NavForesee: A Unified Vision-Language World Model for Hierarchical Planning and Dual-Horizon Navigation Prediction

Fei Liu* Shichao Xie* Minghua Luo Zedong Chu[†]

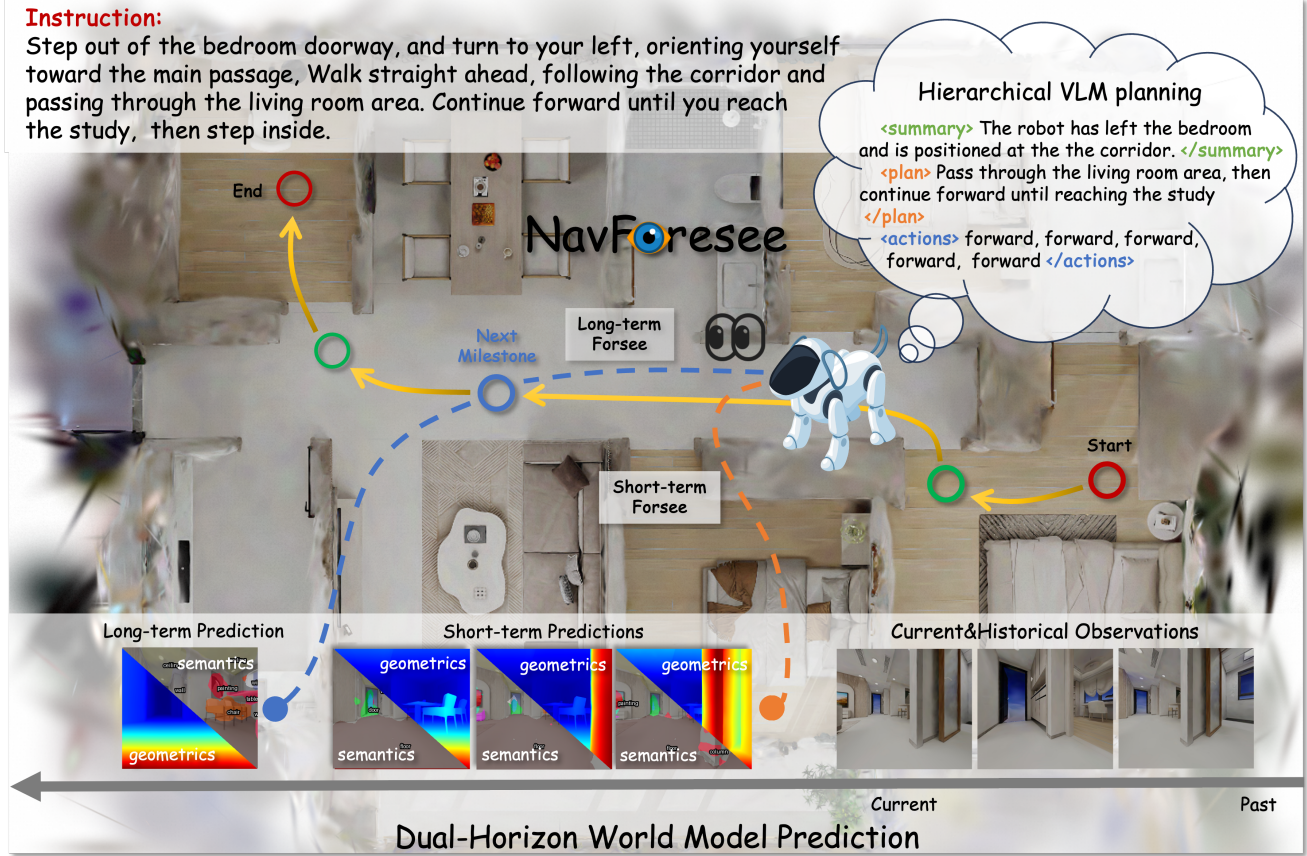Junjun Hu Xiaolong Wu[†] Mu Xu

Amap, Alibaba Group

Fig. 1: NavForesee integrates hierarchical language planning with dual-horizon predictive foresight. The planner decomposes instructions into milestone-based sub-goals, while the world model predicts high-level environmental features for long- and short-term guidance, producing coherent navigation actions.

Abstract—Embodied navigation for long-horizon tasks, guided by complex natural language instructions, remains a formidable challenge in artificial intelligence. Existing agents often struggle with robust long-term planning about unseen environments, leading to high failure rates. To address these limitations, we introduce NavForesee, a novel Vision-Language Model (VLM) that unifies high-level language planning and predictive world model imagination within a single, unified framework. Our approach empowers a single VLM to concurrently perform planning and predictive foresight. Conditioned on the full instruction and historical observations, the model is trained to understand the navigation instructions by decomposing the task, tracking its progress, and formulating the subsequent sub-goal. Simultaneously, it functions as a generative world model, providing crucial foresight by predicting short-term environmental dynamics and long-term navigation milestones. The VLM's structured plan guides its targeted prediction, while the imagined future provides rich context to inform the navigation actions, creating a powerful internal feedback loop of perception-planning/prediction-action. We demonstrate through extensive experiments on the R2R-CE and RxR-CE benchmark that NavForesee achieves highly competitive performance in complex scenarios. Our work highlights the immense potential of fusing explicit language planning with implicit spatiotemporal prediction, paving the way for more intelligent and capable embodied agents.

I. INTRODUCTION

Embodied navigation, a cornerstone challenge in artificial intelligence, has recently witnessed remarkable progress driven by the advent of Vision-Language Models (VLMs)

*Joint first authors   †Corresponding authors

[1], [2], [3], [4], [5]. These models endow agents with the ability to perceive, interpret instructions, and operate in complex environments. Despite these advances, a significant performance gap persists in long-horizon tasks, where agents frequently fail to maintain course, comprehend observations, or make consistently correct decisions. This gap stems from two primary limitations: (1) a planning and memory deficit, as deployable VLMs often have limited context windows and planning capabilities, causing them to get "lost" in the navigation environment[6], [7], [8]; and (2) a lack of predictive foresight, as current models are fundamentally reactive and cannot anticipate future environmental states to guide their actions proactively [9], [10], [11].

Existing research has pursued these challenges on separate fronts. One trajectory enhances VLM reasoning through curated datasets and Chain-of-Thought (CoT) prompting [12], [13]. The other develops world models to predict future states, informing action planning [14], [15]. However, a critical oversight is the disconnection between these paradigms. A VLM-centric agent can suffer from semantic hallucinations, where its plan disconnects from visual reality, while a world model without language guidance can experience semantic drift, its predictions becoming untethered from the instructional goal.

We posit that VLM planning and predictive foresight should not be separate but unified and mutually reinforcing within a single VLM [16]. To this end, we introduce NavForesee as in Figure 1, a unified model that integrates multi-modal understanding with world model generation. Our approach is inspired by human navigation, which is not a continuous, low-level process but a hierarchical one centered on milestones. Humans typically navigate by heading towards a sequence of meaningful landmarks, largely ignoring the minutiae of the path between them. We argue that an artificial agent should do the same. NavForesee adopts this strategy by operating through two synergistic functions: (1) Hierarchical Language planning. As a planner, NavForesee generates a high-level plan by summarizing the navigation task into completed sub-instructions, identifying the current sub-instruction, and formulating the next steps as semantic action "trunks." This grounds the agent's planning in the overall instruction. (2) Dual-Horizon Predictive Foresight. As a world model, NavForesee "imagines" the future on two timescales. For long-term guidance, it predicts the key visual features of the environment at the completion of the current sub-instruction—effectively envisioning the next milestone. For short-term execution, it forecasts immediate future features to enhance local awareness, enabling robust obstacle avoidance and better understanding of environmental dynamics. Inspired by latent-space world models [17], [18], [19], [15], this prediction deliberately avoids computationally expensive pixel-level generation. Instead, NavForesee forecasts a compact set of high-level features—depth, DINOv2, and SAM features—that capture essential geometric and semantic information as in DreamVLA. The predicted features are fed to an action policy module which is simply an MLP to generate continuous waypoints and flags for arriving or not.

