NaviVLN: Learning Task-Centric Hierarchical Representations via Embodied Alignment for Dual-System Vision-Language Navigation

Anonymous authors
Paper under double-blind review

Abstract

Vision-Language Navigation (VLN) integrates spatial perception, instruction understanding, and sequential decision-making. Foundation-model-based policies benefit from multimodal supervision, yet insufficient task relevance and cross-modal correspondence can cause task-agnostic intra-modal redundancy and cross-modal semantic misalignment, limiting sample efficiency and decision consistency. We present **NaviVLN**, a dual-system policy built on **THREAD** (Task-centric Hierarchical Representations via Embodied Alignment for Dual-system), aligning vision, language, and trajectories through data, model, and training. THREAD fuses navigation information into mutually informative visual and textual targets grounding instruction semantics and task progress. Joint Attention integrates implicit visual and textual chain-of-thought during training, shaping shared representations through auxiliary and navigation supervision. Staged training connects these representations to a local trajectory expert; deployment discards auxiliary decoders, using asynchronous intention querying for efficient execution. On R2R-CE Val-Unseen, NaviVLN achieves approximately 10% higher SR than Uni-NaVid using roughly 10% of its reported training data. Ablations show absolute SR gains of 8.49% and 4.30% over action-only and unaligned multimodal baselines, respectively, supporting the effectiveness of full-pipeline task alignment. Further gains with increasing training data suggest strong scaling potential.

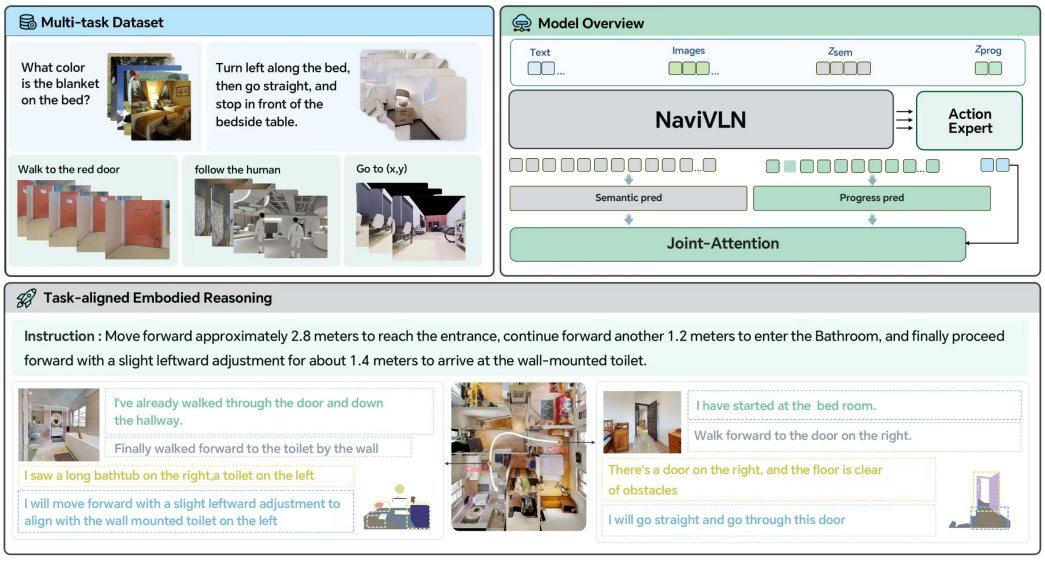


Figure 1: Overview of NaviVLN and THREAD. Top left: multi-task data examples. Top right: semantic and progress predictors in the model architecture, coupled through Joint Attention, shape shared representations for trajectory generation. Bottom: task-aligned embodied reasoning relates instruction context, scene evidence, task progress, and intended motion.

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.

1 Introduction

Vision-Language Navigation (VLN) requires embodied agents to integrate spatial perception, instruction understanding, and sequential decision-making (Anderson et al., 2017; Krantz et al., 2020). Advances in vision-language and video prediction foundation models have encouraged end-to-end policies with broader perceptual and reasoning capabilities, motivating exploration of data formulations and model architectures (Zhang et al., 2024b;a; Wei et al., 2025a).

At the data level, diverse sources and fine-grained annotations improve performance and demonstrate scaling potential (Zhang et al., 2024a; DeepCybo Team, 2026). Although these enrich general-purpose knowledge, their relevance to navigation and correspondence across modalities remains underexplored. Detailed reconstruction and textual analysis (Zhao et al., 2025; Lu et al., 2026) can introduce task-agnostic intra-modal redundancy and underemphasize cross-modal relationships, potentially diverting representational capacity and reducing sample efficiency.

At the model level, many approaches combine specialized models or multiple decoders, encouraging alignment through paired data and joint objectives (Lu et al., 2026; Zuo et al., 2026). However, those prediction branches often operate independently in parallel (Lu et al., 2026), lacking explicit pathways for information exchange during prediction. Combined with navigation ambiguity and differing modal emphases, this separation can cause cross-modal semantic misalignment: individually plausible but inconsistent predictions that introduce ambiguity and conflicting guidance into trajectory generation.

Together, these observations suggest that individual branch enhancement and multi-branch stacking can leave capabilities as loosely connected strands. Drawing inspiration from human reasoning, which relates, cross-validates, and integrates complementary evidence, we investigate how task-relevant information and cross-modal correspondence can be learned together, weaving diverse capabilities into coherent navigation representations.

We introduce **NaviVLN**, a dual-system navigation policy built on **THREAD** (Task-centric Hierarchical Representations via Embodied Alignment for Dual-system). THREAD aligns vision, language, and trajectories throughout data construction, model architecture, and staged training.

At the data level, THREAD fuses navigation information into mutually informative targets. Visual targets combine instruction-relevant semantics with projected trajectory regions; textual targets incorporate movement history, subtask progress, and landmarks. This embeds complementary modal information directly into task-aligned supervision without introducing additional information sources.

