Supplementary Material

A. Implementation Details

A.1. Model Architecture

Base Model We employ Qwen2.5-VL-3B-Instruct [3] 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 [3]. 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.

A.2. 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 5, same as the number of predicted waypoints.

B. Experimental Evaluations

B.1. Hierarchical Planning Evaluation

To evaluate the hierarchical planning capabilities of NavForesee, 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 NavForesee 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.2. 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 five future steps, whereas long-term predictions extrapolate over an adaptive horizon determined by progress to-

Figure 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.

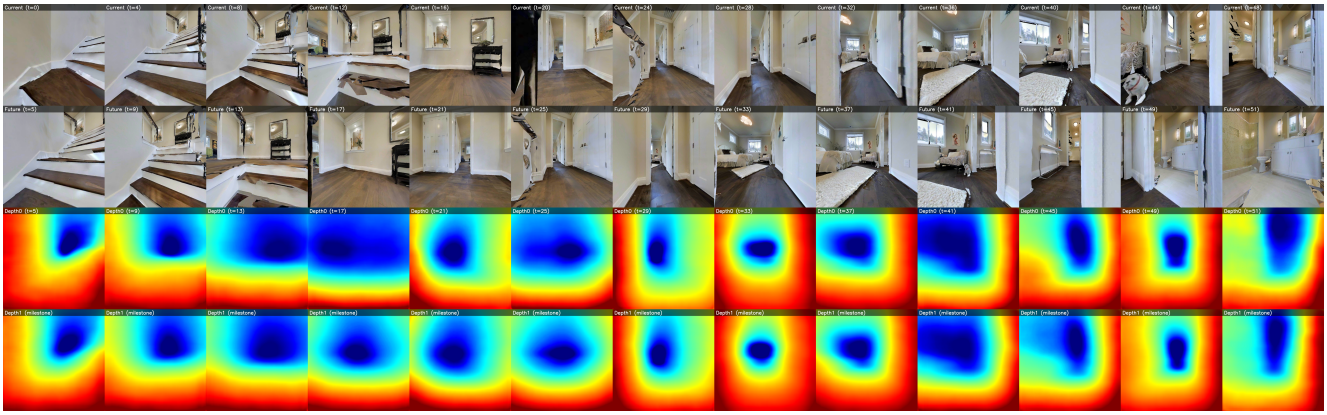

Figure 7. Short-term and long-term depth predictions. From top to bottom: frames with timestamps, future ground truth frames with timestamps, short-term depth predictions for future frames, and long-term depth predictions for milestones. Instruction: "Up the stairs. Turn to the left and enter the second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

wards 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 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.

B.3. Quantitative prediction Quality

We evaluate prediction quality on R2R Val-Unseen (input at T, target at $T+5$). To verify that the model learns temporal dynamics rather than copying inputs, we compare against Identity Baselines that use GT at $T+i$ as predictions for $T+5$. Table 4 shows: 1) **Depth:** Our model outperforms baselines up to $T+2$, capturing short-term geometric dynamics. 2) **Semantics (DINO):** Our model achieves a CosSim of **0.62**, significantly outperforming baselines up to $T+3$.

This indicates strong capability in predicting high-level semantic evolution further into the future.

Table 4. Prediction Quality. Model (Pred) vs. Identity Baselines (copying GT at $T+i$) as predictions for $T+5$.

Type	Metric	**Model Pred**	Identity Baselines (GT vs T+5)				
			T	$T+1$	$T+2$	$T+3$	$T+4$
Depth	SSIM ↑	**0.81**	0.79	<u>0.80</u>	0.82	0.84	0.88
	PSNR ↑	**16.39**	15.03	15.39	<u>15.98</u>	17.00	19.04
	LPIPS ↓	**0.28**	0.32	0.31	<u>0.28</u>	0.25	0.18
Semantic(DINO)	CosSim ↑	**0.62**	0.47	0.50	0.53	<u>0.58</u>	0.67

B.4. 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 5, removing either depth or semantics predictions results in a clear performance drop. The full NavForesee model, which integrates both depth and 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.

Comparing Table 4 with Table 5, we conclude that high-quality semantic foresight is more critical for navigation performance than short-term geometric precision, as it aids in long-term goal anchoring.

Table 5. 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