.

Method	Observation Encoder				R2R Val-Unseen				RxR Val-Unseen			
	Pano.	Odo.	Depth	S.RGB	NE↓	OS↑	SR↑	SPL↑	NE↓	SR↑	SPL↑	nDTW↑
CMA* Hong et al. (2022)	✓	✓	✓		6.20	52.0	41.0	36.0	8.76	26.5	22.1	47.0
BERT* Hong et al. (2022)	✓	✓	✓		5.74	53.0	44.0	39.0	8.98	27.0	22.6	46.7
ETPNav* An et al. (2025)	✓	✓	✓		4.71	65.0	57.0	49.0	5.64	54.7	44.8	61.9
LAWRaychaudhuri et al. (2021)		✓	✓	✓	6.83	44.0	35.0	31.0	10.90	8.0	8.0	38.0
Seq2Seq Krantz et al. (2020)			✓	✓	7.77	37.0	25.0	22.0	12.10	13.9	11.9	30.8
NaVILA Cheng et al. (2024)				✓	5.22	62.5	54.0	49.0	6.77	49.3	44.0	58.8
StreamVLN Wei et al. (2025b)				✓	4.98	64.2	56.9	51.9	6.22	52.9	46.0	61.9
DualVLN Wei et al. (2025a)				✓	4.05	70.7	64.3	58.5	4.58	61.4	51.8	70.0
SEDualVLN(ours)				✓	**3.75**	**73.7**	**67.3**	**62.5**	**4.12**	**63.9**	**52.4**	**72.8**

MLP is employed to project the resulting representations to the same dimensionality as the global visual tokens.

3.2 SYSTEM2: MAPPING–RENDERING–REASONING MLLM

The overall workflow of System 2 is illustrated in Figure 3, comprising three main stages: Mapping, Rendering, and Reasoning.

Mapping. We construct a lightweight 3D map of the environment and a 2D frontier map in real-time. The 2D frontier map guides the agent to explore new regions, following the construction rules detailed in VLFM Yokoyama et al. (2024). Multi-view 3D mapping is computationally expensive and requires many agent steps. To alleviate this, we adopt a feed-forward approach based on LingBot-Map Chen et al. (2026), which incrementally builds the 3D map online.

Rendering. Let $\mathcal{F} = \{\mathbf{f}_1, \ldots, \mathbf{f}_n\}$ denote the set of frontier points in 2D coordinates. Each frontier point $\mathbf{f}_i$ is projected into the 3D map to obtain its corresponding 3D location $\mathbf{F}_i$. For each $\mathbf{F}_i$, we compute a collision-free path from the agent's current position $\mathbf{x}_0$ to $\mathbf{F}_i$ using the A* algorithm: $\mathcal{P}_i = \text{A*}(\mathbf{x}_0, \mathbf{F}_i)$.

To generate a dense set of rendering views along the path, we perform interpolation between consecutive path nodes. Specifically, if the Euclidean distance between two consecutive points $\mathbf{p}_j$ and $\mathbf{p}_{j+1}$ exceeds a threshold d, we insert intermediate points such that the spacing satisfies $\|\mathbf{p}_{j+1} - \mathbf{p}_j\| < d$. Each interpolated point corresponds to a virtual camera pose, and the RGB image at that pose is rendering as $\mathbf{I}_k$. To reduce redundant information and MLLM token usage, we apply a cosine-similarity based pruning:

$$s(\mathbf{I}_k, \mathbf{I}_{k+1}) = \frac{\langle \phi(\mathbf{I}_k), \phi(\mathbf{I}_{k+1}) \rangle}{\|\phi(\mathbf{I}_k)\| \|\phi(\mathbf{I}_{k+1})\|} < \tau \implies \text{keep } \mathbf{I}_{k+1}, \tag{2}$$

where $\phi(\cdot)$ denotes a CLIP Radford et al. (2021) encoder embedding and τ is a similarity threshold. After sampling and redundancy removal, we obtain m RGB frames $\{\mathbf{I}_1, \ldots, \mathbf{I}_m\}$ for each path.

Reasoning. The reasoning stage operates in two phases. In the first phase, self-perception environment enhancement, the GPT-4o OpenAI (2024) is provided with the 3D top-down map and is required to describe the agent's current location and nearby environment. In the second phase, simulated agent motion, the GPT-4o evaluates each frontier path by comparing the Rendering image sequences $\{\mathbf{I}_1, \ldots, \mathbf{I}_m\}$ with the natural language instruction. It then selects the frontier point whose path best aligns with the instruction. This two-stage process allows the agent to integrate spatial context and visual guidance when making navigation decisions.