At the model level, **Joint Attention** integrates implicit visual and textual chain-of-thought (CoT) streams during training, incorporating complementary evidence as predictions form. Each stream feeds its corresponding auxiliary decoder, while joint visual, textual, and navigation supervision shapes shared representations. Staged training connects these representations to a local trajectory expert. Deployment removes auxiliary decoders and Joint Attention, retaining asynchronous intention querying for efficient trajectory generation. Our contributions are threefold:

- **Task-aligned data grounding.** We reformulate navigation data into complementary targets grounding instruction semantics and task progress in scenes and expert trajectories.
- **Cross-modal latent alignment.** Joint Attention couples implicit CoT streams, enabling modality-specific predictions to incorporate complementary task information.
- **Task-centric hierarchical training.** Staged training links VLM reasoning and action through asynchronous latent queries, without online auxiliary decoding.

NaviVLN achieves approximately 10% higher SR than Uni-NaVid on R2R-CE Val-Unseen using roughly 10% of its reported training data. Absolute SR gains of 8.49% and 4.30% over action-only and unaligned multimodal baselines support full-pipeline alignment, while gains across increasing data scales suggest further scaling potential.

2 Related Work

2.1 Vision-Language Navigation

VLN research has progressed from discrete instruction-following benchmarks to continuous environments through VLN-CE (Anderson et al., 2017; Krantz et al., 2020). Early policies explored multimodal pretraining, history modeling, and spatial representations (Hao et al., 2020; Chen et al., 2021; 2022).

Generalization challenges have motivated foundation-model-based navigation, with increasing emphasis on scalable data construction and transferable perceptual and reasoning capabilities (Zhang et al., 2024b;a; Wei et al., 2025a).

2.2 Data Scaling and Auxiliary Supervision

Navigation data construction follows two complementary directions: expanding coverage and enriching supervision. Some approaches collect diverse scenes and trajectories for direct policy learning (Zhang et al., 2024b;a); others organize visual-grounding and trajectory data for staged learning (Wei et al., 2025a). Beyond demonstrations, navigation and adjacent embodied research incorporate image/video question answering, future visual prediction, and CoT annotations (Zhang et al., 2024b; Zhao et al., 2025; Zuo et al., 2026). Fine-grained spatial, temporal, and affordance grounding further broadens supervision (DeepCybo Team, 2026). These advances enrich training signals, while leaving open which information is necessary for navigation and how complementary targets should correspond. We focus on this relationship by constructing mutually informative supervision from existing observations, instructions, and trajectories.

2.3 Multimodal Navigation Architectures

Foundation-model-based navigation spans prompted agents, autoregressive action policies, and hierarchical planning–execution systems (Zhou et al., 2023; Zhang et al., 2024b; Wei et al., 2025a). Auxiliary learning further extends these designs through expert fusion, modality-specific decoders, and consistency objectives. Related approaches explore expert fusion (Bi et al., 2025), auxiliary decoding (Lu et al., 2026), and cross-modal consistency (Zuo et al., 2026). Such coordination may still leave independently formed predictions semantically inconsistent. In our work, we jointly align supervision and latent reasoning through mutually informative targets and Joint Attention across implicit visual and textual CoT streams.

3 Method

Given an instruction I and the current and historical RGB observations $O_{\leq t}$, NaviVLN predicts a continuous local trajectory. THREAD develops task-centric navigation representations through three complementary dimensions: task-aligned data grounding constructs mutually informative visual and textual targets; cross-modal latent alignment integrates implicit CoT streams through Joint Attention; and task-centric hierarchical training connects the learned representations to a local trajectory expert.

3.1 Task-Aligned Data Grounding

Navi-Dataset associates navigation instructions, egocentric observations, and expert trajectories. For instruction-following trajectories, we divide each instruction into ordered subtasks and associate sampled observations with the current subtask. This provides a shared task context for progress-aware textual annotations and task-aligned semantic visual targets (Figure 2).

An offline annotator generates a progress-aware description R_t from the instruction, current subtask, and historical and current observations. It summarizes completed subtasks, identifies the subtask to execute in the current view, describes visible scene evidence, and indicates the next movement direction without specifying action coordinates. Tokenizing R_t produces the textual target Y_t^{text}, connecting movement history, task progress, and current observations to subsequent navigation behavior.

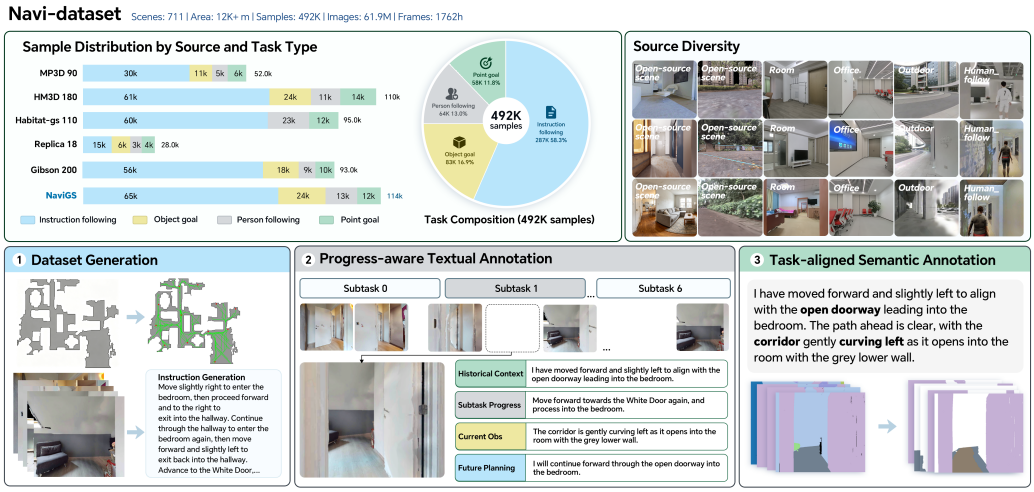


Figure 2: Navi-Dataset composition and task-aligned annotation pipeline. Top: The dataset contains 492K samples across multiple navigation tasks, combining diverse scene sources and environments. Bottom: Navigation trajectories and instructions are generated, followed by progress-aware textual annotation covering movement history, subtask progress, current observations, and future motion. These descriptions guide task-relevant semantic selection, while projected expert trajectories provide spatial grounding in visual targets. The resulting visual and textual supervision encodes complementary information from the same navigation context.

