NavWM: A Unified Navigation World Model for Foresight-Driven Planning

Yanghong Mei[1,3], Longteng Guo[1], Ming-Ming Yu[2], Guiyu Zhao[1,3], Xingjian He[1], and Jing Liu[1,3⋆]

[1] Institute of Automation, Chinese Academy of Sciences
{meiyanghong2024, longteng.guo}@ia.ac.cn
[2] Beihang University
[3] School of Artificial Intelligence, University of Chinese Academy of Sciences

Abstract. Conventional visual navigation policies often struggle with myopic decision-making and mode collapse in complex environments. While world models offer a promising alternative, existing paradigms typically isolate perception, generation, and control, failing to capture their shared spatio-temporal dynamics. In this paper, we propose NavWM, a unified navigation world model that seamlessly integrates latent world reasoning, multimodal action prediction, and controllable visual generation. At its core, NavWM leverages latent world tokens to distill geometric and semantic priors, endowing the agent with robust structural understanding. To overcome the limitations of deterministic policies, we introduce an anchor-based multimodal trajectory forecasting framework that generates a diverse action space. This inherent diversity explicitly empowers the generative world model to act as a robust closed-loop planner, utilizing visual foresight to evaluate and select the optimal path. Extensive experiments across diverse robotics datasets demonstrate that NavWM significantly advances the state-of-the-art, delivering remarkable improvements in both high-fidelity future state generation and zero-shot navigation success.

Keywords: World model · Visual navigation · Visual foresight planning

1 Introduction

Visual navigation [35,41,42,45] stands as a cornerstone of embodied intelligence, underpinning a wide array of real-world applications such as autonomous driving, automated delivery, and robotic services. It requires an agent to rely solely on camera observations to safely maneuver through arbitrary, unstructured environments, actively avoiding pedestrians and obstacles to reach a designated goal. Achieving this complex task demands far more than static scene comprehension; it fundamentally requires the agent to anticipate and reason about future world states while in motion [4].

⋆ Corresponding author: jliu@nlpr.ia.ac.cn

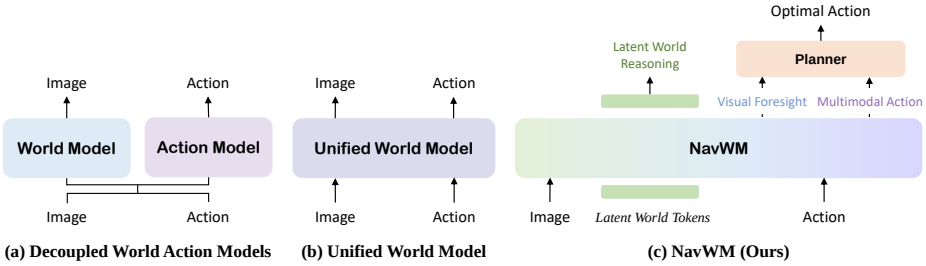

Fig. 1: Comparison of world model paradigms. (a) Decoupled world and action models. (b) A unified world model with independent outputs. (c) Our NavWM integrates latent world reasoning and multimodal actions, enabling the world model to act as a planner via visual foresight.

Conventional visual navigation policies typically learn a direct visuo-motor mapping from visual observations to action sequences, trained on large-scale web videos or embodied trajectories [35, 38, 41, 42]. However, without explicit foresight, these purely reactive policies often suffer from myopic decision-making and limited generalization to unseen environments [21, 44].

To address these limitations, recent works have begun incorporating world models for navigation. While promising, existing paradigms still face three key challenges. First, modular pipelines (e.g., [4, 32]) typically train the navigation policy and the world model separately (Fig. 1(a)). This decoupled design fails to leverage their shared spatiotemporal dynamics, resulting in sub-optimal representations. Second, although a few recent works attempt to unify the navigation policy and the world model [10,15] (Fig. 1(b)), their latent representations often lack explicit scene abstractions. Without structured cues about environmental geometry or semantics, the model must infer spatial regularities directly from raw features, which can hinder stable long-horizon prediction and planning. Finally, real-world navigation exhibits inherently multi-modal futures, where multiple feasible trajectories may lead toward the same goal. However, many existing methods model action prediction as a single trajectory, collapsing the multi-modal action space into one mode. Consequently, the predicted actions become overly concentrated, making the agent prone to sub-optimal decisions or stalling in local minima.

To bridge this critical gap, we present a **Unified Navigation World Model** (**NavWM**), that seamlessly integrates latent world reasoning, controllable visual generation, and multimodal action prediction [4]. Unlike prior disjointed paradigms, NavWM is built upon the core insight that perception, generation, and control fundamentally share a reliance on modeling underlying environmental structures and spatio-temporal dynamics. By unifying these heterogeneous tasks within a shared Bidirectional Mamba backbone, our framework explicitly exploits their inherent synergy during the training phase to foster mutually re-

[4] "Multimodal" denotes predicting multiple spatial trajectory modes [12, 23, 55].

inforcing representations. This unified learning process, in turn, empowers the inference phase by supporting **closed-loop planning** via **visual foresight**. Consequently, the agent can simulate future visual observations for diverse trajectory candidates, enabling it to evaluate goal alignment and filter the optimal path within a simulated "multiverse" of possibilities.

