ERRANDVLN: A BENCHMARK FOR INTENT-DRIVEN LONG-HORIZON CONTINUOUS VISION-AND-LANGUAGE NAVIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing continuous vision-and-language navigation (VLN) benchmarks primarily evaluate single-goal route following, whereas everyday household requests may require agents to visit multiple semantic landmarks in a prescribed order. This mismatch leaves **intent-driven, multi-stage navigation** insufficiently characterized. We introduce **ErrandVLN**, a benchmark that pairs continuous trajectories with diverse role-conditioned requests, allowing the same ordered landmark requirements to be expressed across varied activity contexts. Constructed on InteriorGS environments in NVIDIA Isaac Sim through an automated semantic-geometric pipeline, ErrandVLN comprises 913 scenes, 65,134 unique physical trajectories, and 502,233 instructions, with evaluation splits for unseen scenes, novel trajectory compositions, and unseen instruction styles. To address compounding errors over ordered stages, we propose **VLN-JEPA**: a continuous trajectory proposer (TP-Flow) generates candidate motions, an action-conditioned latent world model (NaviLat) predicts future latent states under these candidate motions without decoding future observations, and a compact 2.35B VLM controller uses these predictions together with the instruction and observation history to generate the next executable action. On ErrandVLN Test Scene-Unseen, VLN-JEPA achieves superior performance across multi-stage milestone completion (20.5% GCR, 16.1% GC-SPL, 29.2% OSR) and competitive task success (12.6% SR), while substantially reducing physical collisions by 40% compared with DualVLN. On a single NVIDIA RTX 4090, it achieves 94 ms per policy-inference step (10.7 FPS) and 5.3 FPS in complete closed-loop execution.

1 INTRODUCTION

Everyday requests in household environments often convey intent-driven activities rather than step-by-step movement instructions. For example, preparing for sleep may involve visiting a chandelier, a bathroom door, and a pillow in sequence, while a cleaning request may involve the same landmarks for a different purpose. These visits are linked by a coherent household errand, with varied requests mapping to shared underlying navigation requirements.

Existing benchmarks study complementary aspects of this setting. R2R (Anderson et al., 2018) and VLN-CE (Krantz et al., 2020) study route following, while REVERIE (Qi et al., 2020) and SOON (Zhu et al., 2021) introduce high-level requests for object localization. GOAT-Bench (Khanna* et al., 2024) evaluates sequential navigation to multimodal goals, and LH-VLN (Song et al., 2025) studies long-horizon object-finding and delivery subtasks. BEHAVIOR-1K (Li et al., 2024) grounds embodied tasks in everyday activities involving manipulation, while SAGE-Bench (Miao et al., 2026) supports physically executable navigation in 3D Gaussian scenes. These advances raise a complementary question: how should navigation be organized when a household activity determines the required visits, and different roles express those requirements through different requests?

We introduce **ErrandVLN**, a benchmark pairing continuous trajectories with role-conditioned household requests. Each trajectory is associated with ordered semantic landmarks and multiple instructions that preserve the required visits while varying their activity context and wording. Built on InteriorGS (Miao et al., 2026) environments in NVIDIA Isaac Sim, ErrandVLN contains 65,134

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.

Table 1: **Comparison of representative embodied navigation benchmarks.** ErrandVLN pairs continuous navigation trajectories with role-conditioned household requests, with each instruction counted as one instance.

Benchmark	Simulator	Scenes	Task Type	Avg. Instr. Length	Avg. Steps	Reported Scale
Single-goal navigation						
R2R	Matterport3D	90	Step-by-step Nav	29	<8	21,567
REVERIE	Matterport3D	90	Object Loco-nav	18	<8	21,702
VLN-CE	Habitat	90	Step-by-step Nav	30	55.88	4,475
SOON	Matterport3D	90	Object Loco-nav	39	10	3,848
SAGE-Bench	Isaac Sim	1,000	Navigation & Exploration	–	–	2,000,000
Multi-stage / iterative navigation						
GOAT-Bench	Habitat	181	Iterative VLN	–	–	725,360
LH-VLN	Habitat	216	Multi-stage VLN	18.17	150.95	3,260
Household / complex tasks						
ErrandVLN (Ours)	**Isaac Sim**	**913**	**Household Errands**	**36.15**	**137.1**	**502,233**

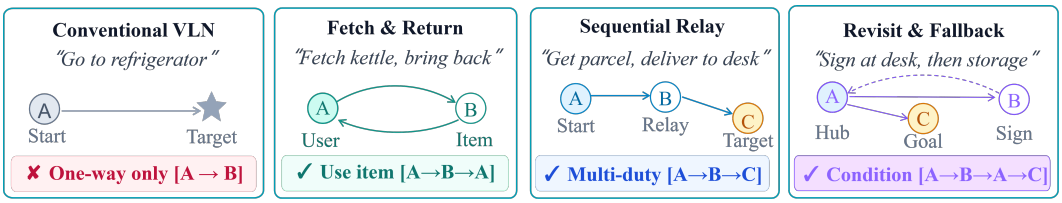

Figure 1: **From route descriptions to household errand requests.** ErrandVLN connects ordered navigation goals through coherent household activities, with role-conditioned requests expressing the corresponding navigation requirements.

unique physical trajectories and 502,233 instructions across 913 scenes. A semantic-geometric pipeline constructs task skeletons, checks their semantic coherence, and generates corresponding trajectories and instructions. The benchmark provides splits for unseen scenes, new trajectories, and scene-conditioned instruction-style variations, with metrics for ordered goal completion and overall navigation success. Table 1 compares representative benchmarks, and Figure 1 illustrates our task setting.