The semantic annotation generator then combines R_t with the current-frame semantic image, the remaining expert trajectory, and camera parameters. It first projects the remaining expert trajectory into the semantic image and selects its traversable path region, then uses R_t to identify task-relevant landmarks. The resulting task-aligned visual target Y_t^{vis} combines instruction-relevant semantics with the projected expert-path region, reducing texture redundancy while retaining navigation-relevant scene and motion information. Together, these mutually informative targets guide the VLM's representation learning from RGB observations, establishing cross-modal correspondence among scene semantics, task progress, and expert trajectories.

3.2 Cross-Modal Latent Alignment

Building on aligned supervision, we align the prediction branches within the model. A shared VLM encodes the instruction and RGB context. Four semantic tokens Z_{sem} and two progress tokens Z_{prog} produce hidden states H_{sem} and H_{prog}, respectively, which the visual and textual predictors use to reconstruct Y_t^{vis} and Y_t^{text}. The predictors retain their distinct targets while exchanging task information at their final layer (Figure 3).

Let F_{sem} and F_{prog} denote the final-layer features of the semantic and progress predictors, and let $\text{KVProj}_{\text{sem}}$ and $\text{KVProj}_{\text{prog}}$ denote their branch-specific key/value projections. In Joint Attention, each branch uses its own latent as the query Q, while its key/value states comprise its own latent, the other branch's stop-gradient latent, and the stop-gradient main-navigation state H_{traj}. The within-branch term preserves modality-specific information; the cross-branch term supplies the other modality's current task cues. Stop-gradient permits forward information exchange while blocking cross-branch gradients along the attention path, limiting unwanted coupling and representation collapse. Both branches also use H_{traj} as the shared navigation reference and supervisory anchor. In Figure 3, the two predictor-to-Joint-Attention arrows correspond to F_{sem} and F_{prog}; the displayed $\mathcal{L}_{\text{sem}}$ and $\mathcal{L}_{\text{prog}}$ denote auxiliary losses. The key/value projections are internal to Joint Attention and are not drawn separately. Formally,

$$Q_{\text{sem}} = F_{\text{sem}} W_Q^{\text{sem}}, \quad (K_{\text{sem}}, V_{\text{sem}}) = \text{KVProj}_{\text{sem}}([F_{\text{sem}}; \text{sg}(F_{\text{prog}}); \text{sg}(H_{\text{traj}})]),$$
$$Q_{\text{prog}} = F_{\text{prog}} W_Q^{\text{prog}}, \quad (K_{\text{prog}}, V_{\text{prog}}) = \text{KVProj}_{\text{prog}}([F_{\text{prog}}; \text{sg}(F_{\text{sem}}); \text{sg}(H_{\text{traj}})]).$$

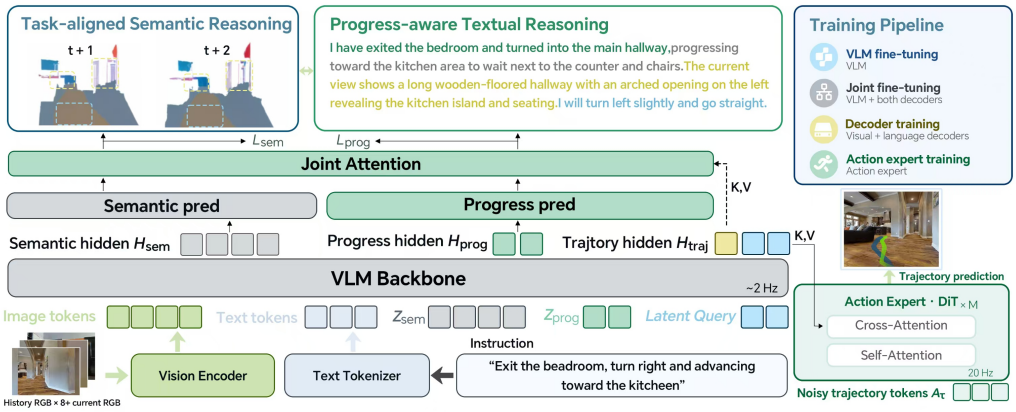


Figure 3: NaviVLN architecture and training framework. The VLM processes observations and instructions to form semantic, progress, and trajectory-related hidden states. During training, visual and textual auxiliary predictors are coupled through Joint Attention, with trajectory-related hidden states providing a shared reference. Joint supervision shapes task-centric representations, while staged training learns a DiT action expert conditioned on latent intention queries. At deployment, auxiliary decoders are removed; asynchronous querying connects slower VLM reasoning with faster local trajectory generation.

Representation learning jointly optimizes the navigation objective and the two auxiliary objectives:

$$\mathcal{L}_{\text{repr}} = \mathcal{L}_{\text{nav}} + \lambda_{\text{prog}}\mathcal{L}_{\text{prog}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}}.$$

Reconstruction losses shape shared VLM representations with progress-aware text and task-aligned visual semantics, while Joint Attention integrates complementary branch information around a common navigation reference.

3.3 Task-Centric Hierarchical Training

Training proceeds in four stages. **Stage 0** adapts the VLM using navigation-language supervision and multitask data (Figure 1). **Stage 1** freezes the VLM and trains auxiliary predictors, projections, and Joint Attention, using H_{traj} to anchor both branches to the navigation task. **Stage 2** jointly optimizes the VLM and auxiliary modules with $\mathcal{L}_{\text{repr}}$, consolidating task-aligned semantic and progress representations. **Stage 3** freezes the VLM and existing task embeddings, training latent queries Z_q and a DiT action expert through flow matching.

The queries extract H_c from this task-aligned latent space, providing semantic and progress context to the expert through cross-attention. Combined with reference and current RGB features, H_c guides prediction of arc-length-sampled local position and orientation increments. At deployment, slower VLM queries and faster closed-loop trajectory updates operate asynchronously. Action-expert training and inference use H_c directly, without decoding navigation answers or auxiliary outputs; only the VLM, latent queries, and action expert are needed for deployment.

4 Experiments

We evaluate NaviVLN through unseen-scene navigation, ablations of the alignment architecture, and deployment on a physical robot. Representation scores and key-frame visualizations further examine how task alignment relates to navigation decisions.