To encourage structured scene representations, we introduce dedicated latent world tokens that capture geometric and semantic cues of the environment. These tokens are supervised using depth and semantic signals derived from visual foundation models, guiding the model to learn spatially consistent representations. Such structured cues provide physically grounded context for long-horizon prediction and improve the stability of future observation synthesis.

In addition, real-world navigation often admits multiple feasible future trajectories toward the same goal. However, many prior methods rely on predicting a single trajectory [15,41,42], which restricts exploration and can trap the agent in sub-optimal behaviors [12]. To address this limitation, NavWM adopts an anchor-based multimodal trajectory prediction framework that generates diverse yet physically plausible motion hypotheses. This multimodal proposal mechanism expands the candidate action space, allowing the world model to evaluate alternative futures through visual foresight and select more reliable navigation decisions.

Extensive experimental results demonstrate that NavWM achieves outstanding performance across a diverse range of robotics datasets (Go Stanford [27], SCAND [29], RECON [40], HuRoN [28], and Tartan Drive [47]). In offline evaluations for single-step prediction, NavWM significantly improves image generation quality, evidenced by a notable increase in PSNR from 14.17 to 17.34 (Tab. 1). In complex, long-horizon navigation tasks, NavWM demonstrates formidable navigation performance alongside robust scene reconstruction capabilities. Notably, it elevates the navigation success rate from 66% to 72% and achieves a remarkable 44% success rate even during zero-shot inference in unseen environments (Tab. 2, Fig. 3). Ablation experiments demonstrate that the joint training of reasoning, control, and world modeling yields mutually reinforcing effects, fundamentally driven by their shared reliance on underlying spatio-temporal dynamics. Moreover, our multi-modal action prediction approach elevates trajectory diversity and strengthens the model's exploration capacity, inherently supporting the use of the generative world model as a high-level supervisory planner.

In summary, the main contributions of our work are threefold:

- We propose NavWM, a unified navigation world model that jointly learns structured perception, controllable world modeling, and multimodal action prediction within a shared architecture.
- We introduce a world-model-based visual foresight planning mechanism that evaluates candidate trajectories by simulating future observations, enabling goal-aligned closed-loop navigation.
- Extensive experiments demonstrate that NavWM achieves superior performance in both visual state simulation and complex navigation, significantly boosting generation quality and success rates across diverse environments.

2 Related Work

2.1 Visual Navigation

Visual navigation stands as a fundamental challenge in robotics, necessitating robust capabilities in scene dynamics modeling and strategic planning [9, 14, 18, 20, 52, 53]. Early works [41, 42] learn a deterministic end-to-end mappings from current states to actions from real-world navigation data, whereas Sridhar et al. [45] adopt diffusion policy to represent stochastic action distributions. Recent approaches [35, 38] learn generalizable visual navigation models from large-scale web videos. Nevertheless, they suffer from deterministic modeling or overly peaked distributions, leading to mode collapse that severely hinders environmental exploration [21]. Moreover, they focus predominantly on reactive low-level execution, devoid of hierarchical decision-making capabilities. Conversely, we present a unified framework designed to forecast multimodal trajectories alongside future states. By deploying the world model as a high-level planner, our approach enables the agent to execute more effective and far-sighted exploration.

2.2 World Models for Navigation

World models [24] capture environmental dynamics and temporal evolution by predicting future states. The complex spatio-temporal dynamics of navigation make it an ideal downstream testbed for world models [15, 18]. Early diffusion-based works [1, 3, 6, 7] have demonstrated remarkable capabilities in high-fidelity simulation. Other works actively exploring viewpoint-conditioned future state generation [17, 25, 26]. Recent works have explored leveraging world models as high-level planners, integrating them into navigation rollouts to explicitly guide action policies [4, 10, 51]. Nevertheless, existing approaches often fail to exploit the underlying capabilities inherently shared by both navigation policies and generative world models: namely, deep scene comprehension and spatio-temporal dynamics modeling. To bridge this gap, this paper proposes a unified framework that explicitly incorporates a latent world reasoning task. By employing a joint training strategy, we enable scene perception, the navigation policy, and the world model to mutually reinforce one another during the learning process.

3 Methodology

3.1 Problem Formulation

Given the current world state, historical observation context, and navigation targets (e.g., Image Goal), NavWM is designed to learn the underlying environmental dynamics. It concurrently predicts the optimal action and the subsequent state transition. Notably, this predictive capability extends to target-agnostic scenarios, allowing the model to imagine future spatiotemporal evolutions under different hypothetical actions. Formally, given the current egocentric RGB