Multi-stage errands require decisions that remain consistent over a long horizon (Song et al., 2025). Local action errors can compound as earlier choices shape subsequent observations, potentially disrupting progress toward later goals (Ross et al., 2011). Video-based policies such as NaVid generate actions from instructions and visual history without explicitly rolling out alternative motions before execution (Zhang et al., 2024b). Predictive navigation models introduce action-conditioned state forecasts to inform decisions (Yao et al., 2025). For errand execution, we seek to make this future information specific to the available choices: predicting the outcomes of multiple candidate motions provides the controller with evidence about their consequences before it generates an action.

We therefore propose **VLN-JEPA**. A **Continuous Trajectory Proposer (TP-Flow)** generates candidate motions, and a **Navigation Latent World Model (NaviLat)** predicts an action-conditioned sequence of future latent states for each candidate in parallel. A vision-language controller combines these candidate-specific predictions with the instruction and observation history to generate the next action, then replans after receiving a new observation. By incorporating predicted consequences into local decisions, VLN-JEPA mitigates sequential error accumulation in long-horizon navigation through efficient latent state forecasting.

On ErrandVLN Test Scene-Unseen, VLN-JEPA achieves 20.5% GCR and 12.6% SR with 26.6 collisions, substantially outperforming existing continuous navigation VLMs in both multi-stage completion and traversal safety. Our contributions are threefold:

- We formulate continuous errand navigation around coherent household activities whose requirements determine the ordered navigation stages.

- We construct **ErrandVLN**, pairing trajectories with role-conditioned requests and providing diagnostic splits for scene, trajectory, and scene-conditioned instruction-style variations.
- We develop **VLN-JEPA**, combining continuous motion proposals with candidate-specific latent predictions to inform local decisions throughout multi-stage errands.

2 Related Work

2.1 From Route Following to Household Errands

Vision-and-language navigation has evolved from visually grounded route following in discrete environments to continuous control and longer-horizon embodied tasks. R2R (Anderson et al., 2018) and RxR (Ku et al., 2020) study instruction following in indoor environments, while REVERIE (Qi et al., 2020) and SOON (Zhu et al., 2021) extend navigation toward high-level object localization. VLN-CE (Krantz et al., 2020) further requires agents to execute low-level actions in continuous environments. Recent benchmarks have expanded this setting along several complementary dimensions. Iterative VLN (IVLN) (Krantz et al., 2023) evaluates agents that maintain memory across ordered tours of instruction-following episodes in persistent environments. ErrandVLN instead represents the ordered navigation requirements of an errand within a single episode. LH-VLN (Song et al., 2025) studies long-horizon execution through object-finding and delivery subtasks. VLN-Verse (Lin et al., 2025) broadens embodied navigation to multiple task settings in Isaac Sim, while SAGE-Bench (Miao et al., 2026) provides semantically annotated and physically executable 3D Gaussian environments. Together, these efforts extend VLN toward longer, more diverse, and more physically executable navigation tasks.

ErrandVLN focuses on task organization around coherent household errands. The purpose of an errand connects its required visits, while role-conditioned requests express shared navigation requirements through different activities and language. The resulting trajectory structure follows the requirements of the household activity, including which landmarks must be visited, in what order, and whether previously visited regions need to be revisited. These task requirements also place new demands on the navigation policy: local actions must remain consistent with both the current observation and the unfinished stages of the errand.

2.2 From Reactive Policies to Predictive Action Selection

To meet the demands of long-horizon errands, a navigation policy must decide not only where the agent is now, but also how its next action may affect subsequent progress. Navigation policies can therefore be distinguished by whether they select actions from observed history or use predictions about future states. VLN-BERT (Hong et al., 2021) and HAMT (Chen et al., 2021) encode visual history, while DUET (Chen et al., 2022b) and ETPNav (An et al., 2024) maintain spatial representations for navigation and path recovery. More recent policies such as NaVid (Zhang et al., 2024b), NaViLa (Cheng et al., 2024), and MapGPT (Chen et al., 2024) directly generate actions from language instructions and visual history. These methods select the current action from observed information, but do not explicitly predict how alternative motions will affect future states before execution.

This limitation motivates future prediction for navigation. PathDreamer (Koh et al., 2021) and DreamWalker (Wang et al., 2023a) generate future visual observations, while FutureView (Li & Bansal, 2023) predicts the semantics of future views. NavMorph (Yao et al., 2025) models the evolution of compact latent states, VISTAv2 (Huang et al., 2025) generates action-conditioned egocentric future views for value estimation. VLN-JEPA focuses on providing a language-conditioned controller with a sequence of predicted latent states for each of several proposed motions.

2.3 From Pixel Prediction to Latent World Models

The representation of predicted outcomes determines what information is available to a navigation policy. Reconstruction-based world models learn latent dynamics together with an observation decoder (Ha & Schmidhuber, 2018). JEPA-style approaches instead predict target representations without reconstructing pixels: I-JEPA predicts masked image representations (Assran et al., 2023), while V-JEPA 2 extends predictive representation learning to video and action-conditioned world