4.1 Unseen-Scene Navigation and Training Scale

We evaluate continuous instruction following on the Val-Unseen splits of R2R-CE and RxR-CE (Krantz et al., 2020; Ku et al., 2020). NaviVLN uses current and historical monocular RGB observations. Success rate (SR) measures whether the agent stops near the goal, while success weighted

by path length (SPL) also accounts for path efficiency. Table 1 compares methods across input settings and training-data scales. Even with a subset of the full training dataset, NaviVLN achieves substantial gains, demonstrating strong sample efficiency and suggesting potential for further scaling.

Table 1: R2R-CE and RxR-CE Val-Unseen navigation results (%) and training-data scale. Pano., Odo., and S.RGB denote panoramic views, odometry, and single-view RGB, respectively; NavFoM uses four RGB views. Methods are grouped by comparison focus: early baselines, navid-family methods, and large-data methods.

						R2R-CE Val-Unseen				RxR-CE	
Method	Data	Pano.	Odo.	Depth	S.RGB	NE ↓	OSR ↑	SR ↑	SPL ↑	SR ↑	SPL ↑
ETPNav(An et al., 2023)	—	✓	✓	✓		4.71	65.0	57.0	49.0	54.8	44.9
ScaleVLN(Wang et al., 2023)	4.9M	✓	✓	✓		4.80	—	55.0	51.0	—	—
NaVid(Zhang et al., 2024b)	1.28M				✓	5.47	49.1	37.4	35.9	—	—
Uni-NaVid(Zhang et al., 2024a)	3.58M				✓	5.58	53.3	47.0	42.7	48.7	40.9
NaVILA(Cheng et al., 2024)	~13.1M				✓	5.22	62.5	54.0	49.0	49.3	44.0
NaviVLN	**330K**				✓	**4.63**	**60.99**	**51.39**	**48.03**	**51.23**	**47.91**
NavFoM(Zhang et al., 2025)	12.7M		✓			4.61	72.1	61.7	55.3	64.4	56.2
ABot-N0(Chu et al., 2026)	16.9M	✓				3.78	70.8	66.4	63.9	69.3	60.0
StreamVLN(Wei et al., 2025b)	~26.3M				✓	4.98	64.2	56.9	51.9	52.9	46.0
InternVLA-N1(Intern Robotics, 2025)	>5M			✓	✓	4.83	63.3	58.2	54.0	53.5	46.1
DualVLN(Wei et al., 2025a)	~12M				✓	4.05	70.7	64.3	58.5	61.4	51.8
Qwen-RobotNav-4B(Qwen Team, 2026)	15.6M				✓	4.22	73.6	66.9	60.5	71.3	61.5

Training-data amounts in Table 1 are reported or estimated from the corresponding papers; dataset composition and exact scale are given in those works. NaviVLN obtains 51.39% SR and 48.03% SPL on R2R-CE, exceeding NaVid and Uni-NaVid in SR by 13.99 and 4.39 percentage points. Compared with Uni-NaVid, NaviVLN achieves approximately 10% higher SR with 330K decision anchors, roughly 10% of Uni-NaVid's reported 3.58M training samples. On RxR-CE, it reaches 51.23% SR and 47.91% SPL, exceeding Uni-NaVid by 2.53 and 7.01 percentage points in SR and SPL, respectively, and NaVILA by 1.93 percentage points in SR. These results are achieved using only a subset of the full dataset; the sample count used for training is approximately 1/47 of Qwen-RobotNav's reported 15.6M mixed samples (Qwen Team, 2026), highlighting competitive performance at the current training scale.

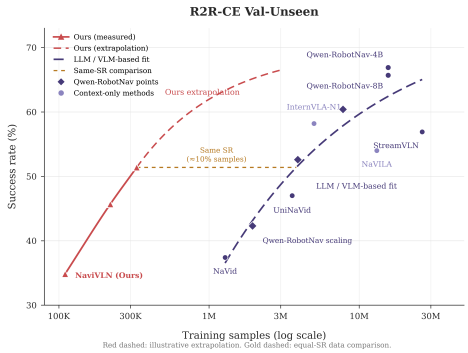


Figure 4: Measured R2R-CE Val-Unseen SR versus training-data scale. The red dashed extension is illustrative, not an additional measurement.

Increasing NaviVLN's training samples from 110K to 220K and 330K raises R2R-CE SR from 34.84% to 45.68% and 51.39%, demonstrating continued gains from additional task-aligned data. As illustrated in Figure 4, these measured improvements

indicate strong sample efficiency and suggest substantial potential for further scaling as more of the full dataset is used.

4.2 Auxiliary Supervision and Architectural Task Alignment

To examine how auxiliary branches align around navigation, we compare indirect coupling (Separate), implicit fusion (Merged), and explicit alignment (Joint), while keeping the visual and textual auxiliary targets fixed. Figure 5(a–c) depicts these predictor designs, while panel (d) presents the representation-alignment comparison.

In **Separate**, independent visual and textual predictors are connected through the shared backbone and auxiliary reconstruction gradients, without explicit fusion during prediction. **Merged** assigns both targets to one vision-language predictor, implicitly integrating modalities through shared context and causal attention. **Joint** retains modality-specific predictors and fuses their latent representations through Joint Attention, using the navigation hidden state H_{traj} as a common anchor. **Joint+loss** adds a cross-branch alignment loss to assess the benefit of an explicit constraint. The first three variants share auxiliary targets and loss settings to isolate predictor organization and fusion; Joint+loss separately evaluates the additional alignment objective.

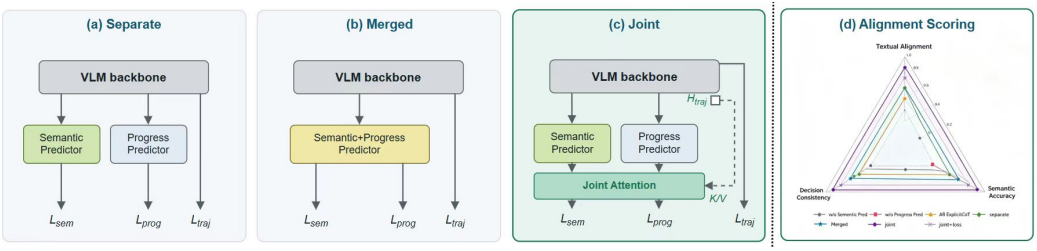


Figure 5: Predictor architectures and representation alignment. Panels (a–c) show Separate, Merged, and Joint predictor designs; panel (d) compares textual alignment, semantic accuracy, and decision consistency. Joint Attention uses H_{traj} as a shared navigation reference.