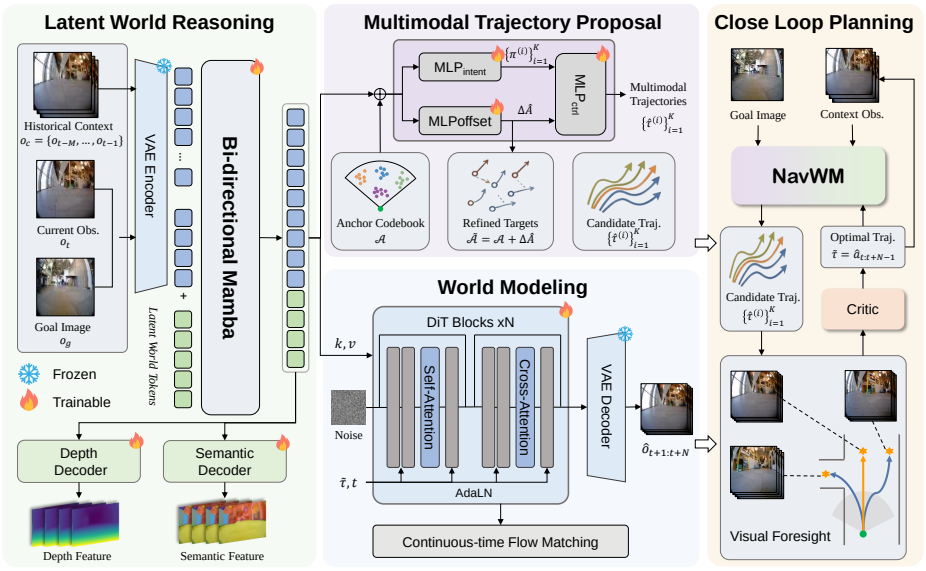

Fig. 2: Overview of the NavWM framework. The unified navigation world model learns latent scene abstractions from historical observations and goals via latent world reasoning, supervised by depth and semantic predictions. It then proposes multimodal trajectory hypotheses and predicts future observations through world modeling, enabling closed-loop navigation planing via visual foresight.

observation $o_t \in \mathbb{R}^{H\times W\times 3}$, a historical context $\mathcal{O}_c = \{o_{t-M}, \ldots, o_{t-1}\}$ within a temporal window M, and a goal image o_g, our objective is to learn a unified model F_θ that approximates the joint conditional distribution of multimodal trajectories $\mathcal{T} = \{\hat{a}^{(k)}_{t:t+N-1}\}_{k=1}^{K}$ and the corresponding future observations $\mathcal{O} = \{\hat{o}^{(k)}_{t+1:t+N}\}_{k=1}^{K}$ over a prediction horizon N:

$$(\mathcal{T}, \mathcal{O}) \sim F_\theta(\cdot|\mathcal{O}_c, o_t, o_g), \tag{1}$$

where the action $a = (\mu, \phi)$ consists of a translational component $\mu \in \mathbb{R}^2$ and a rotational component $\phi \in \mathbb{R}$, defined in the agent's egocentric coordinate frame.

3.2 NavWM Overview

As illustrated in Figure 2, NavWM employs a State Space Model (SSM) as the backbone, which branches into three specialized heads dedicated to latent world reasoning, action prediction and future synthesis, respectively. The historical frames $\mathcal{O}_c$, current observation o_t, and goal o_g are first projected into a compact latent space via a pretrained VAE encoder [5]. Subsequently, these latent embeddings are prepended with a sequence of learnable *Latent World Tokens* and fed into a Bidirectional Mamba backbone [22] for spatiotemporal encoding. We employ an Attentional Pooling Layer to compress the encoded historical tokens

into a fixed-length representation. This simple yet effective strategy significantly enhances computational efficiency for long-horizon inference.

3.3 Latent World Reasoning

To internalize the 3D geometry and semantics essential for foresighted navigation, NavWM introduces a **Latent World Reasoning** mechanism that constructs explicit scene abstractions to ground subsequent planning and future imagination. Specifically, we introduce learnable *Latent World Tokens* to assimilate these environmental priors. For robust supervision, we distill knowledge from visual foundation models (Depth Anything V2 [50] for geometry and SAM [31] for semantics) to generate dense future pseudo-labels, $\mathcal{D}_{gt} = \{d_n\}$ and $\mathcal{S}_{gt} = \{s_n\}$, over the horizon $n \in \{t+1, \ldots, t+N\}$. Following spatiotemporal encoding, these Latent World Tokens are spatially upsampled via a CNN decoder to forecast the scene abstractions, yielding predictions $\hat{\mathcal{D}} = \{\hat{d}_n\}$ and $\hat{\mathcal{S}} = \{\hat{s}_n\}$.

To mitigate the severe scale ambiguity inherent in diverse indoor and outdoor navigation environments, we supervise the geometric predictions using a Scale-Invariant Loss [16], defined as:

$$\mathcal{L}_{si} = \frac{1}{P}\sum_{p=1}^{P} g_p^2 - \frac{\lambda}{P^2}\left(\sum_{p=1}^{P} g_p\right)^2, \tag{2}$$

where $g_p = \log \hat{d}_p - \log d_p$, p indexes the valid pixels within a single frame, and P denotes the total number of pixels. Concurrently, the semantic feature alignment is supervised via a standard Mean Squared Error (MSE) loss. The final Latent World Reasoning objective is formulated as the weighted sum of these two components, averaged over the temporal prediction horizon:

$$\mathcal{L}_{reason} = \lambda_{depth}\mathcal{L}_{si}(\mathcal{D}_{gt}, \hat{\mathcal{D}}) + \lambda_{sem}\mathcal{L}_{mse}(\mathcal{S}_{gt}, \hat{\mathcal{S}}). \tag{3}$$

By explicitly optimizing this auxiliary objective, the model effectively internalizes the underlying geometric and semantic priors of the scene.

3.4 Multimodal Trajectory Proposal

Reliable foresight planning requires exploring multiple plausible future actions rather than committing to a single deterministic trajectory, which is prone to sub-optimal local minima. To provide the world model planner with a diverse and expressive action space, NavWM adopts a multimodal trajectory proposal module that predicts a set of physically plausible future motion hypotheses.