Table 2: Effect of Global Spatial Enhancement Strategy on VLN Performance.

Target Rep.	Layer	R2R Val-Unseen				RxR Val-Unseen			
		NE↓	OS↑	SR↑	SPL↑	NE↓	SR↑	SPL↑	nDTW↑
×	×	4.10	67.3	61.3	58.7	4.43	58.5	49.5	64.3
VGGT	1	3.95	72.5	66.4	60.4	4.23	62.8	49.1	70.5
VGGT	8	3.88	72.7	66.8	61.6	4.17	63.5	51.8	72.3
VGGT	16	3.80	73.4	67.0	62.3	4.15	63.7	52.1	72.5
VGGT	**24**	**3.75**	**73.7**	**67.3**	**62.5**	**4.12**	**63.9**	**52.4**	**72.8**
VGGT	32	3.84	72.8	66.7	61.7	4.18	63.4	51.9	72.6

Table 3: Effect of Local Spatial Enhancement Strategy (LSES) on VLN Performance.

Method	R2R Val-Unseen				RxR Val-Unseen			
	NE↓	OS↑	SR↑	SPL↑	NE↓	SR↑	SPL↑	nDTW↑
w/o LSES	3.95	70.3	63.8	61.5	4.23	59.8	51.0	66.5
SEDualVLN (Full)	**3.75**	**73.7**	**67.3**	**62.5**	**4.12**	**63.9**	**52.4**	**72.8**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Simulation Benchmark Setup. We evaluate our method on two general VLN-CE benchmarks, R2R-CE Krantz et al. (2020) and RxR-CE Ku et al. (2020). Both datasets are based on Matterport3D reconstructions, with navigation episodes exceeding 10 meters. The agent is allowed to navigate freely using low-level actions, such as moving forward 0.25 meters or turning left 15 degrees, rather than relying on fixed navigation graph nodes. All experiments are evaluated using standard metrics, including Navigation Error (NE), Success Rate (SR), Oracle Success Rate (OS), and Success weighted by Path Length (SPL), normalized Dynamic Time Warping (nDTW), following prior works.

Base Models and Implementation Details. We build System 1 based on the LLaVA-Video Zhang et al. (2024c) 7B model, which employs Qwen2-7B Hui et al. (2024) as the language backbone. We perform several additional modifications: first, during training, we reconstruct the loss function by incorporating a spatial alignment loss; second, we use Local Spatial Enhancement Strategy and extract additional visual tokens from the RGB images. Initially, we fine-tune the model for one epoch using only the oracle VLN trajectories. Subsequently, we use this model to collect DAgger trajectories and continue training for another epoch using both the VLN data and general multimodal data. Each training step processes 128 video clips. The entire training procedure takes approximately 2000 A100 GPU hours. System 1 runs using a single RTX 4090 GPU. System 2 requires no training; it utilizes the pre-trained 3D model LingBot-Map Chen et al. (2026) for 3D mapping, and its deployment also occupies one RTX 4090 GPU.

4.2 EXPERIMENT RESULT

Quantitative analysis. As shown in Table1, we compare our SEDualVLN with previous state-of-the-art methods on the two splits of VLN-CE. These methods can be categorized into four types: (1) multi-sensor-based approaches (CMA Hong et al. (2022), BERT Hong et al. (2022), ETPNav An et al. (2025), Seq2Seq Krantz et al. (2020)), (2) VLM methods using first-person RGB-D images (LAW Raychaudhuri et al. (2021)), (3) VLM methods using only first-person RGB images (NaVILA Cheng et al. (2024), StreamVLN Wei et al. (2025b)), and (4) dual-system architecture methods (DualVLN Wei et al. (2025a)). Experimental results show that, without relying on additional data sources, our method achieves a SR improvement of 3% and 2.5% over DualVLN on the two splits, respectively, reaching a new state-of-the-art level.

Qualitative analysis. As shown in Figure 7, we compare with the state-of-the-art method StreamVLN. The results demonstrate that our model significantly reduces the probability of taking wrong turns at junctions, indicating enhanced spatial awareness. As shown in Figure 8, we visualize navigation and mapping visualization of our SEDualVLN. For more experimental visualization results, please refer to the supplementary video material.