Model architecture. Table 2 first isolates the two auxiliary targets and then compares their architectural organization and additional alignment constraint.

Table 2: R2R-CE Val-Unseen navigation ablation. NE is in meters; all other metrics are percentages.

Configuration	Text	Visual	NE ↓	OSR ↑	SR ↑	SPL ↑
AR answers only	—	—	5.1401	52.39	42.90	40.06
+ Traversability predictor	—	✓	5.1080	53.25	43.83	41.33
+ Progress predictor	✓	—	4.8620	55.16	45.89	42.97
+ Online text reasoning	✓	—	4.6701	59.99	50.86	47.56
Separate	✓	✓	4.7752	55.53	47.09	44.52
Merged	✓	✓	4.8531	54.23	45.40	43.21
Joint+loss	✓	✓	4.7230	58.99	49.39	47.23
Joint	✓	✓	**4.6323**	**60.99**	**51.39**	**48.03**

We first assess individual auxiliary targets. Adding task-aligned semantic or progress-aware textual prediction raises SR from the autoregressive navigation-answer baseline of 42.90% to 43.83% and 45.89%, respectively, confirming their individual benefits. Online textual reasoning further improves performance to 50.86% SR and 47.56% SPL but requires explicit text generation at inference. The following implicit reasoning variants avoid online auxiliary decoding, with performance depending on predictor organization and latent fusion.

Combining both targets in Separate achieves 47.09% SR, demonstrating their complementarity. Merged drops to 45.40%: its causal ordering introduces asymmetric dependencies that may propagate modality-specific biases and intensify competition between generation objectives. Joint achieves the highest SR and SPL of 51.39% and 48.03%, exceeding Separate by 4.30 and 3.51

percentage points and slightly outperforming online textual reasoning. These results support retaining modality-specific predictors while integrating complementary information through Joint Attention around a shared navigation anchor, balancing task specialization and cross-modal consistency. Joint+loss provides no further gains, suggesting limited benefit from additional alignment constraints in this setting.

Representation alignment scores. To assess the task-relevant alignment between predicted content and decision representations, we use three complementary auxiliary scores (Figure 5(d)). Textual Alignment evaluates the progress branch by combining the correctness of its subtask-progress judgment with the accuracy of its generated explanation relative to a reference description. Semantic Accuracy evaluates the visual branch along four dimensions: trajectory geometry, coverage of the path region in image space, stopping timing, and deviation from the intended path. It measures whether scene grounding follows traversable structure. Decision Consistency takes the geometric mean of trajectory, spatial, and textual subscores, reflecting task-level agreement between the two auxiliary supervision streams and the final decision representation. These three scores assess progress, scene grounding, and overall decision alignment, respectively, complementing SR and SPL, which evaluate navigation outcomes.

Figure 5(d) offers complementary evidence: Joint has the strongest textual alignment, semantic accuracy, and decision consistency, whereas the single-predictor and Merged variants score lower and Joint+loss does not improve on Joint. Together with Table 2, these results show that the benefit of auxiliary supervision depends on both its content and its integration around navigation. Joint improves auxiliary prediction quality and decision consistency alongside SR and SPL, supporting architectural task alignment.

Key-frame and attention-map analysis. We compare Separate, Merged, and Joint using the same simulated instruction and observation. Figure 6 presents local trajectory predictions, two-step semantic predictions, progress descriptions, and attention maps from task-aligned semantic reasoning (TSR) and progress-aware textual reasoning (PTR). This comparison illustrates how task-progress judgments and branch attention correspond to predicted navigation behavior.

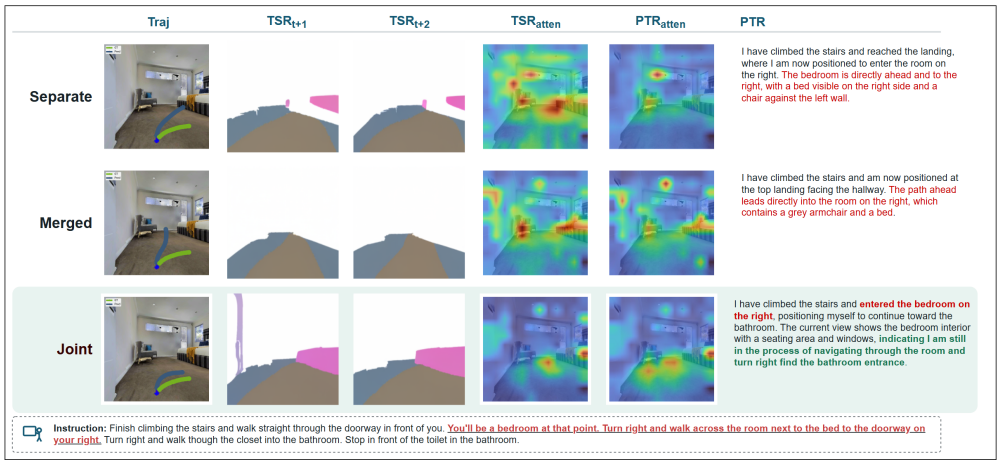


Figure 6: Bedroom-to-bathroom navigation. Separate, Merged, and Joint are compared through local trajectories, two-step semantic predictions, progress descriptions, and attention maps for task-aligned semantic reasoning (TSR) and progress-aware textual reasoning (PTR).

For the same bedroom observation, Joint's progress description recognizes that the agent has entered the bedroom and identifies the bathroom as the next target. Separate and Merged still describe the agent as being on the stair landing and preparing to enter the bedroom, indicating a different judgment of task stage; their local paths and semantic predictions also differ. Joint's projected Semantic and Progress attention is more concentrated on the traversable region ahead, whereas the other configurations show more dispersed or less mutually consistent highlights. This case connects task-stage judgment, attended regions, and local path choice. More cases are shown in the Appendix.