By tightly coupling hierarchical planning with dual-horizon predictive foresight, NavForesee generates coherent, goal-oriented actions, guided by both a long-term vision of its milestones and an immediate awareness of its surroundings.

We conducted extensive experiments on the R2R-CE [20] and RxR-CE [21] benchmarks. Training exclusively on the publicly available R2R-CE and RxR-CE datasets, NavForesee demonstrates highly competitive performance, achieving a Success Rate (SR) of 66.2% and an Oracle Success Rate (OSR) of 78.4% on the R2R-CE benchmark—comparable to state-of-the-art methods. In summary, our key contributions are threefold:

- We propose NavForesee, a VLN framework that unifies vision–language model (VLM) planning with world model prediction for navigation tasks.
- We introduce a hierarchical language planning paradigm that addresses long-instruction, goal-oriented missions by explicitly tracking mission progress and generating concise textual sub-plans.
- We design a dual-horizon world model prediction mechanism for both short-term execution and long-term milestone navigation, implicitly forming a perception–planning and prediction–action loop that guides agent behavior.

II. RELATED WORKS

A. Visual Language Navigation

Vision-and-Language Navigation (VLN) requires an embodied agent to interpret natural language instructions, perceive visual surroundings, and generate a sequence of actions to reach a specified goal. The advent of large-scale pre-trained VLMs has catalyzed significant progress, largely superseding earlier methods based on topological graphs [22], [23], [24], top-down semantic maps [25], [26], [27], or instruction augmentation [28]. Recent works leveraging VLMs can be broadly categorized into two main paradigms. The first uses the VLM as a high-level planner, auto-regressively generating action plans [29], [30], [31] or textual trajectories [32]. While strong in reasoning, this step-by-step generation is prone to error accumulation and slow inference. The second employs the VLM as an end-to-end policy, directly mapping inputs to actions. However, this often leads to overfitting on training scenes and underutilizes the VLM's high-level reasoning capabilities.

To bridge the gap between these two approaches, dual-system architectures have been proposed [6], [33]. These models often adopt a "Fast-and-Slow" reasoning paradigm, combining a deliberative "slow" system for high-level reasoning with a lightweight "fast" reactive controller for low-level execution. Reinforcement learning is frequently employed to align the outputs of both systems and bootstrap the learning of coherent reasoning-action patterns. Despite this progress, a fundamental challenge remains: long, complex reasoning chains (e.g., long CoTs) do not always align with the spatial and dynamic realities of the environment. Furthermore, frequent or periodic elaborate reasoning processes
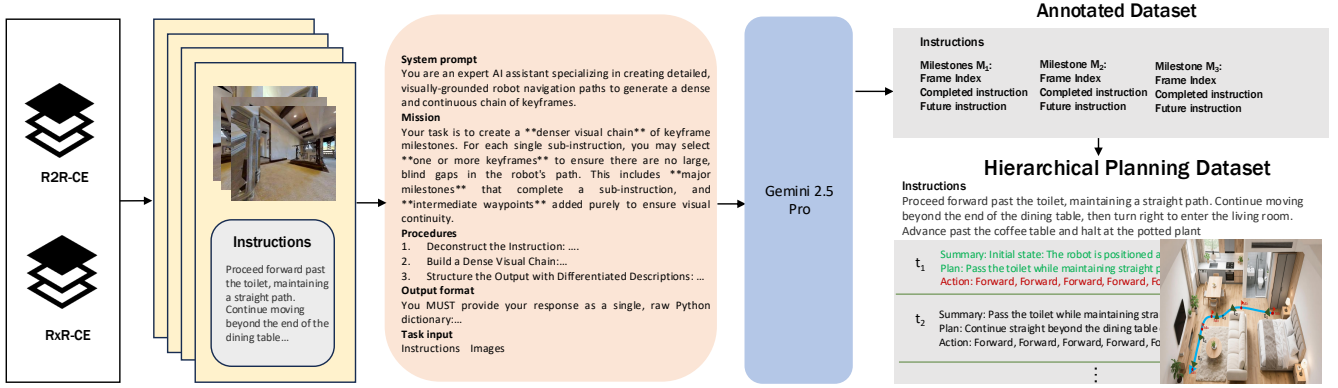
Fig. 2: VLM-driven hierarchical navigation plan dataset generation. Episodes from R2R-CE and RxR-CE are processed by `Gemini 2.5 Pro`, which decomposes long instructions into sub-instructions and identifies keyframe milestones. To generation of waypoint-level reasoning labels, waypoints are sampled between milestones annotated with a navigation summary, future plan, and action (`forward`, `left`, `right`, or `stop`).

may be unnecessary, as human navigation often relies on simpler, high-level semantic plans rather than continuous, detailed deliberation.

B. Navigation World Model

A world model is designed to learn a predictive model of an environment, forecasting future states from historical observations and optional conditioning information, such as actions or instructions. Predictions can be generated in either raw pixel space or a more compact latent space [18]. The concept has gained significant traction recently, propelled by large-scale video generation models like Sora([34], which can produce long-term, consistent, even interactive video sequences from text prompts. A key application of world models in robotics is to serve as a simulation engine, allowing an agent to "imagine" the outcomes of different action sequences and evaluate control policies before execution [16], [19].

In the context of visual navigation, recent works have begun to leverage world models to provide agents with environmental foresight. For instance, NavMorph utilizes a Recurrent State-Space Model (RSSM) to model environmental dynamics in a compact latent space, refining the agent's policy with imagined future states [11]. Similarly, HNR [9] advocates for predicting multi-level semantic features instead of raw pixels, enabling faster and higher-quality imagination to evaluate multiple next-step actions in parallel. Other approaches, like NWM [35], use a controlled video generation model to plan entire trajectories through simulation. Despite their promise, existing world models for navigation face two primary limitations. First, action-conditioned models that rely on extensive trajectory sampling and evaluation are often computationally prohibitive, rendering them infeasible for deployment on resource-constrained agents. Second, and more critically for our work, prior research has focused almost exclusively on learning environmental dynamics, largely neglecting to integrate this predictive capability with the high-level language reasoning abilities of modern VLMs.

This separation leaves a critical gap, which our work aims to address by unifying these two powerful paradigms.

III. METHODS

A. Problem Formulation

We target instruction-guided navigation missions in which an embodied agent must interpret a natural language instruction l and navigate from a given start position to an intended goal location, strictly following the described route. The challenge lies in robustly understanding the instruction, maintaining situational awareness over long horizons, and deciding actions that lead to successful navigation in unseen environments.
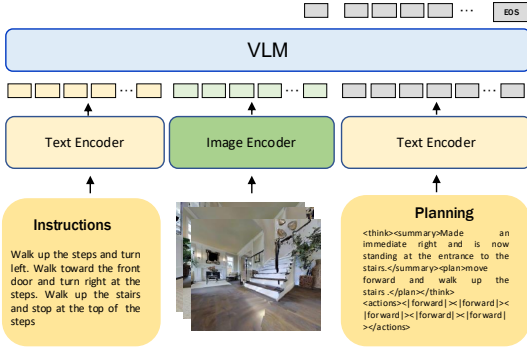
At time step t, the agent perceives the environment and obtains a panoramic RGB observation o_t. It maintains a memory of the past H observations, $O_{t-H:t-1} = [o_{t-H}, \ldots, o_{t-1}]$, to support temporal reasoning. The navigation policy produces a sequence of K future waypoints $w_{t:t+K} \in \mathbb{R}^{K \times 5}$, where each waypoint is defined as

$$w_t = [x_t, y_t, \sin \theta_t, \cos \theta_t, c_t],$$

with (x_t, y_t) denoting planar positions, θ_t the heading angle, and the binary flag c_t indicating whether a `stop` action should be triggered. Unless all predicted actions are marked as `stop`, the agent continuously moves following the generated waypoints.