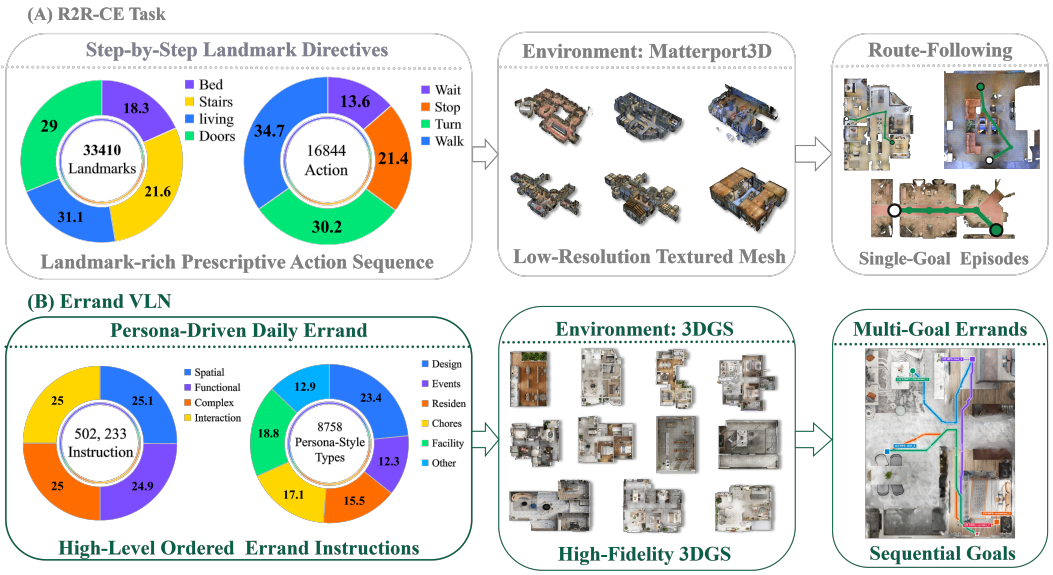

Figure 2: **Comparison between the R2R-CE reference task and ErrandVLN. (Left)** Instruction and persona-conditioned statistics. **(Right)** Representative environments and task structures. The R2R-CE reference task uses single-goal route-following episodes in Matterport3D, whereas ErrandVLN uses ordered multi-goal errands in InteriorGS scenes and supports sequential goals with revisits.

modeling (Assran et al., 2025). LeWorldModel (Maes et al., 2026) combines next-state prediction with Spherical Isotropic Gaussian Regularization to learn latent dynamics without auxiliary decoders.

In navigation, these ideas motivate different ways of encoding future information. PiJEPA (Chahe & Zhou, 2026) predicts latent states along action sequences and evaluates them relative to a goal-image representation. FSC-VLN (Zhang et al., 2026b) aligns future-query representations with future visual embeddings during training, while LatentPilot (Hao et al., 2026) models action-conditioned visual dynamics through latent tokens propagated across time steps. These studies provide foundations for representing future navigation states. For the multiple candidate motions considered in VLN-JEPA, the representation must retain both the correspondence between each motion and its predicted states and the temporal evolution of those states over the prediction horizon.

NaviLat encodes this information as a sequence of action-conditioned future latent states for each trajectory proposed by TP-Flow. Each sequence remains associated with its motion candidate, while successive states represent its predicted evolution over time. This formulation captures candidate-specific future outcomes without reconstructing future images.

3 The ErrandVLN Benchmark

3.1 The ErrandVLN Task

As illustrated in Figure 2, conventional VLN-CE focuses on single-goal path following (Krantz et al., 2020). **ErrandVLN** instead models household errands as long-horizon navigation tasks with an overarching purpose and an ordered sequence of spatial requirements: the agent must visit specific semantic landmarks in a prescribed order to satisfy its navigation requirements. ErrandVLN formalizes household errands as sequential milestone navigation across semantic landmarks, represented as an ordered navigation skeleton:

$$\mathcal{S} = \left[(v_m, \ell_m, p_m^*)\right]_{m=1}^{M}, \qquad 1 \le M \le 4,$$

where stage m specifies a target functional area v_m, a semantic landmark category ℓ_m within that area, and a reachable target position $p_m^* \in \mathbb{R}^2$ near the landmark. The number of stages M is constrained to $M \le 4$ to balance task expressiveness (supporting patterns like loop closures $[A, B, C, A]$ and area revisits) with semantic coherence and tractable instruction generation.

An episode satisfies the ordered navigation requirements when all stage targets are visited in order. SR additionally requires the agent to issue `STOP` within $3.0\,\mathrm{m}$ of the final target, whereas OSR does not require `STOP`. We use a $3.0\,\mathrm{m}$ threshold for SR and OSR, and a stricter $1.5\,\mathrm{m}$ threshold for GCR and GC-SPL. Detailed task protocol and evaluation metrics are provided in Appendix A.1.

3.2 Environment Processing

From each scene, we extract static semantic landmarks by filtering bounding boxes and merging overlapping instances. Qwen-3.5-Plus parses these landmarks into an undirected area graph that represents functional regions and their spatial connectivity. The area graph defines the semantic search space for errand construction, while a two-dimensional occupancy grid supports continuous path planning. We inflate occupied regions by the agent footprint, $r_{\mathrm{agent}} = 0.15\,\mathrm{m}$, to support collision-aware navigation between landmarks. Processing details are provided in Appendix A.3.

3.3 Errand Skeleton Generation

Each ErrandVLN instance represents a household errand comprising an ordered set of spatial requirements embedded within an activity context. To ensure that these requirements correspond to physically executable trajectories, we construct the spatial component through a structured sampling procedure.

We generate candidate navigation skeletons through topology-guided random walks over the area graph, sampling valid semantic landmarks from the visited regions. This procedure yields diverse spatial patterns, including intra-area routines, cross-region transfers, multi-area traversals, loop closures, and regional revisits. A semantic validator filters implausible or trivial landmark combinations, ensuring that the selected sequence forms coherent requirements for a household errand.