4.3 Multi-Scene Real-World Validation

We deploy NaviVLN on a humanoid robot without additional real-world fine-tuning. A VLM running on a four-card A5880 server updates the latent query state H_c from the instruction and RGB history at approximately 2 Hz. The DiT action expert runs on the robot's Orin at approximately 20 Hz, updating local trajectories from the latest state and RGB observation, conditioned on the most recent H_c from the VLM. No auxiliary text is decoded during real-world inference.

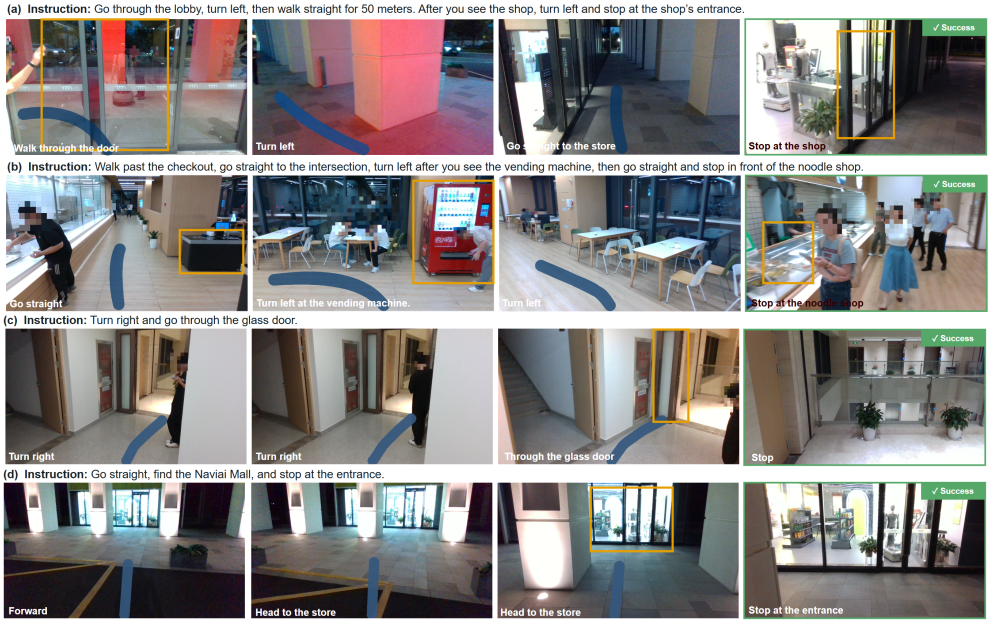


Figure 7: Multi-scene real-world instruction following. Examples demonstrate multi-stage navigation, landmark grounding, and goal localization. Key frames show predicted local trajectories and successful task completion.

As shown in Figure 7, all four real-world cases reach their destinations, covering multi-stage instructions, landmark-guided turns, door traversal, and entrance localization. In the second case, the robot turns left after passing a vending machine and reaches the noodle shop, illustrating task understanding beyond the fixed semantic labels. Although the visual head uses closed-set categories, it provides spatial grounding through task-relevant semantic entities and trajectory regions, while the textual head contributes open-vocabulary scene description and progress understanding. Joint Attention integrates these complementary cues with backbone representations during training, encoding their combined capabilities in task-aligned latents that support navigation without auxiliary decoding at inference.

5 Conclusion

We presented NaviVLN and THREAD, aligning vision, language, and trajectories through task-aligned data grounding, Joint Attention over latent reasoning streams, and staged training. The learned representations guide trajectory generation through latent intention queries. Experiments demonstrate improved navigation performance and sample efficiency, with successful real-world execution. Future work will scale training data and extend the framework to more diverse environments and manipulation tasks.

AI USE STATEMENT

Generative AI tools assisted with Chinese-to-English translation, drafting and editing manuscript text, reviewing notation and reference formatting, preparing figure captions and layout, and discussing the presentation and interpretation of experimental comparisons. The offline annotation pipeline also generated progress-aware textual training descriptions, as detailed in Section 3.1. The authors reviewed and revised AI-assisted material and take responsibility for the research design, experiments, reported results, citations, and final manuscript.

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. https://arxiv.org/abs/2304.03047, 2023.

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. https://arxiv.org/abs/1711.07280, 2017.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, Hongyan Zhao, Hanyu Liu, Zhizhong Su, Lei Ma, Hang Su, and Jun Zhu. Motus: A Unified Latent Action World Model. https://arxiv.org/abs/2512.13030, 2025.

Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History Aware Multimodal Transformer for Vision-and-Language Navigation. https://arxiv.org/abs/2110.13309, 2021.

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. https://arxiv.org/abs/2202.11742, 2022.

An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Zaitian Gongye, Xueyan Zou, Jan Kautz, Erdem Biyik, Hongxu Yin, Sifei Liu, and Xiaolong Wang. NaVILA: Legged Robot Vision-Language-Action Model for Navigation. https://arxiv.org/abs/2412.04453, 2024.

Zedong Chu, Shichao Xie, Xiaolong Wu, Yanfen Shen, Minghua Luo, Zhengbo Wang, Fei Liu, Xiaoxu Leng, Junjun Hu, Mingyang Yin, Jia Lu, Yingnan Guo, Kai Yang, Jiawei Han, Xu Chen, Yanqing Zhu, Yuxiang Zhao, Xin Liu, Yirong Yang, Ye He, Jiahang Wang, Yang Cai, Tianlin Zhang, Li Gao, Liu Liu, Mingchao Sun, Fan Jiang, Chiyu Wang, Zhicheng Liu, Honglin Han, Zhining Gu, Kuan Yang, Jianfang Zhang, Di Jing, Zihao Guan, Wei Guo, Qingguo Liu, Di Yang, Xiangpo Yang, Menglin Yang, Hongguang Xing, Weiguo Li, and Mu Xu. ABot-N0: Technical Report on the VLA Foundation Model for Versatile Embodied Navigation. https://arxiv.org/abs/2602.11598, 2026.

DeepCybo Team. PhysBrain 1.5: From Vision-Language Models to Physical Foundation Models. https://arxiv.org/abs/2609.14973, 2026.

Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. Towards Learning a Generic Agent for Vision-and-Language Navigation via Pre-training. https://arxiv.org/abs/2002.10638, 2020.