To solve this problem, we adopt Qwen2.5-VL as our backbone and extend it with two complementary modules. First, we enable **hierarchical planning** by decomposing the full instruction into sequential sub-instructions, identifying completed ones and predicting the next step under the current context—leveraging the model's language understanding capabilities and pretraining on our constructed dataset. Second, we integrate **world model foresight** for predicting short- and long-term environmental changes, enhancing vision–language coherence and yielding more reliable action policies. Together, these capabilities allow the agent to imitate human navigation behaviors, combining explicit language planning with implicit spatiotemporal prediction.
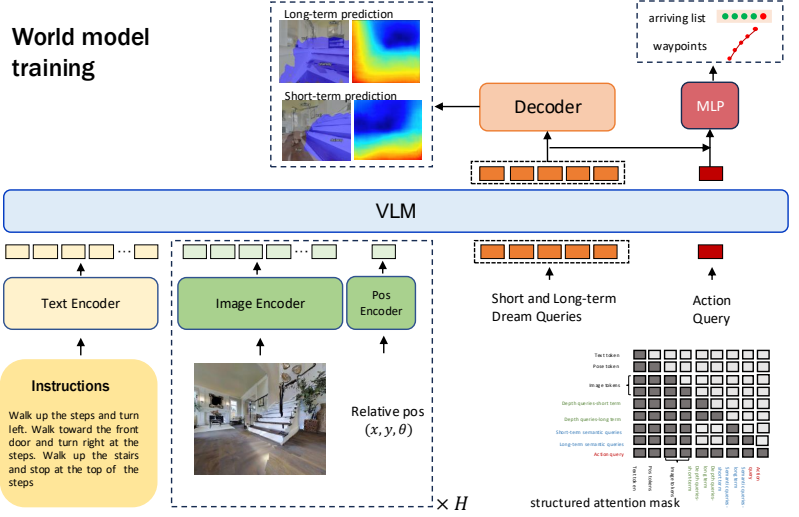
Fig. 3: Overall architecture of NavForesee. The model is built on the Qwen2.5-VL-3B-Instruct backbone, integrating two complementary functionalities: (1) VLM-based hierarchical planning and (2) world model-based dual-horizon visual prediction. For hierarchical planning, textual instruction and visual observations are encoded via Qwen's original multimodal encoders to produce auto-regressive sub-goal plans. For prediction, a position encoder encodes the agent's relative pose, and short- and long-horizon dream queries (depth and semantic subqueries) are appended to multimodal embeddings. These queries, processed through structured attention, feed lightweight convolutional decoders for environmental predictions and an MLP head for navigation actions.

B. VLM-driven Hierarchical Planning Dataset

We construct a hierarchical language planning dataset specifically for instruction-guided navigation missions, leveraging advanced Vision–Language Models (VLMs) for multimodal understanding and sequence analysis. Our goal is to provide training data that captures both short-term execution steps and long-term navigation milestones.

As illustrated in Figure 2, we start from public Vision-and-Language Navigation (VLN) benchmarks—R2R-CE (10k episodes) and RxR-CE (20k episodes)—which provide paired natural language instructions and full image observation sequences. Each episode is processed with Gemini 2.5 Pro, guided by a custom prompt template that specifies the model's role, defines the mission, outlines analytical steps, and enforces an explicit output format. The VLM systematically decomposes each long instruction into a series of sequential sub-instructions, while identifying a dense visual chain of keyframes representing navigation milestones. For paths involving extended travel or sharp turns, we require the inclusion of intermediate milestones to maintain visual continuity in the generated plan. This hierarchical structure enables downstream world models to better learn both short-term and long-term prediction.

For every annotated episode, the output is standardized to include: the milestone frame index, the textual description of the completed sub-instruction, and the upcoming planned instruction. Post-processing involves filtering incomplete annotations, correcting logical inconsistencies in the VLM outputs, and converting each episode into multiple navigation segments. We sample waypoints along each trajectory, with each waypoint forming the endpoint of a segment between milestones. Each sampled waypoint is assigned a *planning label* comprising: (1) a navigation summary (completed sub-instruction), (2) a future plan (next instruction), and (3) a language action (forward, left, right, stop).

This pipeline produces approximately 1.3M training samples from RxR-CE and 0.2M from R2R-CE. To ensure balanced training data, we down-sample over-represented straight-motion cases and augment examples involving stopping actions. The final dataset provides richly annotated, balanced samples for training the hierarchical language planning and predictive modules in NavForesee.

C. Model Architecture

Overall Architecture The overall architecture of NavForesee is illustrated in Figure 3. We adopt Qwen2.5-VL-3B-Instruct [36], a large-scale vision–language model with strong multi-modal understanding capabilities, as the backbone. NavForesee is designed to integrate two complementary functionalities: VLM-based language planning and World model-based visual prediction. Correspondingly, we define two primary training objectives: VLM planning training and world model training. Training data from both tasks are jointly mixed to ensure that the model preserves its multi-modal planning ability, while simultaneously extending its capability to generate visual features. For the VLM planning training, textual planning data are directly fed into Qwen for auto-regressive training, leveraging its original text encoder and image encoder components without modification. For the world model training, we introduce an additional position encoder (pos-encoder) to encode the agent's relative position and orientation from image observations. Two sets of dream queries—corresponding to short- and long-horizon predictions—are appended to the multi-modal embeddings.

Each set of dream queries includes depth and semantics sub-queries, enabling dual-horizon prediction. Furthermore, an action query, alongside the dream queries, is integrated into the multi-modal inputs and processed by Qwen2.5-VL via a structured attention mechanism. Lightweight convolutional layers serve as decoders to transform dream embeddings into environmental predictions (depth and semantics), while a simple MLP predicts action outputs (waypoints, orientation estimates, and arrival flags).

Structured Attention Mask To maintain a clear separation between short- and long-horizon predictions and to avoid cross-type contamination, each dream query type (depth and semantics) is explicitly decomposed into short-horizon and long-horizon components. As illustrated in Figure 3, we design a structured attention mask tailored for dual-horizon prediction. Long-horizon predictions naturally depend on short-horizon predictions, using them as guidance to ensure temporal coherence. Mutual attention between depth and semantics queries is masked to prevent cross-modal leakage or unintended feature mixing. In contrast, the action query attends to all available information—including past context and both horizons of dream queries—enabling it to make globally consistent navigation predictions.

D. Dual-horizon World Model Prediction

Specifically, to enable accurate dual-horizon environmental feature prediction, we employ the world model architecture that serves as guidance for learning the inverse dynamics of a navigation agent. Here, short-term prediction refers to generating forecasts for k steps ahead, while long-term prediction targets navigation milestones, corresponding to an adaptive horizon determined by progress toward the next milestone.