For each validated skeleton, we select an approach pose for each target landmark, connect consecutive poses through shortest-path planning on the occupancy grid, and stitch the resulting segments into continuous trajectories. We discard collision-prone or kinematically irregular paths and remove geometrically redundant trajectories through trajectory-similarity filtering (see Appendix A.2 for realization and quality scoring).

3.4 Intent- and Role-Conditioned Instructions

Each ErrandVLN instance couples an activity context, a role-conditioned request, and an ordered set of spatial requirements. The activity context specifies the domestic purpose of the errand, the role defines the speaker's perspective and pragmatic framing, and the ordered requirements determine the sequence of landmarks evaluated by the benchmark. Given a validated navigation skeleton and semantic scene representation, we synthesize diverse instruction variants that span different motivations, roles, and linguistic realizations while strictly preserving the identical physical trajectory.

To systematically cover everyday household routines, instructions are organized into four canonical task families: **Spatial Navigation**, **Functional/Commonsense Reasoning**, **Interaction & Manipulation**, and **Complex/Composite Tasks**. These families represent distinct cognitive paradigms for contextualizing navigation goals, ranging from explicit spatial relations to multi-stage procedural routines. ErrandVLN evaluates whether an embodied agent can interpret these activity-grounded requests and sequentially visit the specified target milestones. Details of trajectory realization and LLM-based annotation prompts are provided in Appendices A.2 and A.3.

3.5 Dataset Splits

Following standard embodied navigation settings (Anderson et al., 2018), **ErrandVLN** partitions physical environments into Train, Val Scene-Unseen, and Test Scene-Unseen splits, with no physical scenes shared across these splits. Within the training environments, we additionally construct Val Seen-Path-Unseen for unseen trajectories in familiar scenes and Val Seen-Style-Unseen for scene-conditioned persona holdout on reused physical trajectories. These splits evaluate generalization across environments, navigation paths, and linguistic expressions. Split statistics and partitioning details are provided in Appendix A.4.

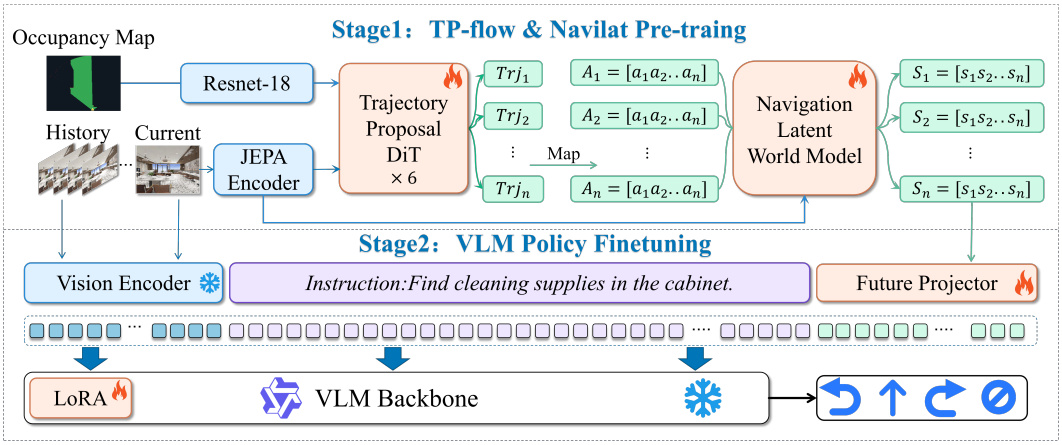

Figure 3: **Overall framework of VLN-JEPA.** Conditioned on recent observations and a depth-derived local traversability map, TP-Flow proposes continuous candidate trajectories. NaviLat predicts candidate-conditioned future latent states for the retained trajectories. These states are projected into Future State Tokens and provided to the VLM controller together with the instruction and visual history to generate the next executable action program.

4 VLN-JEPA: Action-Conditioned Latent Futures for Vision-and-Language Navigation

As illustrated in Figure 3, VLN-JEPA comprises three core modules: a continuous trajectory proposer (TP-Flow), an action-conditioned latent world model (NaviLat), and a language-conditioned VLM controller. In each planning cycle, TP-Flow proposes continuous candidate trajectories, NaviLat rolls out candidate-conditioned future latent states, and the controller uses these predictions to generate the next executable action program under receding-horizon control.

4.1 TP-Flow: Continuous Trajectory Proposer

TP-Flow generates temporally coherent continuous trajectory proposals over a multi-step planning horizon. Conditioned on recent latent history and an egocentric occupancy grid M_t constructed from onboard depth observations, TP-Flow produces $Q = 16$ waypoint sequences via a conditional flow model and retains the $K = 5$ highest-scoring candidates:

$$\{\tau_t^k\}_{k=1}^{K} = \mathrm{TopK}\left(\{\tilde{\tau}_t^q\}_{q=1}^{Q}; M_t\right). \tag{1}$$

The trajectory score combines traversability, forward progress, motion regularity, and geometric penalties. Each selected candidate τ_t^k is converted into a sequence of continuous distance–yaw actions for downstream latent-state forecasting. Map construction, trajectory scoring, and action conversion are detailed in Appendix B.3.

4.2 NaviLat: Navigation Latent World Model

NaviLat models environmental transitions directly within a compact latent space without pixel-level decoding. Given recent latent states $s_{t-h+1:t}$ and a continuous action a_t embedded as $e_a(a_t)$, NaviLat forecasts the subsequent latent state:

$$\hat{s}_{t+1} = f_\theta\left(s_{t-h+1:t}, e_a(a_t)\right). \tag{2}$$

The network is trained on environment transitions using next-state prediction with feature-variance regularization. During inference, NaviLat autoregressively rolls out $J = 5$ steps for each of the $K = 5$ candidates proposed by TP-Flow, yielding $K \times J = 25$ action-conditioned future latent states $\hat{s}_{t+j}^k$ for the controller (see Appendix B.2).