Intern Robotics. InternVLA-N1: An Open Dual-System Vision-Language Navigation Foundation Model with Learned Latent Plans. https://internrobotics.github.io/internvla-n1.github.io/static/pdfs/InternVLA_N1.pdf, 2025.

Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the Nav-Graph: Vision-and-Language Navigation in Continuous Environments. https://arxiv.org/abs/2004.02857, 2020.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-Across-Room: Multilingual Vision-and-Language Navigation with Dense Spatiotemporal Grounding. https://arxiv.org/abs/2010.07954, 2020.

Jinghui Lu, Jiayi Guan, Zhijian Huang, Jinlong Li, Guang Li, Lingdong Kong, Yingyan Li, Han Wang, Shaoqing Xu, Yuechen Luo, Fang Li, Chenxu Dang, Junli Wang, Tao Xu, Jing Wu, Jianhua Wu, Xiaoshuai Hao, Wen Zhang, Tianyi Jiang, Lingfeng Zhang, Lei Zhou, Yingbo Tang, Jie Wang, Yinfeng Gao, Xizhou Bu, Haochen Tian, Yihang Qiu, Feiyang Jia, Lin Liu, Yigu Ge, Hanbing Li, Yuannan Shen, Jianwei Cui, Hongwei Xie, Bing Wang, Haiyang Sun, Jingwei Zhao, Jiahui Huang, Pei Liu, Zeyu Zhu, Yuncheng Jiang, Zibin Guo, Chuhong Gong, Hanchao Leng, Kun Ma, Naiyan Wang, Guang Chen, Kuiyuan Yang, Hangjun Ye, and Long Chen. Xiaomi OneVL: One-Step

Latent Reasoning and Planning with Vision-Language Explanation. https://arxiv.org/abs/2604.18486, 2026.

Qwen Team. Qwen-RobotNav Technical Report: A Scalable Navigation Model Designed for an Agentic Navigation System. https://arxiv.org/abs/2606.18112, 2026.

Zun Wang, Jialu Li, Yicong Hong, Yi Wang, Qi Wu, Mohit Bansal, Stephen Gould, Hao Tan, and Yu Qiao. Scaling Data Generation in Vision-and-Language Navigation. https://arxiv.org/abs/2307.15644, 2023.

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. https://arxiv.org/abs/2512.08186, 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. https://arxiv.org/abs/2507.05240, 2025b.

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. https://arxiv.org/abs/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. https://arxiv.org/abs/2402.15852, 2024b.

Jiazhao Zhang, Anqi Li, Yunpeng Qi, Minghan Li, Jiahang Liu, Shaoan Wang, Haoran Liu, Gengze Zhou, Yuze Wu, Xingxing Li, Yuxin Fan, Wenjun Li, Zhibo Chen, Fei Gao, Qi Wu, Zhizheng Zhang, and He Wang. Embodied Navigation Foundation Model. https://arxiv.org/abs/2509.12129, 2025.

Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, Ankur Handa, Ming-Yu Liu, Donglai Xiang, Gordon Wetzstein, and Tsung-Yi Lin. CoT-VLA: Visual Chain-of-Thought Reasoning for Vision-Language-Action Models. https://arxiv.org/abs/2503.22020, 2025.

Gengze Zhou, Yicong Hong, and Qi Wu. NavGPT: Explicit Reasoning in Vision-and-Language Navigation with Large Language Models. https://arxiv.org/abs/2305.16986, 2023.

Jing Zuo, Lingzhou Mu, Fan Jiang, Chengcheng Ma, Mu Xu, and Yonggang Qi. FantasyVLN: Unified Multimodal Chain-of-Thought Reasoning for Vision-Language Navigation. https://arxiv.org/abs/2601.13976, 2026.

A Supplementary Cases

A.1 Turn Left into the Second Room

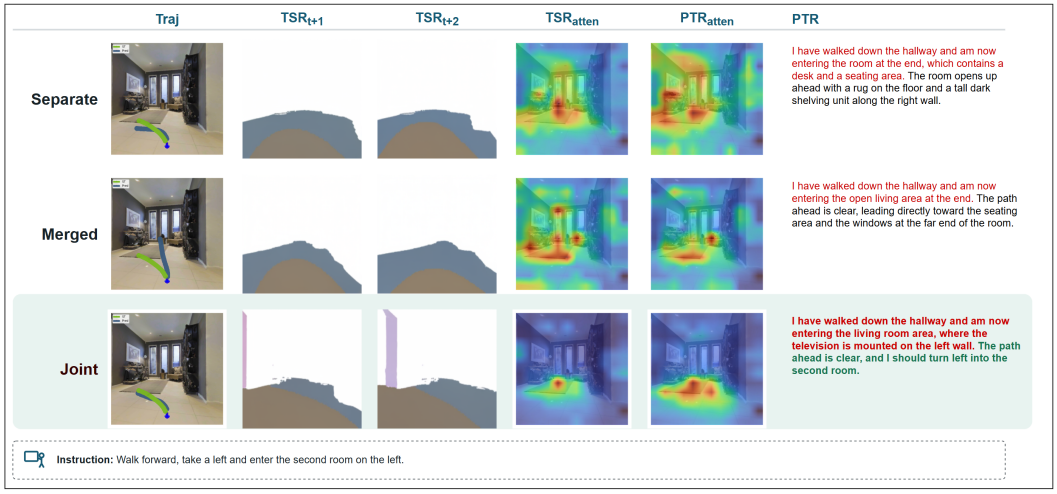


Figure A1: Turn left into the second room. Separate, Merged, and Joint are compared under the same instruction and observation; columns have the same meaning as in Figure 6.

The instruction calls for a left turn into the second room. Joint's progress description identifies the upcoming turn, and its local path enters the room on the left. Separate and Merged mainly describe the area ahead without clearly identifying this next subgoal. Joint's Semantic and Progress attention is concentrated on the forward floor and turning region, while the other configurations show more dispersed or misaligned highlights. As in Figure 6, the task-stage judgment, attended regions, and local path correspond in the Joint configuration.