Following prior work on trajectory prediction [23], we model future uncertainty as arising from two complementary factors: *intent uncertainty*, corresponding to the high-level navigation target, and *control uncertainty*, reflecting low-level motion variations toward that target. Accordingly, the trajectory proposal process is decomposed into two stages: anchor-based target prediction and target-conditioned trajectory generation.

Anchor-based Target Prediction. We first predict a set of candidate navigation targets represented as relative 2D positions in the agent's egocentric coordinate frame. Specifically, we construct an anchor codebook $\mathcal{A} \in \mathbb{R}^{K\times 2}$ that captures the most representative target locations over horizon N. The anchors are obtained offline by applying K-Means clustering followed by Farthest Point Sampling on ground-truth target coordinates. Given the scene representation $\mathcal{H}$ produced by the world model encoder, we estimate the likelihood of each anchor and its corresponding spatial refinement. Concretely, the anchor coordinates are concatenated with the scene features and passed through two lightweight MLP heads: an intent head that predicts the anchor probabilities $\hat{\boldsymbol{\pi}} \in \mathbb{R}^K$, and a control head that regresses continuous spatial offsets $\Delta\hat{\mathbf{A}} \in \mathbb{R}^{K\times 2}$:

$$\hat{\boldsymbol{\pi}} = \mathrm{Softmax}(\mathrm{MLP}_{intent}([\mathcal{H} \parallel \mathcal{A}])), \quad \Delta\hat{\mathbf{A}} = \mathrm{MLP}_{control}([\mathcal{H} \parallel \mathcal{A}]). \tag{4}$$

The refined targets are therefore expressed as $\widetilde{\mathcal{A}} = \mathcal{A} + \Delta\hat{\mathbf{A}}$. During training, we identify the anchor k^* that best matches the ground-truth and optimize a joint objective comprising an intent classification loss and an offset regression loss:

$$\mathcal{L}_{anchor} = -\log(\hat{\boldsymbol{\pi}}_{k^*}) + \mathcal{L}_{huber}(\Delta\hat{x}_{k^*}, \Delta x_{gt}) + \mathcal{L}_{huber}(\Delta\hat{y}_{k^*}, \Delta y_{gt}), \tag{5}$$

where $(\Delta x_{gt}, \Delta y_{gt})$ is the ground-truth offset relative to the matched anchor.

Target-Conditioned Trajectory Prediction. Given the refined targets $\widetilde{\mathcal{A}}$, we further predict a full sequence of future waypoints for each candidate target. Instead of autoregressively generating the trajectory step-by-step, which incurs significant computational overhead, we adopt a direct prediction strategy that assumes conditional independence across future timesteps [8, 13, 37, 55]. Specifically, the scene features $\mathcal{H}$ and refined targets $\widetilde{\mathcal{A}}$ are fed into an MLP to directly regress the waypoint sequence $\hat{\mathcal{T}}_k$ associated with each candidate target. To avoid the non-differentiable *argmax* operation during training, supervision is applied only to the trajectory corresponding to the matched anchor k^*. The trajectory loss is defined as:

$$\mathcal{L}_{traj} = \mathcal{L}_{huber}(\hat{\mathcal{T}}_{k^*}, \mathcal{T}_{gt}). \tag{6}$$

The final action proposal objective combines the anchor prediction and trajectory regression losses:

$$\mathcal{L}_{action} = \mathcal{L}_{anchor} + \lambda_{traj}\mathcal{L}_{traj}, \tag{7}$$

where λ_{traj} balances the two components. The resulting module produces a diverse set of trajectory candidates that serve as action hypotheses for subsequent visual foresight evaluation.

3.5 World Modeling for Visual Foresight

To enable foresight-based planning, NavWM learns a world model that predicts future visual observations conditioned on candidate trajectories. Given the scene

Table 1: Offline evaluation. Average performance of action prediction and visual generation across four primary datasets. NavWM achieves state-of-the-art performance as both a navigation policy and a world model.

Method	Navigation				Visualization			
	ATE ↓	RPE ↓	AOE ↓	MAOE ↓	SSIM ↑	PSNR ↑	LPIPS ↓	DreamSim ↓
GNM [41]	1.519	0.622	18.165	23.961	-	-	-	-
ViNT [42]	1.424	0.617	17.074	23.831	-	-	-	-
NoMaD [45]	1.070	0.376	12.463	17.645	-	-	-	-
Diamond [2]	-	-	-	-	0.296	8.857	0.425	0.161
Anloe-7B [11]	1.735	0.749	18.288	24.878	0.288	9.176	0.513	0.163
NWM [4]	0.642	0.211	10.496	16.534	0.422	14.343	0.295	0.091
UniWM [15]	0.302	0.116	9.468	13.221	0.435	14.172	0.282	0.102
NavWM (Ours)	**0.207**	**0.066**	**8.152**	**12.855**	**0.507**	**17.340**	**0.243**	**0.084**

representation and predicted motion, the model simulates how the environment would evolve under different action hypotheses. As illustrated in Figure 2, the input visual latents and predicted actions are processed by a stack of Conditional Diffusion Transformer (CDiT) blocks [4], which provide a more efficient alternative to directly applying a vanilla Diffusion Transformer (DiT) [39]. To explicitly encode camera motion, the predicted actions are first transformed into dense optical flow fields [43], denoted as $\mathcal{F}$. The motion representation is spatially pooled and fused with the diffusion timestep embedding, forming a joint conditioning signal that modulates intermediate activations through adaptive layer normalization (AdaLN) [49].