4.3 Navigation Policy with Future-State Awareness

To condition the VLM on predicted physical outcomes, a lightweight projector P maps the future latent states into Future State Tokens. Together with the instruction I and visual observation history $o_{\leq t}$, these tokens form the input to the policy π_ω:

$$u_t = \pi_\omega \left(I,\, o_{\leq t},\, P\left(\{\hat{s}^k_{t+j}\}^{K,J}_{k=1,j=1} \right) \right), \tag{3}$$

where u_t is the next executable continuous action program, serialized as a continuous action string, or `STOP`. Under a receding-horizon control protocol, the agent executes the generated program, receives the next observation, and replans from that observation. The token serialization, prompt templates, and action-decoding protocol are detailed in Appendix B.4.

5 Experiments

5.1 Evaluation Protocol

Baselines and comparison models. On **ErrandVLN**, we compare VLN-JEPA with representative continuous-navigation VLMs and zero-shot generalist MLLMs. External baselines use publicly released checkpoints and their native observation/action interfaces without ErrandVLN fine-tuning or adaptation; matched-data controls and module ablations follow the same Qwen3.5-2B training protocol as VLN-JEPA. Detailed configurations are provided in Appendix C.1.

Environments and metrics. We evaluate continuous vision-and-language navigation in R2R-CE and ErrandVLN. R2R-CE uses Habitat-Sim with Matterport3D environments (Krantz et al., 2020; Savva et al., 2019; Chang et al., 2017), while ErrandVLN uses NVIDIA Isaac Sim with photorealistic InteriorGS scenes (Miao et al., 2026). For R2R-CE, we report NE, OSR, SR, and SPL (Anderson et al., 2018). For ErrandVLN, we report GCR, GC-SPL, SR, OSR, PL, and Col. Complete evaluation definitions are provided in Appendix A.1.

Implementation details. VLN-JEPA training comprises two stages: Stage 1 first trains the state representation and NaviLat, then trains TP-Flow on expert paths; Stage 2 fine-tunes the policy controller with a Future State Projector and LoRA adapters. All stages use AdamW with BF16 mixed precision. Detailed training configurations and additional diagnostics are provided in Appendix B and Appendix C.

5.2 Main Results

Performance on ErrandVLN. As shown in Table 2, zero-shot generalist models struggle on ErrandVLN, peaking at 3.6% SR and 5.2% GCR. Scaling parameter count yields no monotonic improvement, as Qwen3.5-2B outperforms both Gemma-4-E4B and Qwen3.5-9B in GCR, indicating that capacity alone cannot resolve ordered multi-stage constraints.

Among specialized models, the compact 2.35B VLN-JEPA leads in multi-stage metrics, achieving 20.5% GCR, 29.2% OSR, and 16.1% GC-SPL. While the 8.38B DualVLN records a higher terminal SR of 15.1%, its truncated trajectory of 5.7 m alongside 44.3 collisions reflects premature stopping near the origin, leaving downstream milestones unreached at 12.9% GCR. In contrast, VLN-JEPA navigates deeper to 11.3 m with a 40% reduction in collisions, down to 26.6, confirming that action-conditioned latent forecasting sustains safe multi-stage progress.

Stratified analyses across horizons, revisits, and task families in Appendix C.2 reveal that scene familiarity maintains resilient sub-goal progression against topological backtracking, whereas novel scenes suffer steeper compounding errors. Trajectory rollouts and failure modes are detailed in Appendix C.7.

Performance on R2R-CE. To evaluate the cross-backbone transferability of our prospective latent framework, we augment the frozen JanusVLN backbone with frozen TP-Flow and NaviLat pretrained on ErrandVLN, fine-tuning only the LoRA adapters and the Future State Projector on R2R-CE (see Appendix B.7 for implementation details). As shown in Table 3, integrating prospective latent rollouts

Table 2: Comparison of navigation performance on the ErrandVLN Test Scene-Unseen split.

Model	Params	SR ↑	GCR ↑	OSR ↑	GC-SPL ↑	PL (m)	Col ↓
Vision-Language Models (Zero-Shot)							
Qwen3.5-2B	2B	3.6	5.2	8.4	4.0	5.4	90.6
Gemma-4-E4B	4B	0.0	2.9	4.6	2.8	2.2	71.1
Qwen3.5-9B	9B	2.0	3.1	5.0	3.0	1.3	24.6
VLM Based Navigation Models							
NaVid[RSS'24] (Zhang et al., 2024b)	7.04B	0.4	10.5	16.9	7.4	2.2	12.9
NaVILA[RSS'25] (Cheng et al., 2024)	8.46B	3.2	16.0	24.4	13.4	10.1	446.5
StreamVLN[ICRA'26] (Wei et al., 2025)	8.03B	11.2	11.2	18.5	10.7	5.8	57.9
DualVLN[ICLR26] (Wei et al., 2026)	8.38B	**15.1**	12.9	23.1	12.4	5.7	44.3
VLN-JEPA (Ours)	**2.35B**	12.6	**20.5**	**29.2**	**16.1**	11.3	26.6

Table 3: Comparison of representative VLN methods on R2R-CE Val-Unseen. On R2R-CE, following the baseline setup of JanusVLN (Zeng et al., 2026), we equip JanusVLN with TP-Flow and NaviLat to demonstrate generalizability across backbones.