A.2 Navigation to a Library Entrance

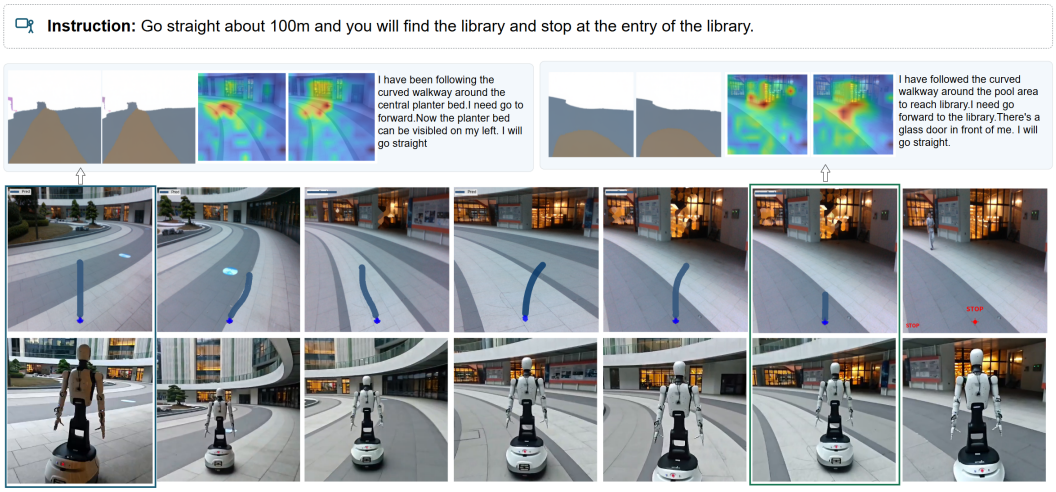


Figure A2: Navigation to a library entrance. Egocentric path predictions and third-person observations show the robot following a curved path toward the library and stopping near its entrance. Two enlarged regions show semantic predictions, attention maps, and progress descriptions at different task stages.

A.3 Navigation from a Bookstore to a Storage Room

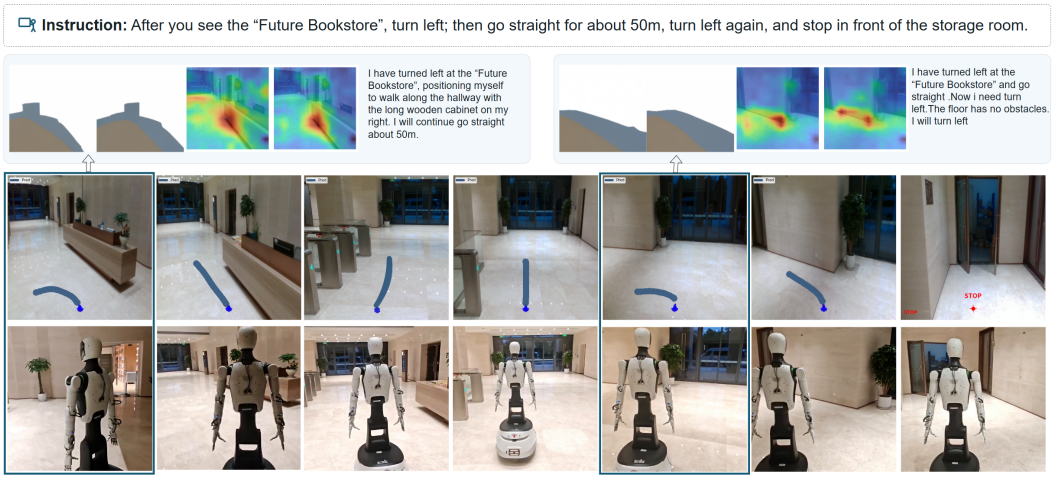


Figure A3: Navigation from a bookstore to a storage room. The robot turns left at the bookstore, proceeds along a corridor, turns left again, and stops near the storage-room entrance. Two enlarged regions show semantic predictions, attention maps, and progress descriptions as the task progresses.

A.4 Additional Simulated Navigation Sequences

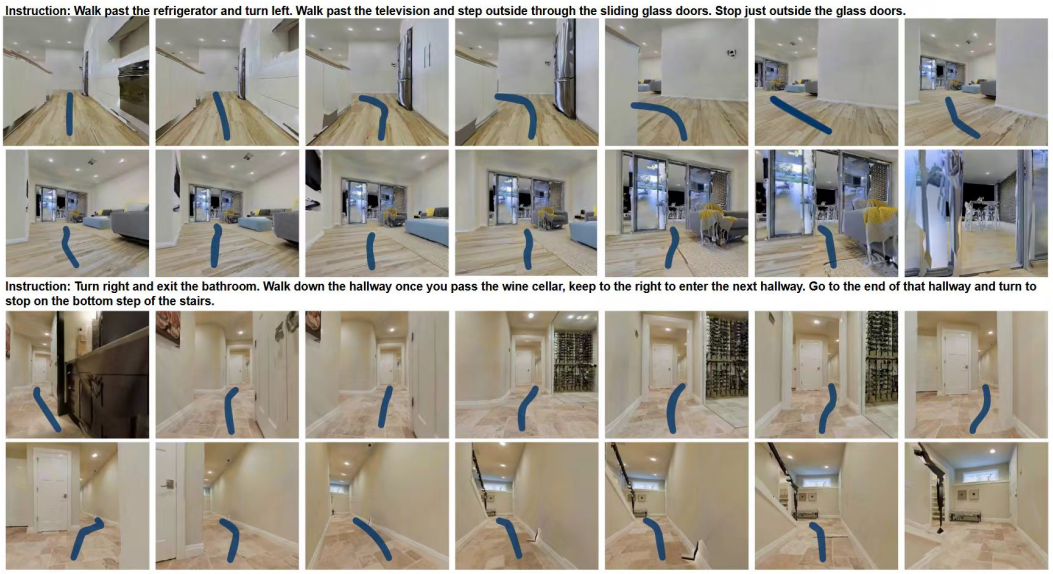


Figure A4: Additional simulated navigation sequences. The upper sequence follows a route past a refrigerator and television before exiting through a sliding glass door. The lower sequence traverses a hallway, turns toward a second corridor, and stops at the stairs. Blue curves show predicted local paths in successive RGB observations.