Meanwhile, the scene features $\mathcal{H}$ produced by the unified world model backbone are injected via cross-attention, providing contextual guidance for future synthesis. The resulting latent tokens are then unpatchified into spatial latent maps and decoded by the VAE decoder to generate realistic future observations.

We optimize the NavWM world model via continuous-time Flow Matching [33,34], formulating future synthesis as a deterministic mapping from a Gaussian prior to the target latent distribution. Specifically, given a ground-truth latent z_1 and a noise sample $z_0 \sim \mathcal{N}(0, \mathbf{I})$, we define a probability path via linear interpolation $z_\tau = \tau z_1 + (1-\tau) z_0$, where $\tau \sim \mathcal{U}(0,1)$. The network v_θ is trained to regress the constant vector field $u_\tau = z_1 - z_0$ that drives the flow from noise to data, supervised by the following objective:

$$\mathcal{L}_{visual} = \mathbb{E}_{\tau\sim\mathcal{U}(0,1), z_0\sim\mathcal{N}, z_1\sim q}\left[\left\|v_\theta(z_\tau, \tau, \mathcal{F}, \mathcal{H}) - (z_1 - z_0)\right\|_2^2\right], \tag{8}$$

This formulation ensures straighter synthesis trajectories and highly efficient sampling during inference.

3.6 Two-Stage Joint Training

NavWM is trained with a two-stage optimization strategy to leverage the complementary strengths of perception, control, and visual foresight while mitigating the train–inference discrepancy. In the first stage, all components are jointly optimized using teacher forcing. Ground-truth trajectories are provided to the

CDiT world modeling module, allowing the perception, action prediction, and visual synthesis branches to learn mutually consistent spatiotemporal representations. The second stage addresses the exposure bias caused by the mismatch between training and inference. Specifically, we freeze the SSM backbone and trajectory proposal heads, and fine-tune the CDiT conditioned on trajectories sampled from the predicted distribution rather than the ground truth. This stage improves the robustness of the world model under imperfect action predictions.

To balance the heterogeneous gradients arising from different learning objectives, we adopt the uncertainty-weighted multi-task formulation proposed in [30]. The overall training objective is defined as

$$\mathcal{L} = \sum_{k \in \mathcal{T}} \left(\frac{1}{2} \exp(-s_k) \mathcal{L}_k + \frac{1}{2} s_k \right), \tag{9}$$

where $\mathcal{T} = \{\text{visual}, \text{action}, \text{reason}\}$ denotes the set of training tasks, and s_k represents the learnable log-variance associated with task k. This adaptive weighting mechanism automatically balances the relative contributions of different objectives and eliminates the need for manual loss weighting.

3.7 Closed-Loop Planning with Visual Foresight

Selecting the optimal action from diverse trajectory candidates traditionally relies on hand-crafted heuristics. Instead, NavWM employs its world model as a visual supervisor to evaluate each path directly. By simulating future visual observations, we accurately measure goal alignment, enabling the agent to pinpoint and execute the optimal path. Specifically, from K predicted trajectories, C diverse candidates are first selected via Non-Maximum Suppression (NMS) and queried against the world model to roll out future observations $\hat{o}$ for each waypoint. To evaluate the navigational quality of these simulated states, we introduce a scoring function $\mathcal{R}(\hat{o}, o_g)$:

$$\mathcal{R}(\hat{o}, o_g) = \exp\left(-\left(\alpha \cdot \text{LPIPS}(\hat{o}, o_g) + \beta \cdot \text{DreamSim}(\hat{o}, o_g)\right)\right), \tag{10}$$

where α and β balance the low-level perceptual details (measured by LPIPS [54]) and the high-level semantic layout (measured by DreamSim [19]). A higher value indicates a more accurate alignment with the goal observation o_g. Subsequently, we identify the simulated observation achieving the highest score and execute the initial action from its parent trajectory to advance the navigation.

4 Experiments and Results

4.1 Experimental Setting

Datasets. We train NavWM using a mixture of robotics datasets that cover both indoor and outdoor environments (Go Stanford [27], SCAND [29], RECON [40], and HuRoN [28]).Containing precise real-world robot location and