Method	Observation				R2R Val-Unseen			
	Pano.	Odo.	Depth	S.RGB	NE↓	OSR↑	SR↑	SPL↑
CMA [ECCV'20] (Krantz et al., 2020)	✓	✓	✓		7.14	42.5	34.2	30.1
VLN↻BERT [CVPR'21] (Hong et al., 2021)	✓	✓	✓		6.85	46.8	38.6	34.2
GridMM [ICCV'23] (Wang et al., 2023b)	✓	✓	✓		5.11	61.0	49.0	41.0
ScaleVLN [ICCV'23] (Wang et al., 2023c)	✓	✓	✓		4.80	–	55.0	51.0
ETPNav [TPAMI'24] (An et al., 2024)	✓	✓	✓		4.71	65.0	57.0	49.0
InstructNav [CoRL'24] (Long et al., 2024)	✓	✓	✓	✓	6.89	–	31.0	24.0
WS-MGMap[NIPS'22] (Chen et al., 2022a)		✓	✓	✓	6.28	47.6	38.9	34.3
LAW[EMNLP'21] (Raychaudhuri et al., 2021)		✓	✓	✓	6.83	44.0	35.0	31.0
CM2[CVPR'22] (Georgakis et al., 2022)		✓	✓	✓	7.02	41.5	34.3	27.6
Seq2Seq [ECCV'20] (Krantz et al., 2020)			✓	✓	7.77	37.0	25.0	22.0
NavMorph [ICCV'25] (Yao et al., 2025)			✓	✓	5.75	56.9	47.9	33.2
monoVLN [ICCV'25] (Lu et al., 2025)			✓	✓	8.29	37.7	31.8	26.8
InternVLA-N1 [arXiv'25] (Team, 2025)			✓	✓	4.83	63.3	58.2	54.0
NaVid [RSS'24] (Zhang et al., 2024b)				✓	5.47	49.1	37.4	35.9
MapNav [ACL'25] (Zhang et al., 2025)				✓	4.93	53.0	39.7	37.2
Uni-NaVid [RSS'25] (Zhang et al., 2024a)				✓	4.96	57.4	51.8	47.7
NaVILA [RSS'25] (Cheng et al., 2024)				✓	5.22	62.5	54.0	49.0
JanusVLN [ICLR'26] (Zeng et al., 2026)				✓	4.78	65.2	60.5	56.8
StreamVLN [ICRA'26] (Wei et al., 2025)				✓	4.90	63.6	56.4	50.2
NavFoM [ICLR'26] (Zhang et al., 2026a)				✓	5.01	64.9	56.2	51.2
D3D-VLP [CVPR'26] (Wang et al., 2026)				✓	4.73	67.2	61.3	56.1
LatentPilot [ECCV'26] (Hao et al., 2026)				✓	**4.41**	66.3	62.0	58.0
JanusVLN + TP-Flow + NaviLat			✓	✓	4.59	**69.2**	**62.5**	**59.7**

yields consistent gains over the baseline JanusVLN, improving SR from 60.5% to 62.5%, OSR from 65.2% to 69.2%, and SPL from 56.8% to 59.7%.

Across all published continuous-navigation baselines on R2R-CE Val-Unseen, the adapted model attains leading performance, achieving 62.5% SR, 69.2% OSR, and 59.7% SPL alongside a low navigation error of 4.59 m. By outperforming strong contemporary methods such as LatentPilot (62.0% SR, 58.0% SPL) and StreamVLN (56.4% SR, 50.2% SPL), these results confirm that action-conditioned latent forecasting transfers effectively across simulated environments and provides informative prospective guidance to heterogeneous VLM backbones.

5.3 Ablation and Future-Token Sensitivity

Ablation of TP-Flow and NaviLat. Table 4 shows that controller fine-tuning itself provides a substantial improvement over the zero-shot base VLM: GCR increases from 5.2% to 13.0%, while

Table 4: Core architectural ablations on ErrandVLN Test Scene-Unseen, with a zero-shot reference. All fine-tuned variants use the same controller training examples and optimization settings. The w/o Both variant fine-tunes only the base VLM's LoRA adapters, without TP-Flow or NaviLat.

Method / Variant	**SR**↑	**GCR**↑	**GC-SPL** ↑	**PL** (m)	**Col** ↓
Base VLM (zero-shot)	3.6	5.2	4.0	5.4	90.6
w/o Both (SFT)	5.2	13.0	11.4	6.3	56.7
w/o NaviLat	9.8	16.2	12.9	19.9	27.0
w/o TP-Flow	11.2	18.1	14.3	12.5	38.5
VLN-JEPA	**12.6**	**20.5**	**16.1**	11.3	**26.6**

Table 5: Future Token perturbation analysis on ErrandVLN Test Scene-Unseen.

Variant	**SR**↑	**GCR**↑	**GC-SPL** ↑	**PL** (m)	**Col** ↓
Original Future	**12.6**	**20.5**	**16.1**	11.3	**26.6**
Shuffled Future	8.8	14.7	10.9	7.5	31.4
Zero Future	10.2	16.8	12.3	8.8	28.5

collisions decrease from 90.6 to 56.7. The proposed modules yield further gains beyond this training effect. The two modules contribute complementary improvements. Without NaviLat, the controller receives TP-Flow trajectories as text and achieves 9.8% SR and 16.2% GCR, but produces longer paths ($19.9\,\mathrm{m}$) than the full model ($11.3\,\mathrm{m}$). With TP-Flow removed, fixed action branches preserve future-state prediction but increase collisions from 26.6 to 38.5. These results indicate that TP-Flow provides safer and more suitable candidate motions, whereas NaviLat helps the controller compare candidate actions using their predicted future states. Combining both modules therefore improves task completion while maintaining shorter and safer trajectories.