For visual feature prediction within Qwen2.5-VL, we introduce two sets of learnable dream queries, namely the short-term $Q_S \in \mathbb{R}^{L \times d}$ and and long-term $Q_L \in \mathbb{R}^{L \times d}$ to, which extract temporally aligned feature embeddings specialized for prediction at distinct horizons. To enhance the model's capability in capturing spatial-temporal correlations and learning environmental dynamics, we further integrate position-orientation state embeddings $s_{t-H:t}$ for each input frame through an encoder $h(.)$. These dream queries are concatenated with textual instruction embeddings l and visual observation sequences $O_{t-H:t}$ and processed by the Qwen2.5-VL backbone $f(.)$. Specially,

$$E_S = f(l, O_{t-H:t}, h(s_{t-H:t})|Q_S),$$
$$E_L = f(l, O_{t-H:t}, h(s_{t-H:t}), Q_S|Q_L)$$

where causal attention masking ensures auto-regressive generation: short-term embeddings are produced first, and long-term embeddings are conditioned on short-term predictions.

We design lightweight decoders to interpret E_L and E_S into predicted depth d_p, and high-level semantics c_p (e.g. derived from DINOV2, SAM). Short-term predictions correspond to a fixed horizon k whereas long-term predictions adaptively extrapolate over M_t steps, dependent on the agent's progress toward the next milestone:

$$p_{t+k} = D(E_S) = [d_p(t), c_p(t)],$$
$$p_{t+M_t} = D(E_L) = [d_p(t+M_t), c_p(t+M_t)]$$

E. Predictive Action Policy Learning

Given two temporally order states o_t and o_{t+1}, the intermediate action $\hat{a}(t)$ can be inferred via inverse dynamics. We leverage this principle to learn an action policy conditioned on the instruction l, historical visual observations $O_{t-H:t}$ and the dual-horizon predictive latent features E_S and E_L generated by the world model. To enhance the encoding of task-relevant context for action prediction, we introduce a learnable action query Q_a. This query is concatenated with the dream queries and multi-modal input embeddings to form a unified action embedding. The Qwen2.5-VL backbone processes these embeddings to produce the contextual representation for action inference, which is subsequently projected into the action space:

$$E_a = f(l, O_{t-H:t}, h(s_{t-H:t}), Q_S, Q_L|Q_a)$$
$$\hat{a}_{t:t+k} = M_{inv}(E_S, E_L|E_a)$$

where E_a is the action embedding and M_inv denotes the inverse dynamics model. Notably, in our action policy learning pipeline, the action embedding E_a is extracted through the Qwen2.5-VL backbone, while action predictions are primarily conditioned on the dual-horizon predictive features, ensuring that decisions are informed by both past observations and forecasted environmental dynamics

F. Close the Planning/Prediction and Action Loop

For VLM planning training, we finetune Qwen2.5-VL model based on the constructed dataset in an auto-regressive manner independently to build a powerful model capable of conducting hierarchical navigation.

For world model prediction and action policy learning, the training tasks are classified depth prediction, semantic feature prediction and action prediction. Depth prediction error L_d is measured using the Scale-invariant Logarithmic Loss (SiLogLoss) at the pix-level level. The semantics feature prediction error L_c and action error L_a are computed using mean squared error (MSE). The overall training loss L comprise L_d, L_c and L_a

$$L = \alpha L_d + \beta L_c + L_a$$

where α and β are weighting hyperparameters balancing the tasks.

IV. EXPERIMENTAL EVALUATION

We evaluate our model in continuous environment of the Habitant simulator on the R2R-CE and RxR-CE datasets.

R2R-CE dataset is derived from the Matterport3D indoor environments, discretized for path planning but operationalized in the Habitat simulator under a continuous navigation setting. It provides fine-grained, step-by-step natural language instructions, allowing for detailed guidance at each navigation step. In the simulator, the embodied agent can

TABLE I: Comparison with other methods on the Val-Unseen split of R2R-CE and RxR-CE

Method	Observation				R2R-CE Val-Unseen				RxR-CE Val-Unseen		
	S.RGB	Pano.	Depth	Odo	NE↓	OS↑	SR↑	SPL↑	NE↓	SR↑	SPL↑
HPN+DN* [37]		✓	✓	✓	6.31	40.0	36.0	34.0	-	-	-
CMA* [38]		✓	✓	✓	6.20	52.0	41.0	36.0	8.76	26.5	22.1
Sim2Sim [39]		✓	✓	✓	6.07	52.0	43.0	36.0	8.76	26.5	22.1
GridMM* [8]		✓	✓	✓	5.11	61.0	49.0	41.0	-	-	-
DreamWalker* [40]		✓	✓	✓	5.53	59.0	49.0	44.0	-	-	-
Reborn* [41]		✓	✓	✓	5.40	57.0	50.0	46.0	5.98	48.6	42.0
ETPNav* [42]		✓	✓	✓	4.71	65.0	57.0	49.0	5.64	54.7	44.8
HNR* [9]		✓	✓	✓	4.42	67.0	61.0	51.0	5.50	56.3	46.7
AG-CMTP [43]		✓	✓	✓	7.90	39.0	23.0	19.0	-	-	-
R2R-CMTP [43]		✓	✓	✓	7.90	38.0	26.0	22.0	-	-	-
Instruc-Nav [31]		✓	✓		6.89	-	31.0	24.0	-	-	-
LAW [44]	✓		✓	✓	6.83	44.0	35.0	31.0	10.90	8.0	8.0
CM2 [45]	✓		✓	✓	7.02	41.0	34.0	27.0	-	-	-
WS-MGMap [46]	✓		✓	✓	6.28	47.0	38.0	34.0	-	-	-
AO-Planner [47]		✓	✓		5.55	59.0	47.0	33.0	-	-	-
Seq2Seq [48]	✓		✓		7.77	37.0	25.0	22.0	12.10	13.9	11.9
CMA [48]	✓		✓		7.37	40.0	32.0	30.0	-	-	-
NA Vid [49]	✓				5.47	49.0	37.0	35.0	-	-	-
Uni-NA Vid [50]	✓				5.58	53.5	47.0	42.7	6.24	48.7	40.9
NaVILA [51]	✓				5.22	62.5	54.0	49.0	6.77	49.3	44.0
Stream VLN [52]	✓				4.98	64.2	56.9	51.9	6.22	52.9	46.0
CorrectNav [53]	✓				4.24	67.5	65.1	**62.3**	**4.09**	**69.3**	**63.3**
NavForesee(Ours)		✓			**3.94**	**78.4**	**66.2**	59.7	4.20	66.3	53.2

TABLE II: Performance comparison between VLM planning and dual-horizon world model prediction

Index	VLM planning	Long-term prediction	Short-term prediction	SR↑	OSR↑	NE↓	SPL↑
1	✓	✓	✓	66.2%	78.4%	3.94	59.7%
2	✗	✓	✓	48.8%	75.5%	5.61	39.4%
3	✓	✗	✓	58.6%	76.4%	4.47	50.1%
4	✗	✗	✗	52.6%	67.4%	5.53	46.7%

execute turns as small as 15° and perceives the scene through a 90° horizontal field-of-view.

RxR-CE is a large-scale, multilingual VLN dataset comprising about 126K human-annotated instructions. Compared to R2R-CE, RxR-CE covers more diverse and complex trajectories, increasing the difficulty of the navigation tasks. The agent in this setting uses a coarser minimum turn increment of 30° and a narrower 79° horizontal field-ofview, which demands more deliberate movement planning for effective scene coverage.

We evaluate navigation performance using standard metrics including success rate (SR), oracle success rate (OS), success weighted by path length (SPL), and navigation error (NE).