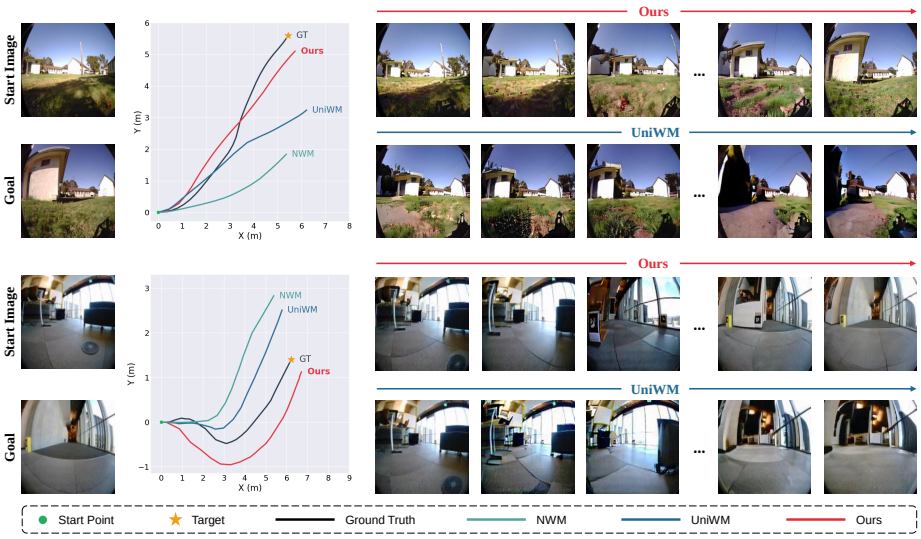

Fig. 3: Qualitative comparison across indoor and outdoor environments. Left: Trajectory plots illustrating the alignment of predicted paths from different methods with the Ground Truth (GT). Right: Visual generation results of camera viewpoint control during navigation.

rotation, they can be easily converted into navigation actions (i.e., relative camera pose transformations). Paired with corresponding RGB observations, they provide a suitable resource for training navigation policies and world models. Following prior works [4,15,45], we discard trajectories shorter than three steps and backward movements, partitioning the remaining data into semantically coherent segments based on the continuous video stream. Additionally, we incorporate the TartanDrive [47] as a zero-shot testbed for unseen environments.

Evaluation Metrics. For the offline evaluation of the navigation policy, we employ Absolute Trajectory Error (ATE) and Relative Pose Error (RPE) to assess pose consistency [46], while utilizing Average Orientation Error (AOE) and Maximum Average Orientation Error (MAOE) to quantify overall trajectory similarity [35]. Regarding the generative performance of the world model, pixel-level synthesis quality and structural similarity are measured via PSNR and SSIM [48], whereas LPIPS [54] and DreamSim [19] are adopted to evaluate perceptual similarity in the latent space.

Implementation Details. In the default setup, the model operates at a resolution of 256×256, receiving a context of 4 past frames to forecast the waypoints and visual states of the next 4 frames. We optimize 1.5B-parameter NavWM of using the AdamW optimizer [36], employing a learning rate schedule that warms up from 5×10^{-5} to a peak of 1×10^{-4} over the first 500 steps, followed by a

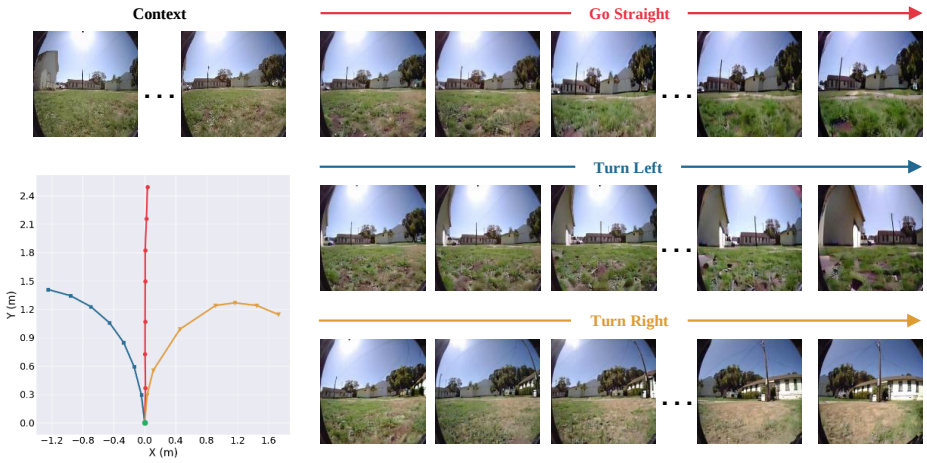

Fig. 4: Predicting diverse futures. NavWM synthesizes multiple possible visual futures conditioned on varying trajectory inputs without goal conditioning. The green dot marks the starting position, while the red, blue, and yellow curves correspond to straight, left, and right trajectories, respectively.

cosine decay to a minimum of 1×10^{-5}. Using 8 NVIDIA A100 GPUs with a global batch size of 64, we employ teacher forcing in the first stage to train the future synthesis head with ground-truth trajectories for 100,000 steps, applying a goal mask with a 15% probability. In the second stage, the future synthesis head is exclusively fine-tuned for 50,000 steps, conditioned instead on the predicted trajectories.

4.2 Offline Evaluation

Baselines. For trajectory prediction, we select conventional navigation policies ViNT [42] and NoMaD [45] as our baselines. ViNT employs a Transformer architecture to learn an end-to-end visuo-motor mapping, whereas NoMaD utilizes a diffusion head to output a non-deterministic policy. Regarding visualization performance, we benchmark against Diamond [2] alongside state-of-the-art methods NWM [4] and UniWM [15]. NWM is a DiT-based world model; to evaluate its navigation metrics, we leverage its Standalone Planning capability to guide NoMaD. UniWM, fine-tuned from the unified multimodal model Anole [11], serves as an autoregressive model tailored for navigation and visualization.