Future Token Perturbation Analysis. As reported in Table 5, we evaluate the controller's sensitivity to prospective representations through two inference-time perturbations without retraining.

In *Zero Future*, replacing all 25 Future Tokens with zero vectors assesses **token-content sensitivity**. This intervention reduces SR from 12.6% to 10.2% and GCR from 20.5% to 16.8%, demonstrating that the controller actively conditions on and extracts predictive signals from the future latent states rather than relying solely on past observations and instructions. In *Shuffled Future*, randomly permuting the candidate-major token blocks across positions assesses **structural token-order and position sensitivity**. Because TP-Flow organizes proposals according to a consistent rank and geometric layout, shuffling disrupts this canonical structural ordering, degrading SR to 8.8% (GCR 14.7%) and increasing collisions from 26.6 to 31.4. Together, these results confirm that the controller is sensitive both to the informative latent content of future predictions and to their canonical structural layout across candidate slots.

6 Conclusion

We presented ErrandVLN, a continuous vision-and-language navigation benchmark for household errands featuring ordered semantic goals and instruction-conditioned revisits. ErrandVLN captures domestic intent through an automated pipeline spanning topological task sampling, physical clearance validation, geometric deduplication, and scene-conditioned instruction generation. We also introduced VLN-JEPA, an embodied navigation framework that couples continuous trajectory proposals with action-conditioned latent world modeling, supplying the VLM controller with prospective future context without observation decoding. On Test Scene-Unseen, VLN-JEPA achieves superior performance across goal completion (GCR), task success (SR), and collision avoidance. Together, ErrandVLN and VLN-JEPA establish an intent-driven evaluation paradigm and demonstrate the effectiveness of prospective latent planning for continuous embodied navigation.

AI USE STATEMENT

Generative AI tools were used to assist with Chinese-to-English translation, drafting and editing manuscript text, suggesting paper structure, summarizing and locating relevant literature, formatting references, and creating manuscript-native schematic figures. AI tools were not used to generate the dataset statistics or experimental measurements reported in this manuscript. AI-assisted claims and references were checked against the supplied source materials or retrieved publication records. The authors take responsibility for the final content, including text and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

ErrandVLN is constructed entirely from simulated indoor environments and does not involve human subjects, personal data, or real-world private information. All navigation trajectories and action supervision are generated algorithmically within the simulated environments. Language models are used only for semantic plausibility screening and instruction synthesis, and no personal or sensitive information is involved in this process. The underlying scene assets remain subject to their respective licenses and attribution requirements, which will be preserved in any dataset release. We acknowledge that the simulated environments and synthetic instructions may not fully capture the diversity of real-world homes and human navigation behavior, which may limit direct transfer to physical robots.

REPRODUCIBILITY STATEMENT

We document the benchmark construction, model training, and evaluation procedures in the main text and appendices. Section 3 and Appendix A describe the benchmark generation pipeline, dataset partitioning, task protocol, and evaluation metrics. Section 4 and Appendix B detail the model architecture, training stages, optimization settings, and dataset-specific adaptation procedures. Section 5 and Appendix C describe the experimental protocols, baseline configurations, and additional analyses. We plan to release the benchmark data, construction and evaluation code, split manifests, trained checkpoints, and training and evaluation configurations to support reproduction of the reported results.

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*, 47(7):5130–5145, 2024.

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 *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 3674–3683. IEEE, 2018.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Amirhosein Chahe and Lifeng Zhou. Policy-guided world model planning for language-conditioned visual navigation. *arXiv preprint arXiv:2603.25981*, 2026.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

Jiaqi Chen, Bingqian Lin, Ran Xu, Zhenhua Chai, Xiaodan Liang, and Kwan-Yee Wong. Mapgpt: Map-guided prompting with adaptive path planning for vision-and-language navigation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9796–9810, 2024.

Peihao Chen, Dongyu Ji, Kunyang Lin, Runhao Zeng, Thomas Li, Mingkui Tan, and Chuang Gan. Weakly-supervised multi-granularity map learning for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 35:38149–38161, 2022a.

Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History aware multimodal transformer for vision-and-language navigation. *Advances in neural information processing systems*, 34:5834–5847, 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. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16516–16526. IEEE, 2022b.

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.

Georgios Georgakis, Karl Schmeckpeper, Karan Wanchoo, Soham Dan, Eleni Miltsakaki, Dan Roth, and Kostas Daniilidis. Cross-modal map learning for vision and language navigation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15439–15449. IEEE, 2022.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3):440, 2018.

Haihong Hao, Lei Chen, Mingfei Han, Changlin Li, Dong An, Yuqiang Yang, Zhihui Li, and Xiaojun Chang. Latentpilot: Scene-aware vision-and-language navigation by dreaming ahead with latent visual reasoning. *arXiv preprint arXiv:2603.29165*, 2026.

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*, pp. 1643–1653, 2021.

Yanjia Huang, Xianshun Jiang, Xiangbo Gao, Mingyang Wu, and Zhengzhong Tu. VISTAv2: World imagination for indoor vision-and-language navigation. *arXiv preprint arXiv:2512.00041*, 2025.