A. Comparison with State-of-the-Art Methods

Table 1 reports the performance of the proposed method compared with other approaches on the R2R-CE and RxR-CE datasets. Overall, NavForesee delivers competitive results against state-of-the-art (SOTA) methods. Specifically, on the val unseen split of the R2R-CE dataset, NavForesee achieves SOTA performance by improving SR by 1.1%, OSR by 10.9%, and reducing NE by 0.3 m. This improvement can be attributed to the integration of the world model prediction module, which enables the agent to better capture environmental dynamics, avoid obstacles, and explore the surroundings more effectively.

In contrast, NavForesee performs slightly worse than SOTA methods on RxR-CE, indicating limited generalization to more complex environments. It is worth noting that we train soly on NavForesee on R2R-CE and RxR-CE datasets, whereas other methods exploit diverse and large-scale datasets to enhance generalization. Although NavForesee does not consistently outperform all baselines, it achieves the highest OSR across both datasets. This demonstrates the value of incorporating world model prediction into VLN agents and implies its promising potential for future vision-and-language navigation tasks.

B. Ablation Study

As shown in Table II, removing any of the three key modules—hierarchical VLM planning, long-term prediction, or short-term prediction—results in clear performance degradation. The full NavForesee model, which combines all modules, achieves the highest SR (66.2%), OSR (78.4%), lowest NE (3.94), and best SPL (59.7%), validating the benefit of their integration. Without VLM planning, the success rate drops sharply to 48.8% and the SPL decreases by more than 16 points, reflecting the importance of explicit instruction decomposition and progress tracking for efficient navigation. Disabling long-term prediction also leads to a noticeable reduction in SR (58.6%) and higher NE, highlighting the role
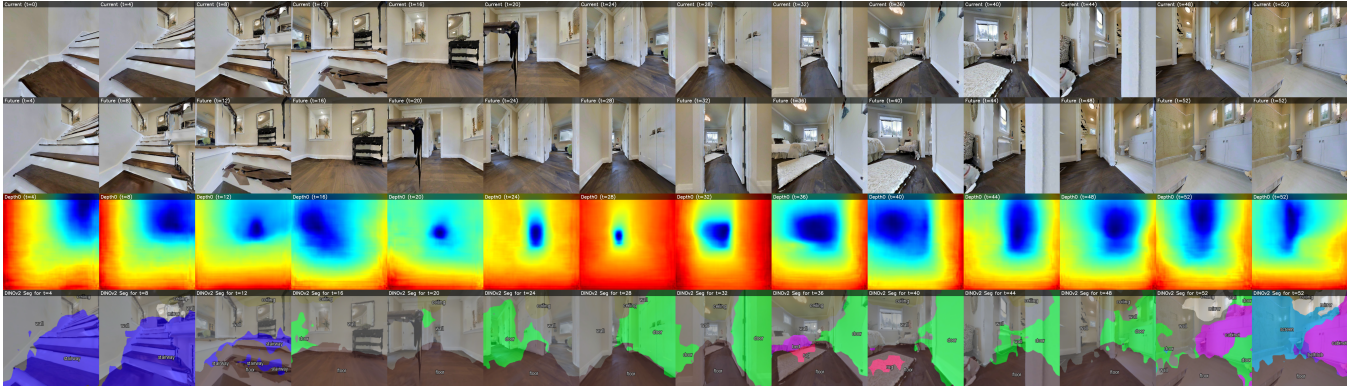
Fig. 4: Short-term depth and semantics predictions. From top to bottom: frames with timestamps, future ground truth frames with timestamps, future depth prediction for future frames, semantics predictions for future frames. Semantic features are DinoV2 features and visualized by a pretrained segmentation head. Instructions: UP the stairs. Turn to the left and enter into the second open door on the left. Walk towards the foot of the bed. Turn right and enter the open door to the bathroom
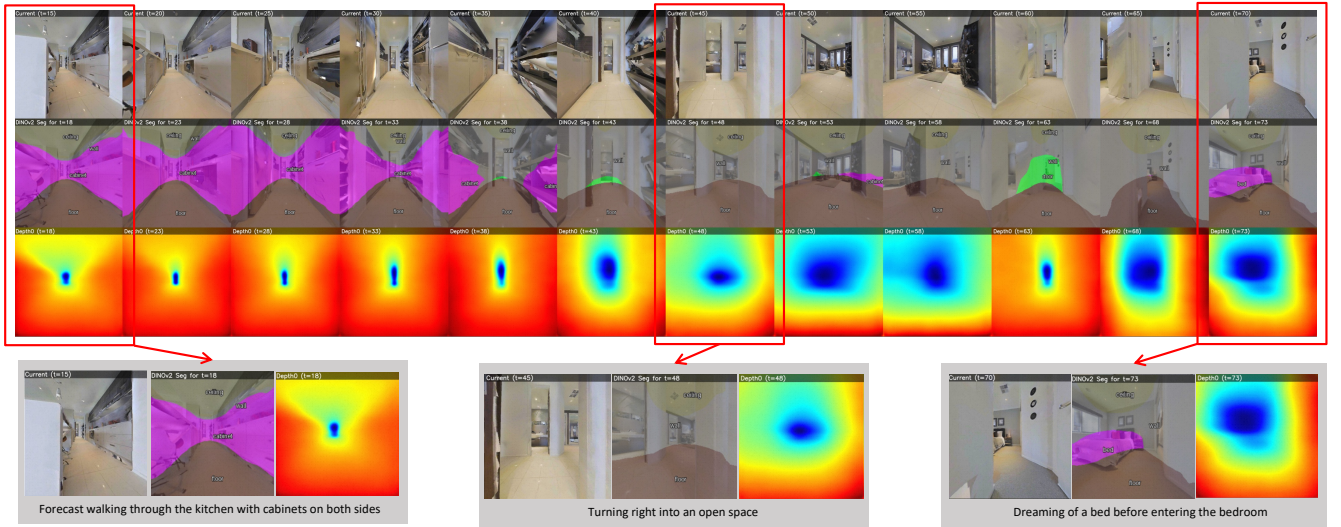

Fig. 5: NavForesee's geometric-semantic feature imagination across different motion modes. The model accurately predicts environmental dynamics in straight motion, generalizes effectively to turning scenarios, and infers detailed object geometry and depth distribution from minimal visual input, such as a brief glimpse into a room

of milestone foresight in providing strategic guidance over extended trajectories. When all three modules are removed, navigation quality deteriorates the most, confirming that planning and both prediction horizons together are crucial for accurate, efficient long-horizon navigation.

C. Qualitative Analysis

Figure 4 illustrates the short-term depth and semantic feature predictions generated by our world model over the course of a complete navigation episode, forecasting up to four future steps. Although the predicted depth maps appear somewhat coarse—owing to the constraints of pixel-level supervised training on R2R-CE and RxR-CE—they nonetheless preserve the scene's global geometry and spatial layout, faithfully capturing agent movements such as ascending or descending staircases, entering and exiting rooms, and making sharp or gradual turns. This ability to retain high-level spatial coherence despite reduced pixel detail ensures that

the model's predictions remain informative for downstream navigation decisions. The semantics predictions, obtained via a pretrained segmentation head, exhibit strong alignment with ground truth labels, successfully reflecting dynamic environmental changes in synchrony with the agent's actions.

Figure 5 further provides detailed examples that showcase NavForesee's ability to imaginatively anticipate semantic features across diverse motion patterns. In addition to delivering accurate environment dynamics predictions when following a straightforward trajectory, NavForesee demonstrates remarkable generalization by reliably extrapolating future geometric and semantic structures when performing more complex navigational behaviors such as turns. In the final scenario, the agent receives only a brief partial observation—a quick glance into a room—yet the model is able to produce a vivid and coherent internal imagination of the room's layout. This includes accurately inferring the relative shape and position of the bed, as well as estimating the depth

distribution across the room, thus indicating its capacity to reason about unseen spatial regions.