Comparison with SOTA Methods. Tab. 1 details the offline evaluation. To assess the fundamental navigation performance independently of the generative planner, we perform random sampling across the multimodal trajectory predictions to evaluate their baseline accuracy. As shown, NavWM consistently surpasses both the NWM-guided NoMaD and the current state-of-the-art method,

Table 2: Navigation rollout results. This table reports metrics for Image Goal Navigation. NavWM achieves better performance in rollout success rate (SR) and trajectory accuracy in both Seen and Unseen environments.

Method	Seen			Unseen		
	SR ↑	**ATE ↓**	**RPE ↓**	**SR ↑**	**ATE ↓**	**RPE ↓**
ViNT [42]	0.24	1.948	0.817	0.11	2.423	0.812
NoMaD [45]	0.29	1.711	0.601	0.16	2.242	0.789
ANOLE-7B [11]	0.16	2.237	0.911	0.11	2.334	0.932
NWM [4]	0.58	1.141	0.360	0.23	1.712	0.647
UniWM [15]	0.66	0.645	0.229	0.36	1.122	0.375
NavWM (Ours)	**0.72**	**0.381**	**0.161**	**0.44**	**0.765**	**0.260**

UniWM, establishing superior trajectory prediction accuracy (ATE: 0.302 → 0.207, RPE: 0.116 → 0.066). Moreover, our visual generation quality outperforms both the diffusion-based NWM and the autoregressive UniWM in terms of structural and perceptual similarity (PSNR: 14.343 → 17.340, LPIPS: 0.282 → 0.243), emerging as the new state-of-the-art.

4.3 Navigation Rollout

In this phase, we evaluate the model's performance on the Image Goal Navigation task. This requires the model to navigate to the goal image's location relying solely on the initial observation, achieved through an interleaved process of action prediction and visual forecasting [15]. We segment the dataset trajectories based on semantic coherence, yielding an average segment length of 43 steps. Navigation is deemed successful when the agent arrives within a 0.5m radius of the goal, based on which we calculate the Success Rate (SR). Additionally, we measure trajectory consistency using ATE and RPE.

Navigation Performance. As shown in Tab. 2, NavWM achieves superior performance in navigation tasks across both seen and unseen environments. Specifically, when performing random sampling on multimodal trajectory outputs, NavWM attains a navigation success rate of 0.63 in Seen environments, establishing a strong baseline. Crucially, when leveraging the world model as a planner to supervise action generation, NavWM achieves state-of-the-art results in both seen (SR: 0.74) and unseen (SR: 0.44) settings, while also delivering optimal trajectory quality.

Qualitative Analysis. Fig. 3 visualizes the navigation trajectories and corresponding visual generation results for different methods. As can be observed, NavWM achieves superior alignment with the ground truth trajectories. Furthermore, Fig. 4 presents the simulated rollouts of diverse potential futures generated by NavWM. By leveraging these diverse trajectory outputs in synergy with a

Table 3: Module-level ablation study. Effects of latent World Reasoning (WR), Action Prediction (AP), and World Model (WM). Joint training of the three branches significantly improves model performance.

WR	AP	WM	Navigation		Visualization			
			ATE ↓	RPE ↓	SSIM ↑	PSNR ↑	LPIPS ↓	DreamSim ↓
	✓		0.513	0.187	-	-	-	
	✓	✓	0.338	0.137	0.414	14.286	0.249	0.087
✓		✓	-	-	0.482	16.627	0.293	0.112
✓	✓	✓	**0.254**	**0.113**	**0.516**	**17.622**	**0.240**	**0.086**

world model planner, NavWM can correct errors in real-time, whereas other methods suffer severely from cumulative drift. Finally, visual inspection reveals that while competing methods tend to lose track of the target over long-horizon trajectories, NavWM maintains robustness by consistently reconstructing the target state.

4.4 Ablations

We conduct ablation studies on the combined RECON and HuRoN datasets to investigate the following questions:

- **Q1.** Do perception, action, and vision truly synergize in a unified framework?
- **Q2.** How effective is the anchor-based multimodal trajectory prediction?
- **Q3.** Does employing a world model as a generative planner effectively enhance trajectory quality?
- **Q4.** To what extent does increasing candidate count improve performance?

Unified Framework vs. Decoupled Modules. To answer **Q1**, we conduct module ablation study on the latent world reasoning, action prediction, and world model, with the quantitative results summarized in Tab. 3. To assess the raw multimodal policy, we compute metrics directly on probabilistically sampled trajectories without world model planning. As shown, integrating world reasoning and visual foresight drastically enhances trajectory quality (ATE 0.513 $\rightarrow$ 0.164 and RPE 0.187 $\rightarrow$ 0.113). Conversely, the inclusion of scene abstractions and action prediction respectively boosts the structural (PSNR 14.286 $\rightarrow$ 17.622) and perceptual (LPIPS 0.293 $\rightarrow$ 0.240) fidelity of the generated images. Furthermore, our training stage ablation demonstrates that two-stage fine-tuning yields significant improvements in overall image reconstruction quality (Fig. 5). These results validate the efficacy of our unified framework and training paradigm.

Multimodal Diversity and Feasibility. Addressing **Q2**, we compare our anchor-based multimodal trajectory prediction method against Gaussian Mixture Modeling (GMM) and the diffusion-based NoMaD. Additionally, we introduce Average Pairwise Distance (APD) to explicitly evaluate trajectory diversity. As shown in Tab. 4, under raw probabilistic sampling without the planner,

Table 4: Effectiveness of multimodal action and world model guidance. We compare our anchor-based trajectory generation approach against Gaussian Mixture Modeling (GMM) and the diffusion-based head under raw sampling (×) and world model planning (✓).