4.3 Ablation Study

To demonstrate that each system and module of SEDualVLN is indispensable, we conduct extensive ablation studies, including: (1) evaluating the effect of spatial enhancement strategy at different attention layers of the VLM on the navigation performance; (2) assessing the effect of System 1's channel connectivity extraction module; (3) comparing the navigation results when System 1 and System 2 are run independently versus collaboratively; and (4) analyzing the impact of different general-purpose MLLM.

Table 4: Impact of System 1 and System 2 Collaboration on VLN.

Method	Freq. (S1:S2)	R2R Val-Unseen					RxR Val-Unseen				
		NE↓	OS↑	SR↑	SPL↑	AT↓	NE↓	SR↑	SPL↑	nDTW↑	AT↓
(a) system1	-	4.25	68.2	62.8	58.7	**48.9**	4.43	58.5	49.5	64.3	**53.6**
(b) system2	-	4.34	65.7	59.8	57.9	294.3	4.53	56.6	49.1	70.5	313.8
(c) system1+system2	10:1	3.83	72.7	67.0	61.6	59.9	4.17	63.5	51.8	72.3	67.3
(d) system1+system2	**20:1**	**3.75**	**73.7**	**67.3**	**62.5**	57.3	**4.12**	**63.9**	**52.4**	**72.8**	63.9
(e) system1+system2	30:1	3.80	72.3	66.5	60.9	52.8	4.15	63.4	52.1	72.5	59.8

Table 5: Effect of different general-purpose MLLM on VLN.

Method	R2R Val-Unseen				RxR Val-Unseen			
	NE↓	OS↑	SR↑	SPL↑	NE↓	SR↑	SPL↑	nDTW↑
Gemini Team (2023)	3.82	73.4	66.9	61.9	4.19	63.4	52.1	71.9
Qwen2.5-VL Bai et al. (2025)	3.78	73.6	67.1	62.1	4.17	63.7	52.2	72.4
GPT-4o OpenAI (2024)	**3.75**	**73.7**	**67.3**	**62.5**	**4.12**	**63.9**	**52.4**	**72.8**

Effect of Global Spatial Enhancement Strategy at Different Attention Layers of the VLM. We analyze whether the spatial enhancement strategy effectively improves the agent's global spatial awareness, and we further study its effect when supervised at different layers of the VLM model. The experiments are conducted on two datasets for both training and evaluation. As shown in Table 2, the results indicate that supervising the Global Spatial Enhancement Strategy at layer 24 yields the best performance. Specifically, compared to the model without supervision, improving the SR by 6.0% and SPL by 3.8% on the R2R Val-Unseen split, and the SR by 14.4% and SPL by 11.3% on the RxR Val-Unseen split. These results indicate that supervising the Spatial Enhancement Strategy at mid-to-deep layers is most effective in enhancing the agent's spatial.

Effect of Local Spatial Enhancement Strategy. As shown in Table 3. Incorporating Local Spatial Enhancement Strategy (LSES) consistently improves navigation metrics across both R2R and RxR Val-Unseen sets. On R2R, NE is reduced by 0.20, SR improves by 3.5%, SPL increases by 1.0% , and OS rises by 3.4%. For RxR, NE decreases by 0.11, SR improves by 4.1%, and nDTW increases significantly by 6.3%. The SPL increases by 1.4% Overall, these results suggest that agents benefit from LSES during navigation, and LSES significantly improves success rates at decision points such as intersections.

Impact of System 1 and System 2 Collaboration on VLN. As shown in Table 4, to further demonstrate the effectiveness of the dual-system approach, we introduce an additional metric, Average Time (AT, in seconds), which measures the mean time taken by the agent to complete a navigation eposide. Comparing (a) and (d), we observe that the dual-system achieves significantly higher success rates than the pure VLM system, with SR improvements of 4.5% and 5.4%, albeit at the cost



Figure 4: Real-world experiments using the Unitree G2 robot. Guided by natural language instructions (top), SEDualVLN enables the robot to successfully navigate complex campus environments to reach specific targets (e.g., the red barrier and the bench).

of slightly increased time. Comparing (b) and (d), the dual-system outperforms the general MLLM single-system, with SR gains of 7.5% and 7.3%, while the average completion time is reduced by a factor of six. Additionally, we examine the metrics under different switching frequency ratios, as shown in (c), (d), and (e), and find that a frequency ratio of 20:1 allows our SEDualVLN to best balance SR and time efficiency.