V. Conclusion

We proposed NavForesee, a vision–language navigation framework that unifies hierarchical language planning with dual-horizon predictive world modeling. By decomposing long instructions into sub-goals and anticipating both short-term dynamics and long-term milestones, NavForesee forms an implicit perception–planning and prediction–action loop. Experiments on R2R-CE and RxR-CE show strong performance—66.2% SR and 78.4% OSR on R2R-CE—comparable to state-of-the-art despite training only on public data. Qualitative results further reveal solid depth and semantics predictions that guide agent decisions in complex scenarios. These findings highlight the benefit of equipping embodied agents with foresight: by "foreseeing" future states, NavForesee effectively fuses language planning with spatiotemporal imagination to improve visual-language navigation.

References

[1] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," March 2023. [Online]. Available: https://lmsys.org/blog/2023-03-30-vicuna/

[2] S. Chen, X. Chen, C. Zhang, M. Li, G. Yu, H. Fei, H. Zhu, J. Fan, and T. Chen, "Ll3da: Visual interactive instruction tuning for omni-3d understanding, reasoning, and planning," 2023. [Online]. Available: https://arxiv.org/abs/2311.18651

[3] A.-M. Halacheva, J.-N. Zaech, X. Wang, D. P. Paudel, and L. V. Gool, "Gaussianvlm: Scene-centric 3d vision-language models using language-aligned gaussian splats for embodied reasoning and beyond," 2025. [Online]. Available: https://arxiv.org/abs/2507.00886

[4] H. Huang, Y. Chen, Z. Wang, R. Huang, R. Xu, T. Wang, L. Liu, X. Cheng, Y. Zhao, J. Pang, and Z. Zhao, "Chat-scene: Bridging 3d scene and large language models with object identifiers," 2024. [Online]. Available: https://arxiv.org/abs/2312.08168

[5] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," 2023. [Online]. Available: https://arxiv.org/abs/2304.08485

[6] Q. Liu, T. Huang, Z. Zhang, and H. Tang, "Nav-r1: Reasoning and navigation in embodied scenes," 2025. [Online]. Available: https://arxiv.org/abs/2509.10884

[7] H. Zhou, J. Yu, and W. Yang, "Dual memory units with uncertainty regulation for weakly supervised video anomaly detection," 2023. [Online]. Available: https://arxiv.org/abs/2302.05160

[8] Z. Wang, X. Li, J. Yang, Y. Liu, and S. Jiang, "Gridmm: Grid memory map for vision-and-language navigation," 2023. [Online]. Available: https://arxiv.org/abs/2307.12907

[9] Z. Wang, X. Li, J. Yang, Y. Liu, J. Hu, M. Jiang, and S. Jiang, "Lookahead exploration with neural radiance representation for continuous vision-language navigation," 2024. [Online]. Available: https://arxiv.org/abs/2404.01943

[10] X. Zhao, W. Cai, L. Tang, and T. Wang, "Imaginenav: Prompting vision-language models as embodied navigator through scene imagination," 2024. [Online]. Available: https://arxiv.org/abs/2410.09874

[11] X. Yao, J. Gao, and C. Xu, "Navmorph: A self-evolving world model for vision-and-language navigation in continuous environments," 2025. [Online]. Available: https://arxiv.org/abs/2506.23468

[12] B. Lin, Y. Nie, Z. Wei, J. Chen, S. Ma, J. Han, H. Xu, X. Chang, and X. Liang, "Navcot: Boosting llm-based vision-and-language navigation via learning disentangled reasoning," 2025. [Online]. Available: https://arxiv.org/abs/2403.07376

[13] S. Wang, Y. Wang, W. Li, X. Cai, Y. Wang, M. Chen, K. Wang, Z. Su, D. Li, and Z. Fan, "Aux-think: Exploring reasoning strategies for data-efficient vision-language navigation," 2025. [Online]. Available: https://arxiv.org/abs/2505.11886

[14] Q. Zhao, Y. Lu, M. J. Kim, Z. Fu, Z. Zhang, Y. Wu, Z. Li, Q. Ma, S. Han, C. Finn, A. Handa, M.-Y. Liu, D. Xiang, G. Wetzstein, and T.-Y. Lin, "Cot-vla: Visual chain-of-thought reasoning for vision-language-action models," 2025. [Online]. Available: https://arxiv.org/abs/2503.22020

[15] Y. Huang, J. Zhang, S. Zou, X. Liu, R. Hu, and K. Xu, "Ladi-wm: A latent diffusion-based world model for predictive manipulation," 2025. [Online]. Available: https://arxiv.org/abs/2505.11528

[16] J. Cen, C. Yu, H. Yuan, Y. Jiang, S. Huang, J. Guo, X. Li, Y. Song, H. Luo, F. Wang, D. Zhao, and H. Chen, "Worldvla: Towards autoregressive action world model," 2025. [Online]. Available: https://arxiv.org/abs/2506.21539

[17] E. Karypidis, I. Kakogeorgiou, S. Gidaris, and N. Komodakis, "Dino-foresight: Looking into the future with dino," 2024. [Online]. Available: https://arxiv.org/abs/2412.11673

[18] F. Baldassarre, M. Szafraniec, B. Terver, V. Khalidov, F. Massa, Y. LeCun, P. Labatut, M. Seitzer, and P. Bojanowski, "Back to the features: Dino as a foundation for video world models," 2025. [Online]. Available: https://arxiv.org/abs/2507.19468

[19] W. Zhang, H. Liu, Z. Qi, Y. Wang, X. Yu, J. Zhang, R. Dong, J. He, F. Lu, H. Wang, Z. Zhang, L. Yi, W. Zeng, and X. Jin, "Dreamvla: A vision-language-action model dreamed with comprehensive world knowledge," 2025. [Online]. Available: https://arxiv.org/abs/2507.04447

[20] J. Krantz, E. Wijmans, A. Majumdar, D. Batra, and S. Lee, "Beyond the nav-graph: Vision-and-language navigation in continuous environments," 2020. [Online]. Available: https://arxiv.org/abs/2004.02857

[21] A. Ku, P. Anderson, R. Patel, E. Ie, and J. Baldridge, "Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding," 2020. [Online]. Available: https://arxiv.org/abs/2010.07954

[22] Y. Hong, C. Rodriguez-Opazo, Y. Qi, Q. Wu, and S. Gould, "Language and visual entity relationship graph for agent navigation," 2020. [Online]. Available: https://arxiv.org/abs/2010.09304

[23] Z. Deng, K. Narasimhan, and O. Russakovsky, "Evolving graphical planner: Contextual global planning for vision-and-language navigation," 2020. [Online]. Available: https://arxiv.org/abs/2007.05655

[24] S. Chen, P.-L. Guhur, M. Tapaswi, C. Schmid, and I. Laptev, "Think global, act local: Dual-scale graph transformer for vision-and-language navigation," 2022. [Online]. Available: https://arxiv.org/abs/2202.11742

[25] M. Z. Irshad, N. C. Mithun, Z. Seymour, H.-P. Chiu, S. Samarasekera, and R. Kumar, "Sasra: Semantically-aware spatio-temporal reasoning agent for vision-and-language navigation in continuous environments," 2021. [Online]. Available: https://arxiv.org/abs/2108.11945

[26] G. Georgakis, K. Schmeckpeper, K. Wanchoo, S. Dan, E. Miltsakaki, D. Roth, and K. Daniilidis, "Cross-modal map learning for vision and language navigation," 2022. [Online]. Available: https://arxiv.org/abs/2203.05137

[27] P. Chen, D. Ji, K. Lin, R. Zeng, T. H. Li, M. Tan, and C. Gan, "Weakly-supervised multi-granularity map learning for vision-and-language navigation," 2022. [Online]. Available: https://arxiv.org/abs/2210.07506

[28] S. Zhang, Y. Qiao, Q. Wang, L. Guo, Z. Wei, and J. Liu, "Flexvln: Flexible adaptation for diverse vision-and-language navigation tasks," 2025. [Online]. Available: https://arxiv.org/abs/2503.13966

[29] Y. Long, X. Li, W. Cai, and H. Dong, "Discuss before moving: Visual language navigation via multi-expert discussions," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 17 380–17 387.