Mukul Khanna*, Ram Ramrakhya*, Gunjan Chhablani, Sriram Yenamandra, Theophile Gervet, Matthew Chang, Zsolt Kira, Devendra Singh Chaplot, Dhruv Batra, and Roozbeh Mottaghi. Goat-bench: A benchmark for multi-modal lifelong navigation. In *CVPR*, 2024.

Jing Yu Koh, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. Pathdreamer: A world model for indoor navigation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14718–14728. IEEE, 2021.

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*, pp. 104–120. Springer, 2020.

Jacob Krantz, Shurjo Banerjee, Wang Zhu, Jason Corso, Peter Anderson, Stefan Lee, and Jesse Thomason. Iterative vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2210.03087.

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.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Wensi Ai, Benjamin Martinez, Hang Yin, Michael Lingelbach, Minjune Hwang, Ayano Hiranaka, Sujay Garlanka, Arman Aydin, Sharon Lee, Jiankai Sun, Mona Anvari, Manasi Sharma, Dhruva Bansal, Samuel Hunter, Kyu-Young Kim, Alan Lou, Caleb R Matthews, Ivan Villa-Renteria, Jerry Huayang Tang, Claire Tang, Fei Xia, Yunzhu Li, Silvio Savarese, Hyowon Gweon, C. Karen Liu, Jiajun Wu, and Li Fei-Fei. Behavior-1k: A human-centered, embodied ai benchmark with 1,000 everyday activities and realistic simulation. *arXiv preprint arXiv:2403.09227*, 2024.

Jialu Li and Mohit Bansal. Improving vision-and-language navigation by generating future-view image semantics. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10803–10812. IEEE, 2023.

Sihao Lin, Zerui Li, Xunyi Zhao, Gengze Zhou, Liuyi Wang, Rong Wei, Rui Tang, Juncheng Li, Hanqing Wang, Jiangmiao Pang, et al. Vlnverse: A benchmark for vision-language navigation with versatile, embodied, realistic simulation and evaluation. *arXiv preprint arXiv:2512.19021*, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

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.

Renjie Lu, Yu Zhou, Hao Cheng, Jingke Meng, and Wei-Shi Zheng. monovln: Bridging the observation gap between monocular and panoramic vision and language navigation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9477–9486. IEEE, 2025.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. Leworld-model: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026.

Bingchen Miao, Rong Wei, Zhiqi Ge, Shiqi Gao, Jingzhe Zhu, Renhan Wang, Siliang Tang, Jun Xiao, Rui Tang, Juncheng Li, et al. Towards physically executable 3d gaussian for embodied navigation. In *International Conference on Learning Representations*, volume 2026, pp. 22945–22965, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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*, pp. 9982–9991, 2020.

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*, 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 Geoffrey Gordon, David Dunson, and Miroslav Dudík (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9338–9346. IEEE, 2019.

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

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

InternNav Team. InternVLA-N1: An open dual-system navigation foundation model with learned latent plans, 2025.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Hanqing Wang, Wei Liang, Luc Van Gool, and Wenguan Wang. Dreamwalker: Mental planning for continuous vision-language navigation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10839–10849. IEEE, 2023a.

Zihan Wang, Xiangyang Li, Jiahao Yang, Yeqi Liu, and Shuqiang Jiang. Gridmm: Grid memory map for vision-and-language navigation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15579–15590. IEEE, 2023b.

Zihan Wang, Seungjun Lee, Guangzhao Dai, and Gim Hee Lee. D3d-vlp: Dynamic 3d vision-language-planning model for embodied grounding and navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32463–32474, 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. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986. IEEE, 2023c.

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

Meng Wei, Chenyang Wan, Peng Peng, Xiqian Yu, Yuqiang Yang, Delin Feng, Wenzhe Cai, Chenming Zhu, Tai Wang, Jiangmiao Pang, et al. Ground slow, move fast: A dual-system foundation model for generalizable vision-language navigation. In *International Conference on Learning Representations*, volume 2026, pp. 12380–12396, 2026.

Xuan Yao, Junyu Gao, and Changsheng Xu. Navmorph: A self-evolving world model for vision-and-language navigation in continuous environments. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5536–5546. IEEE, 2025.

Shuang Zeng, Dekang Qi, Xinyuan Chang, Feng Xiong, Shichao Xie, Xiaolong Wu, Shiyi Liang, Mu Xu, and Xing Wei. Janusvln: Decoupling semantics and spatiality with dual implicit memory for vision-language navigation. In *International Conference on Learning Representations*, volume 2026, pp. 33001–33026, 2026.

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.

Jiazhao Zhang, Anqi Li, Yunpeng Qi, Minghan Li, Jiahang Liu, Shaoan Wang, Haoran Liu, Gengze Zhou, Yuze Wu, Xingxing Li, et al. Embodied navigation foundation model. In *International Conference on Learning Representations*, volume 2026, pp. 127293–127322, 2026a.

Lingfeng Zhang, Xiaoshuai Hao, Qinwen Xu, Qiang Zhang, Xinyao Zhang, Pengwei Wang, Jing Zhang, Zhongyuan Wang, Shanghang Zhang, and Renjing Xu. Mapnav: A novel memory representation via annotated semantic maps for vlm-based vision-and-language navigation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13032–13056, 2025.

Lingfeng Zhang, Zhanguang Zhang, Liheng Ma, Tongtong Cao, and Yingxue Zhang. Anticipate before acting: Future-state-conditioned vision-language navigation. *arXiv preprint arXiv:2607.18042*, 2026b.

Fengda Zhu, Xiwen Liang, Yi Zhu, Qizhi Yu, Xiaojun Chang, and Xiaodan Liang. Soon: Scenario oriented object navigation with graph-based exploration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12689–12699, 2021.