Effect of different general-purpose MLLM on VLN. As shown in Table5, using Gemini results in two benchmarks SR of 66.9% and 63.4%. The Qwen2.5-VL model performs slightly better, achieving an SR of 67.1% and 63.7%. However, GPT-4o demonstrates the highest efficacy. This indicates that GPT-4o provides superior spatial reasoning and zero-shot generalization capabilities.

4.4 Real-World Experiments

To validate the effectiveness of our proposed SEDualVLN, we deployed a Unitree G2 robot in complex campus environments with dynamic obstacles. As shown in Figure 4, driven by our SEDualVLN, the robot successfully executed multi-step natural language instructions (e.g., navigating to a red barrier or a bench) by accurately perceiving landmarks and avoiding pedestrians in real-time. By seamlessly integrating high-level semantic planning with low-level locomotion control, SEDualVLN enabled the robot to complete all tasks without human intervention. These real-world deployments conclusively demonstrate that our method effectively bridges the sim-to-real gap, ensuring robust and precise navigation in unstructured physical environments.

5 Limitation

Our main limitation lies in deployment on physical devices. This is because the latest real-time 3D mapping features still face constraints in terms of real-world real-time performance, which leads to distorted 3D mapping results and consequently impacts the entire system.

6 Conclusion

We presented SEDualVLN, a spatially-enhanced dual-system Vision-Language Navigation framework that successfully integrates the strengths of end-to-end VLM and zero-shot MLLM pipeline. To overcome the challenges of long-horizon navigation and limited spatial grounding, SEDualVLN employs a fast-slow coordinated approach. Both systems are explicitly designed to prioritize and enhance spatial awareness. Our results on the VLN-CE benchmarks establish a new state-of-the-art. By cultivating a stronger sense of direction and dynamic reasoning capabilities, SEDualVLN paves the way for more robust and capable navigation agents in complex, unseen environments.

AI USE STRATEGYMENT

In this work, we used generative AI tools (e.g., large language models) for manuscript language editing, grammar refinement, and code debugging. We have not used generative AI tools for generating core research ideas, designing methodologies, or creating experimental data. We have thoroughly reviewed all AI-assisted text and code to ensure their accuracy, correctness, and originality. We take full responsibility for the final content of this work, including all text, claims, and artifacts.

REFERENCES

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 (TPAMI)*, pp. 5130–5145, 2025.

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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025.

Kehan Chen, Dong An, Yan Huang, Rongtao Xu, Yifei Su, Yonggen Ling, Ian Reid, and Liang Wang. Constraint-aware zero-shot vision-language navigation in continuous environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 47(11):10441–10456, 2025.

Lin-Zhuo Chen, Jian Gao, Yihang Chen, Ka Leong Cheng, Yipengjing Sun, Liangxiao Hu, Nan Xue, Xing Zhu, Yujun Shen, Yao Yao, et al. Geometric context transformer for streaming 3d reconstruction. *arXiv preprint arXiv:2604.14141*, 2026.

Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16537–16547, 2022.

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

Dylan Goetting, Himanshu Gaurav Singh, and Antonio Loquercio. End-to-end navigation with vision language models: Transforming spatial reasoning into question-answering. *arXiv preprint arXiv:2411.05755*, 2024.

Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Wang. Vision-and-language navigation: A survey of tasks, methods, and future directions. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7606–7623, 2022.

Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. Vln bert: A recurrent vision-and-language bert for navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1643–1653, 2021.

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 (CVPR)*, pp. 15439–15449, 2022.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Deciphering cross-modal alignment in large vision-language models with modality integration rate. *arXiv preprint arXiv:2410.07167*, 2024.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2.5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

Jingyang Huo, Qiang Sun, Boyan Jiang, Haitao Lin, and Yanwei Fu. Geovln: Learning geometry-enhanced visual representation with slot attention for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23212–23221, 2023.

Zhiqiang Jiang and Xin Wang. Spatialgpt: Zero-shot vision-and-language navigation via spatial cot over structured spatial memory. In *Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems*, pp. 423–435, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023.

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

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)*, pp. 4392–4412, 2020.