[30] P. Chen, X. Sun, H. Zhi, R. Zeng, T. H. Li, G. Liu, M. Tan, and C. Gan, "a^2nav: Action-aware zero-shot robot navigation by exploiting vision-and-language ability of foundation models," 2023. [Online]. Available: https://arxiv.org/abs/2308.07997

[31] Y. Long, W. Cai, H. Wang, G. Zhan, and H. Dong, "Instructnav: Zero-shot system for generic instruction navigation in unexplored environment," 2024. [Online]. Available: https://arxiv.org/abs/2406.04882

[32] Y. Wang, Y. Fang, T. Wang, Y. Feng, Y. Tan, S. Zhang, P. Liu, Y. Ji, and R. Xu, "Dreamnav: A trajectory-based imaginative framework for zero-shot vision-and-language navigation," 2025. [Online]. Available: https://arxiv.org/abs/2509.11197

[33] X. Xue, J. Hu, M. Luo, X. Shichao, J. Chen, Z. Xie, Q. Kuichen, G. Wei, M. Xu, and Z. Chu, "Omninav: A unified framework for prospective exploration and visual-language navigation," 2025. [Online]. Available: https://arxiv.org/abs/2509.25687

[34] Y. Liu, K. Zhang, Y. Li, Z. Yan, C. Gao, R. Chen, Z. Yuan, Y. Huang, H. Sun, J. Gao, L. He, and L. Sun, "Sora: A review on background, technology, limitations, and opportunities of large vision models," 2024. [Online]. Available: https://arxiv.org/abs/2402.17177

[35] A. Bar, G. Zhou, D. Tran, T. Darrell, and Y. LeCun, "Navigation world models," 2025. [Online]. Available: https://arxiv.org/abs/2412.03572

[36] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, "Qwen2.5-vl technical report," 2025. [Online]. Available: https://arxiv.org/abs/2502.13923

[37] J. Krantz, A. Gokaslan, D. Batra, S. Lee, and O. Maksymets, "Waypoint models for instruction-guided navigation in continuous environments," 2021. [Online]. Available: https://arxiv.org/abs/2110.02207

[38] Y. Hong, Z. Wang, Q. Wu, and S. Gould, "Bridging the gap between learning in discrete and continuous environments for vision-and-language navigation," 2022. [Online]. Available: https://arxiv.org/abs/2203.02764

[39] J. Krantz and S. Lee, "Sim-2-sim transfer for vision-and-language navigation in continuous environments," 2022. [Online]. Available: https://arxiv.org/abs/2204.09667

[40] H. Wang, W. Liang, L. V. Gool, and W. Wang, "Dreamwalker: Mental planning for continuous vision-language navigation," 2023. [Online]. Available: https://arxiv.org/abs/2308.07498

[41] D. An, Z. Wang, Y. Li, Y. Wang, Y. Hong, Y. Huang, L. Wang, and J. Shao, "1st place solutions for rxr-habitat vision-and-language navigation competition (cvpr 2022)," 2022. [Online]. Available: https://arxiv.org/abs/2206.11610

[42] D. An, H. Wang, W. Wang, Z. Wang, Y. Huang, K. He, and L. Wang, "Etpnav: Evolving topological planning for vision-language navigation in continuous environments," 2024. [Online]. Available: https://arxiv.org/abs/2304.03047

[43] K. Chen, J. K. Chen, J. Chuang, M. Vázquez, and S. Savarese, "Topological planning with transformers for vision-and-language navigation," 2020. [Online]. Available: https://arxiv.org/abs/2012.05292

[44] S. Raychaudhuri, S. Wani, S. Patel, U. Jain, and A. X. Chang, "Language-aligned waypoint (law) supervision for vision-and-language navigation in continuous environments," 2021. [Online]. Available: https://arxiv.org/abs/2109.15207

[45] G. Georgakis, K. Schmeckpeper, K. Wanchoo, S. Dan, E. Miltsakaki, D. Roth, and K. Daniilidis, "Cross-modal map learning for vision and language navigation," 2022. [Online]. Available: https://arxiv.org/abs/2203.05137

[46] P. Chen, D. Ji, K. Lin, R. Zeng, T. H. Li, M. Tan, and C. Gan, "Weakly-supervised multi-granularity map learning for vision-and-language navigation," 2022. [Online]. Available: https://arxiv.org/abs/2210.07506

[47] J. Chen, B. Lin, X. Liu, L. Ma, X. Liang, and K.-Y. K. Wong, "Affordances-oriented planning using foundation models for continuous vision-language navigation," 2024. [Online]. Available: https://arxiv.org/abs/2407.05890

[48] J. Krantz, E. Wijmans, A. Majumdar, D. Batra, and S. Lee, "Beyond the nav-graph: Vision-and-language navigation in continuous environments," 2020. [Online]. Available: https://arxiv.org/abs/2004.02857

[49] J. Zhang, K. Wang, R. Xu, G. Zhou, Y. Hong, X. Fang, Q. Wu, Z. Zhang, and H. Wang, "Navid: Video-based vlm plans the next step for vision-and-language navigation," 2024. [Online]. Available: https://arxiv.org/abs/2402.15852

[50] J. Zhang, K. Wang, S. Wang, M. Li, H. Liu, S. Wei, Z. Wang, Z. Zhang, and H. Wang, "Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks," 2025. [Online]. Available: https://arxiv.org/abs/2412.06224

[51] A.-C. Cheng, Y. Ji, Z. Yang, Z. Gongye, X. Zou, J. Kautz, E. Bıyık, H. Yin, S. Liu, and X. Wang, "Navila: Legged robot vision-language-action model for navigation," 2025. [Online]. Available: https://arxiv.org/abs/2412.04453

[52] M. Wei, C. Wan, X. Yu, T. Wang, Y. Yang, X. Mao, C. Zhu, W. Cai, H. Wang, Y. Chen, X. Liu, and J. Pang, "Streamvln: Streaming vision-and-language navigation via slowfast context modeling," 2025. [Online]. Available: https://arxiv.org/abs/2507.05240

[53] Z. Yu, Y. Long, Z. Yang, C. Zeng, H. Fan, J. Zhang, and H. Dong, "Correctnav: Self-correction flywheel empowers vision-language-action navigation model," 2025. [Online]. Available: https://arxiv.org/abs/2508.10416

Supplementary Material

I. IMPLEMENTATION DETAILS

A. Model Architecture

Base Model We employ Qwen2.5-VL-3B-Instruct [36] as the backbone of NavForesee. It adopts the Qwen2.5 LLM as its text decoder and integrates a vision encoder. The vision encoder utilizes a Vision Transformer (ViT) architecture to encode visual observations, while the text decoder is responsible for generating the hierarchical planning outputs and action trunk predictions. Detailed descriptions of Qwen2.5-VL can be found in [36]. For hierarchical planning, we directly use the original multimodal encoders and text decoder of Qwen2.5-VL without any modifications. For world model prediction and action policy learning, we introduce a position encoder to represent the agent's relative position and orientation derived from image observations. Lightweight decoders transform the dream query embeddings into environmental predictions (depth and semantics), while a simple MLP predicts action outputs (waypoints, orientation estimates, and arrival flags).