Tianfu Li, Wenbo Chen, Haoxuan Xu, Xinhu Zheng, and Haoang Li. P^3nav: End-to-end perception, prediction and planning for vision-and-language navigation. *arXiv preprint arXiv:2603.17459*, 2026.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5971–5984, 2024.

Chenghao Liu, Zhimu Zhou, Jiachen Zhang, Minghao Zhang, Songfang Huang, and Huiling Duan. Msnav: Zero-shot vision-and-language navigation with dynamic memory and llm spatial reasoning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 20112–20116, 2026a.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision (ECCV)*, pp. 38–55, 2024.

Xiangchen Liu, Hanghan Zheng, Jeil Jeong, Minsung Yoon, Lin Zhao, Zhide Zhong, Haoang Li, and Sung-Eui Yoon. Dygeovln: Infusing dynamic geometry foundation model into vision-language navigation. *arXiv preprint arXiv:2603.21269*, 2026b.

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.

Kailin Lyu, Kangyi Wu, Pengna Li, Xiuyu Hu, Qingyi Si, Cui Miao, Ning Yang, Zihang Wang, Long Xiao, Lianyu Hu, Jingyuan Sun, and Ce Hao. Himemvln: Enhancing reliability of open-source zero-shot vision-and-language navigation with hierarchical memory system. *arXiv preprint arXiv:2603.14807*, 2026.

OpenAI. Gpt-4o system card, 2024.

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 (CVPR)*, 2020.

Yanyuan Qiao, Wenqi Lyu, Hui Wang, Zixu Wang, Zerui Li, Yuan Zhang, Mingkui Tan, and Qi Wu. Open-nav: Exploring zero-shot vision-and-language navigation in continuous environment with open-source llms. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6710–6717, 2025.

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 (ICML)*, pp. 8748–8763, 2021.

Sonia Raychaudhuri, Saim Wani, Shivansh Patel, Unnat Jain, and Angel Chang. Language-aligned waypoint (LAW) supervision for vision-and-language navigation in continuous environments. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4018–4028, 2021.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 627–635. PMLR, 2011.

Xinshuai Song, Weixing Chen, Yang Liu, Weikai Chen, Guanbin Li, and Liang Lin. Towards long-horizon vision-language navigation: Platform, benchmark and method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12078–12088, 2025.

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 (NAACL)*, pp. 2610–2621, 2019.

Gemini Team. Gemini: A family of highly capable multimodal models, 2023.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025a.

Yunheng Wang, Yuetong Fang, Taowen Wang, Yixiao Feng, Yawen Tan, Shuning Zhang, Peiran Liu, Yiding Ji, and Renjing Xu. Dreamnav: A trajectory-based imaginative framework for zero-shot vision-and-language navigation, 2025b.

Meng Wei, Chenyang Wan, Jiaqi Peng, Xiqian Yu, Yuqiang Yang, Delin Feng, Wenzhe Cai, Chenming Zhu, Tai Wang, Jiangmiao Pang, and Xihui Liu. Ground slow, move fast: A dual-system foundation model for generalizable vision-and-language navigation, 2025a.

Meng Wei, Chenyang Wan, Xiqian Yu, Tai Wang, Yuqiang Yang, Xiaohan Mao, Chenming Zhu, Wenzhe Cai, Hanqing Wang, Yilun Chen, Xihui Liu, and Jiangmiao Pang. Streamvln: Streaming vision-and-language navigation via slowfast context modeling. *arXiv preprint arXiv:2507.05240*, 2025b.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 2024.

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

Lu Yue, Yue Fan, Shiwei Lian, Yu Zhao, Jiaxin Yu, Liang Xie, and Feitian Zhang. Spatial-vln: Zero-shot vision-and-language navigation with explicit spatial perception and exploration. *arXiv preprint arXiv:2601.12766*, 2026.

Shuang Zeng, Dekang Qi, Xinyuan Chang, Feng Xiong, Shichao Xie, Xiaolong Wu, Shiyi Liang, Mu Xu, Xing Wei, and Ning Guo. Janusvln: Decoupling semantics and spatiality with dual implicit memory for vision-language navigation. *arXiv preprint arXiv:2509.22548*, 2025.

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

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

Jiwen Zhang, Zejun Li, Siyuan Wang, Xiangyu Shi, Zhongyu Wei, and Qi Wu. Spatialnav: Leveraging spatial scene graphs for zero-shot vision-and-language navigation. *arXiv preprint arXiv:2601.06806*, 2026.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024. *arXiv preprint arXiv:2410.02713*, 2024c.

Yue Zhang, Ziqiao Ma, Jialu Li, Yanyuan Qiao, Zun Wang, Joyce Chai, Qi Wu, Mohit Bansal, and Parisa Kordjamshidi. Vision-and-language navigation today and tomorrow: A survey in the era of foundation models, 2024d.

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

Yu Zhong, Zihao Zhang, Rui Zhang, Lingdong Huang, Haihan Gao, Shuo Wang, Da Li, Ruijian Han, Jiaming Guo, Shaohui Peng, Di Huang, and Yunji Chen. Run, ruminate, and regulate: A dual-process thinking system for vision-and-language navigation. In *AAAI*, pp. 18845–18854, 2026.

A TECHNICAL DETAILS

In this section, we first build upon the previously mentioned qualitative experiment Section 4.2 to further illustrate how the two-stage reasoning of the MLLM in System 2 is performed. We then provide the prompts used in both System 1 and System 2. For complete video visualizations and additional experimental segments, please refer to the supplementary video material.

As shown in Figure 5 and Figure 6, the reasoning of the MLLM is divided into two stages. In the first stage, we guide the MLLM to analyze the environmental structure, object relationships, and possible orientations based on the top-down map. In the second stage, we provide the MLLM with additional cues through Rendering views, enabling it to complete the selection of waypoints.

Below are prompt templates for VLM and MLLM.

VLM Prompts

You are an autonomous navigation assistant. Your task is to ¡Instruction¿. Generate a sequence of actions to follow this instruction using the available commands: TURN LEFT (←) or TURN RIGHT (→) by 15 degrees, MOVE FORWARD (↑) by 25 centimeters, or STOP.
Input: [Images]

Model Output: JSON List
[
"Action": "←↑↑→"
]

MLLM Reasoning Stage1 Prompts

Task: You are an Environment Understanding Assistant. Analyze the 3D Top-Down map and the accompanying language instruction. Summarize the environment information for navigation planning.

Input: [3D Top-Down Map + Language Instruction]

Constraints:

- *Extract the key spatial information of the environment.*
- *Focus on **Location**, **Relationship**, and **Possible directions**.*
- *Keep the output concise and structured.*

Model Output: JSON
[
"Location": "The agent is navigating an indoor environment resembling a kitchen-dining room layout.",
"Relationship": "Ovens are located on the left side, with dining table and chairs positioned along the main corridor.",
"Possible directions": "The agent's normal path is likely to follow the main corridor from the kitchen area toward the dining room, while staying aligned with walls and furniture."
]

MLLM Reasoning Stage2 Prompts

Task: You are a Navigation Planner Assistant. Analyze the images of each waypoint (Frontier 1, Frontier 2, Frontier 3, etc.) and select the most suitable waypoint for the agent to proceed. Provide clear reasoning for your choice based on visual cues such as corridor alignment, furniture arrangement, and spatial orientation.

Input: [Waypoint Images for Frontier 1, Frontier 2, Frontier 3, ...]

Constraints:

- *Evaluate each frontier marked in the images.*
- *Explicitly explain why the selected waypoint has the highest probability.*
- *Keep reasoning concise (2-3 sentences max).*
- *Do not suggest waypoints not shown in the images.*

Model Output: JSON List
[
"Selected waypoint": "F3",
"Reasoning": "F3 is along the main corridor between the kitchen and dining area, making it the closest and most direct path to the dining room entrance. It aligns with the room's primary axis, and the arrangement of furniture and walls provides clear orientation cues. Frontiers F1 and F2 lead to side areas with less clear access to the main corridor."
]