Dream Query Design Two sets of dream queries (short-term and long-term), along with an action query, are appended to the multimodal embeddings. Each set of dream queries contains depth and semantics subqueries, enabling dual-horizon prediction. We use DINOv2 and SAM features as semantic representations. Thus, there are six query subsets in total—depth, DINOv2, and SAM for both short-term and long-term horizons—with each subset consisting of 64 tokens. The action query consists of a single token dedicated to action prediction.

World Model Decoders We design task-specific lightweight world model decoders to transform dream embeddings into depth maps, semantic features, and actions. For depth and semantics predictions, we employ decoder architectures with identical design: dream embeddings and a set of learnable masks are processed by a 2-layer ViT-based decoder to produce predicted features. Additionally, we apply the decoder from VQ-VAE to render depth features into depth maps.

Action Prediction The action prediction module takes the action embedding produced by Qwen2.5-VL as input and generates predicted waypoints, orientation estimates, and arrival flags. First, a 2-layer transformer processes the action embedding to capture dependencies on the world model's dream embeddings. Then, the processed action embedding is passed to the action prediction head, which outputs the final navigation predictions, including waypoints, orientation estimates, and arrival flags. The action prediction head consists of a simple MLP with two linear layers and a ReLU activation in between.

B. Training Details

We interleave the VLM planning training data and world model training data to jointly train NavForesee. The training batch size is set to 4, and the number of image observations is flexible, up to a maximum length of 20. Depth and semantic features are precomputed and loaded during training. We use the AdamW optimizer with an initial learning rate of 1×10^{-5}. Depth and semantics predictions are weighted with $\alpha = 0.25$ and $\beta = 0.3$. The model is trained for a total of 3 epochs on 64 NVIDIA H20 GPUs, with ViT parameters frozen. The fixed short-term prediction horizon is set to 4, and the number of predicted waypoints is set to 5.

II. EXPERIMENTAL EVALUATIONS

A. Hierarchical Planning Evaluation

To evaluate the hierarchical planning capabilities of Nav-Foresee, we conduct experiments on the Val-Unseen split of the R2R-CE and RxR-CE datasets. An example is illustrated in Figure 6. We perform hierarchical planning for each step of an episode. NavForesee generates a navigation summary, plan, and actions strictly following the output format specified in the prompt template. Apart from the initial position, NavForesee consistently identifies milestones along the route, summarizes completed sub-instructions, and formulates the next sub-instruction in alignment with the overall instruction context. This demonstrates that Nav-Foresee effectively leverages its multimodal understanding capabilities to decompose complex navigation tasks into manageable sub-goals, thereby enabling more structured and efficient navigation. Notably, the hierarchical planning module is jointly trained with the world model prediction and action policy learning modules, indicating that NavForesee maintains strong language planning capabilities even when extended with additional functionalities. Furthermore, the hierarchical plans are precise and concise, which greatly benefits subsequent navigation decisions.

B. Short-term and Long-term Prediction Evaluation

Figure 7 illustrates the short-term and long-term depth predictions produced by our world model over a complete navigation episode. Short-term predictions forecast up to four future steps, whereas long-term predictions extrapolate over an adaptive horizon determined by progress towards the next milestone. Compared to short-term predictions, long-term depth predictions may be less accurate in capturing detailed depth at milestone locations, since milestone positions are unknown during inference. At the beginning of the episode, the long-term predictions effectively capture the scene when the agent ascends the stairs. As the agent approaches the first milestone (the doorway), the long-term predictions degrade slightly, likely due to the increased uncertainty of longer horizons and the absence of explicit milestone information. In such cases, long-term predictions tend to track short-term outputs, because long-term queries can attend to short-term queries. Nevertheless, the long-term predictions maintain the overall scene layout and depth distribution, providing

Fig. 6: Hierarchical planning examples generated by NavForesee for the instruction "Go up the stairs and straight forward the doorway. Turn right, move forward, and enter the doorway on the right. Move forward into the bedroom and stop in front of the toilet". From top to bottom: frames with timestamps, global navigation map, and navigation planning outputs. NavForesee accurately identifies milestones along the route, summarizes completed sub-instructions, and generates the next sub-instruction in accordance with the instruction context.
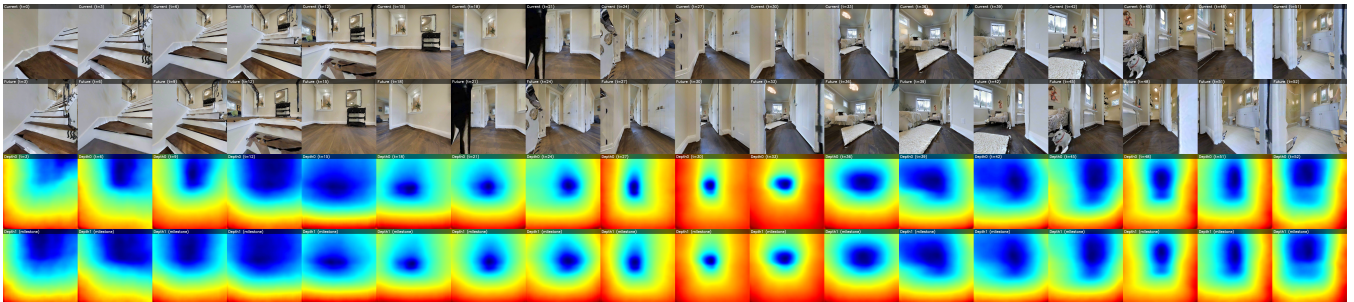

Fig. 7: Short-term and long-term depth predictions. From top to bottom: frames with timestamps, future ground truth frames with timestamps, short-term depth predictions for future frames, and long-term depth predictions for milestones. Instruction: "Up the stairs. Turn to the left and enter the second open door on the left. Walk towards the foot of the bed. Turn right and enter the open door to the bathroom."

valuable guidance for strategic navigation. This demonstrates that NavForesee's world model effectively anticipates environmental changes over both short and long horizons, enhancing the agent's planning and action capabilities in complex scenarios.

C. Ablation Study on Depth and Semantics Predictions

We conduct ablation studies to evaluate the individual contributions of depth and semantics predictions in the world model. As shown in Table III, removing either depth or semantics predictions results in a clear performance drop. The full NavForesee model, which integrates both depth and

TABLE III: Performance comparison between depth prediction and semantics prediction

Index	Depth	Semantics	SR ↑	OSR↑	NE ↓	SPL↑
1	✓	✓	66.2%	78.4%	3.94	59.7%
2	✗	✓	61.8%	76.7%	4.37	54.9%
3	✓	✗	60.0%	76.2%	4.59	52.9%

semantics predictions, achieves the highest SR (66.2%), OSR (78.4%), lowest NE (3.94), and best SPL (59.7%), validating the benefit of their combination. Without depth prediction, the SR drops to 61.8% and SPL decreases by 4.8 points,

highlighting the importance of depth information for spatial reasoning and obstacle avoidance. Disabling semantics predictions leads to an even larger SR reduction (60.0%) and higher NE, underscoring the critical role of semantic features in recognizing landmarks and guiding navigation. These findings confirm that both depth and semantics predictions are essential for accurate and efficient navigation.