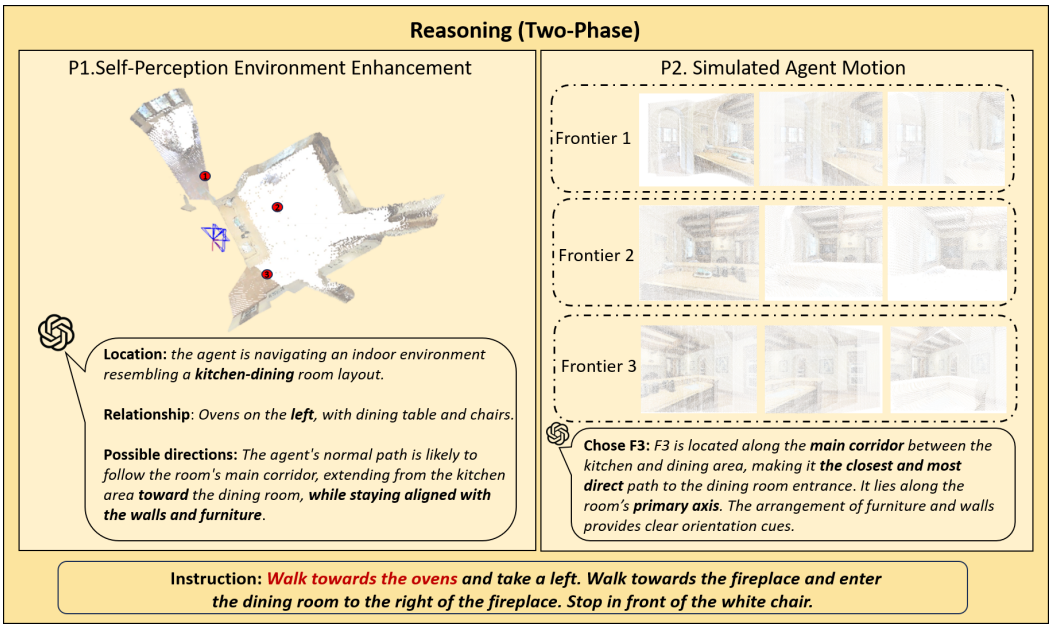

Figure 5: The specific details of the MLLM reasoning are presented in the Case study.

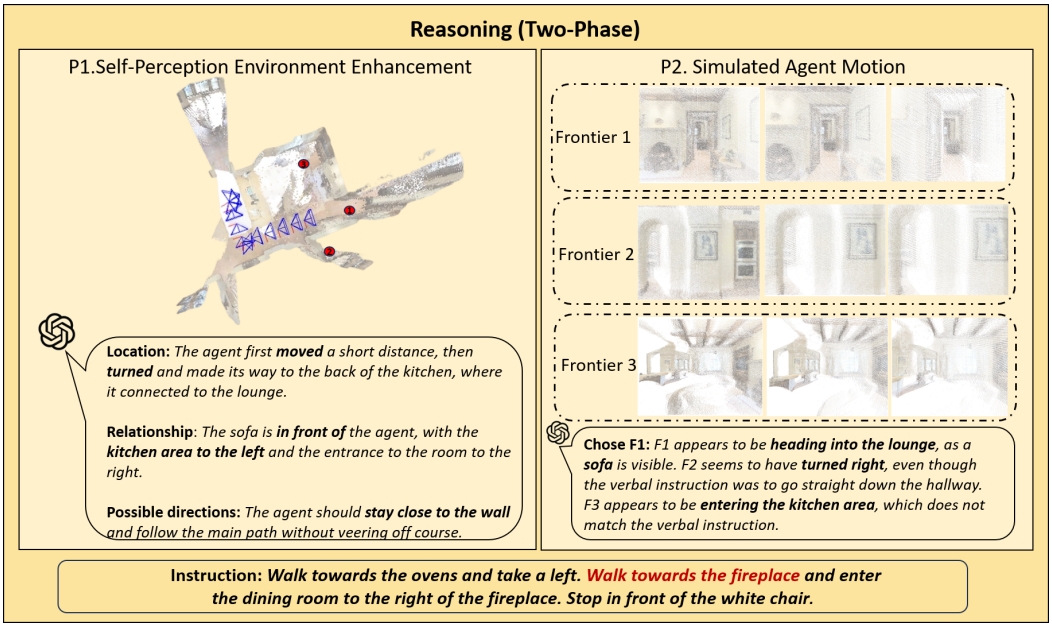

Figure 6: The specific details of the MLLM reasoning are presented in the Case study.

B Experiment Details

B.1 Qualitative Analysis

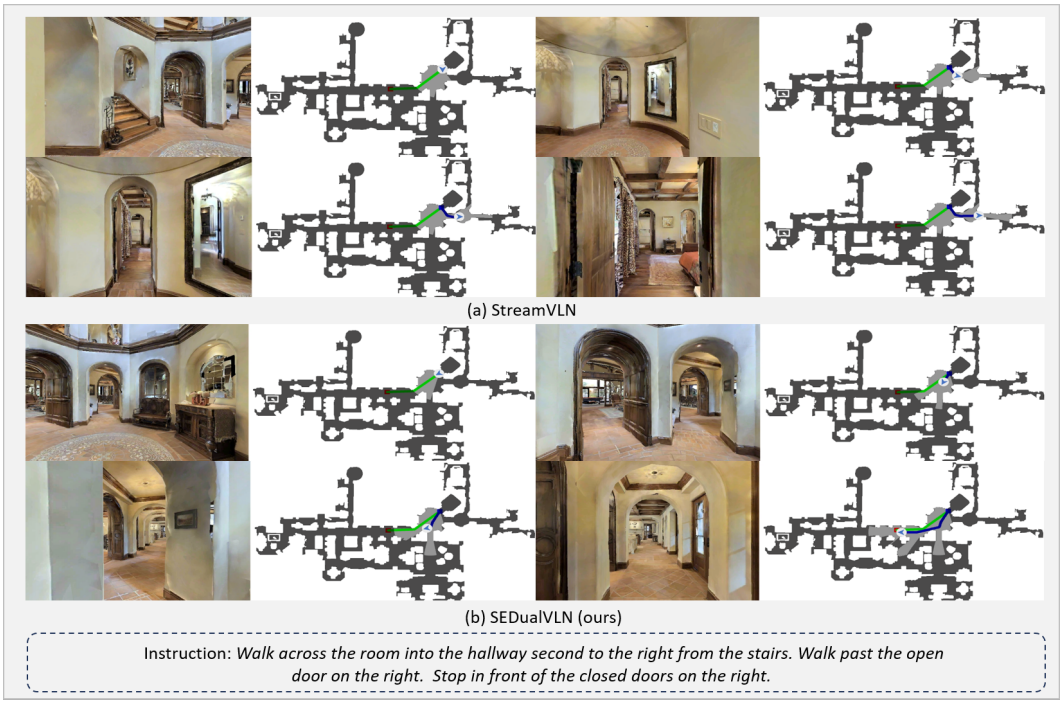

Figure 7: Comparative experiment. SEDualVLN did not take a wrong turn at the beginning like StreamVLN; instead, it selected the correct path at the fork and successfully completed the task.

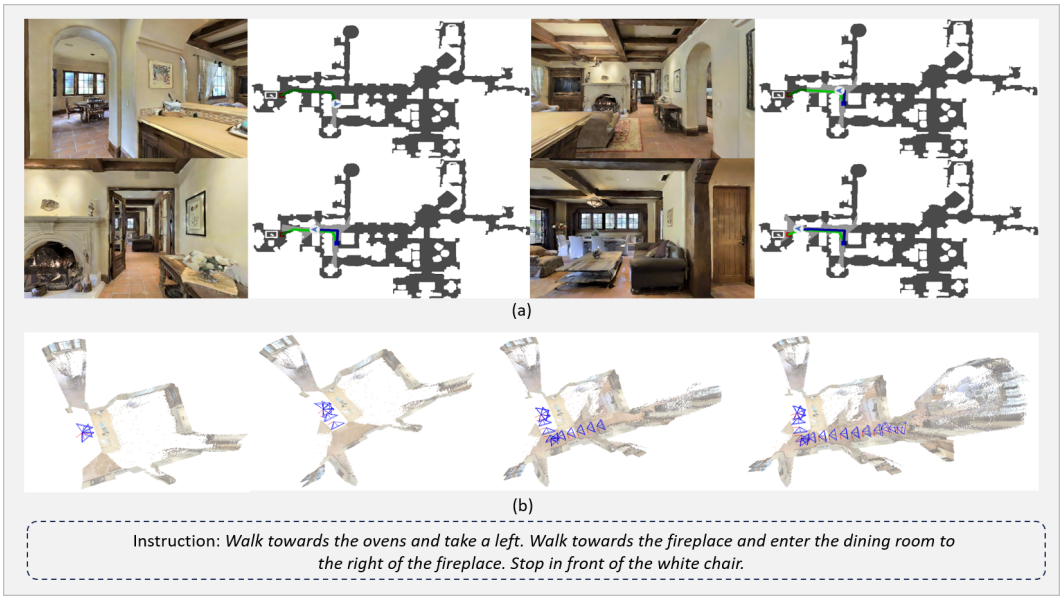

Figure 8: Case study. Navigation and mapping visualization of our SEDualVLN.