Methods	WM Planning	APD ↑	ATE ↓	RPE ↓	AOE ↓	MAOE ↓
GMM	×	-	0.88	0.26	12.03	17.54
	✓	0.31	0.79	0.23	11.62	16.84
Diffusion	×	-	0.86	0.25	11.79	17.18
	✓	0.59	0.58	0.18	10.56	14.50
Anchor-based (Ours)	×	-	0.85	0.24	11.33	16.86
	✓	**1.49**	**0.36**	**0.14**	**8.03**	**12.34**

NavWM achieves the best navigation performance while concurrently maintaining the highest APD. This proves that our method effectively mitigates the mode collapse inherent in prior baselines, simultaneously maximizing topological diversity and physical feasibility to significantly boost exploration capabilities without compromising accuracy.

Efficacy of World Model Planning. We compare trajectories generated via raw probabilistic sampling against those refined by the predictive world model planner, thereby answering **Q3**. As demonstrated in Tab. 4, supervision from the NavWM world model effectively enhances the overall navigation performance. Furthermore, we observe that these performance gains are significantly more pronounced for our anchor-based multimodal prediction compared to the standard GMM and diffusion-based NoMaD baselines. This proves that the trajectories generated by NavWM achieve superior environmental exploration, naturally providing a highly compatible action space for high-level planners. Consequently, this synergistic framework effectively mitigates the risk of getting trapped in local optima and ensures robust, foresighted decision-making in complex, long-horizon tasks.

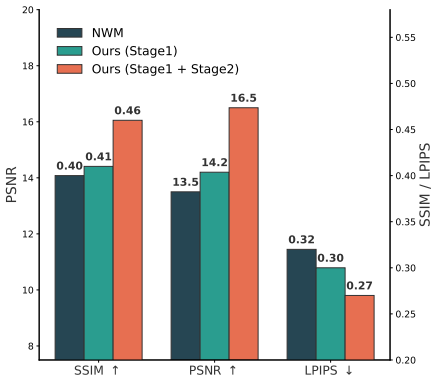

Fig. 5: Training stage ablation. Performing stage two fine-tuning significantly improves image reconstruction quality.

Impact of Candidate Trajectory Count. To address **Q4**, we ablated the number of candidate trajectories K to analyze its effect on navigation metrics

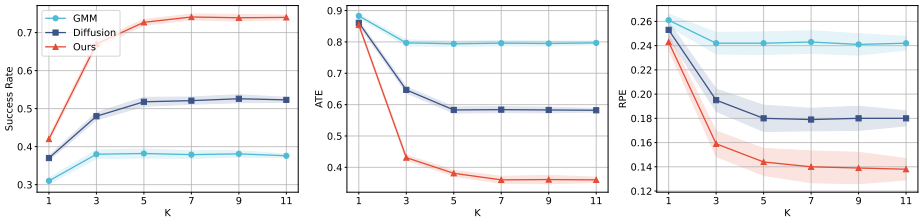

Fig. 6: Navigation performance vs. candidate trajectory count (K). NavWM shows sustained gains up to $K \approx 7$, outperforming Diffusion ($K \approx 5$) and GMM (earliest saturation), balancing diversity and accuracy.

under the condition that the world model acts as the planner. As shown in Fig. 6, navigation performance improves as K increases, which is most evident in NavWM. However, this improvement eventually saturates at a higher K. Specifically, the performance plateaus at approximately 7 for NavWM and 5 for the diffusion-based strategy, while the GMM reaches saturation at an even lower value. This implies that NavWM's multimodal predictions maintain high fidelity while offering greater diversity, facilitating more efficient environmental exploration and allowing the system to benefit from a larger set of candidates.

5 Conclusion

We present NavWM, a unified world model that overcomes the limitations of deterministic navigation policies by jointly optimizing latent scene reasoning, multimodal action prediction, and visual generation. By distilling geometric and semantic priors through Latent World Tokens, NavWM achieves robust structural understanding. Crucially, our multimodal trajectory framework generates a diverse action space that circumvents local minima and empowers the generative world model to perform optimal closed-loop planning via visual foresight. Extensive evaluations confirm that NavWM significantly advances the state-of-the-art in both high-fidelity future simulation and zero-shot navigation. While its generative capabilities in highly dynamic scenes with moving objects remain a current limitation, NavWM ultimately establishes a powerful new paradigm for unifying perception, generation, and control.

Acknowledgements

This research is supported by Artificial Intelligence-National Science and Technology Major Project (2023ZD0121200), the National Natural Science Foundation of China (62531026, 62437001, 62436001), the Natural Science Foundation of Jiangsu Province under Grant BK20243051 and the Strategic Priority Research Program of Chinese Academy of Sciences under Grant XDB1350103.

References

1. Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., et al.: Cosmos world foundation model platform for physical ai. arXiv preprint arXiv:2501.03575 (2025) 4
2. Alonso, E., Jelley, A., Micheli, V., Kanervisto, A., Storkey, A.J., Pearce, T., Fleuret, F.: Diffusion for world modeling: Visual details matter in atari. Advances in Neural Information Processing Systems **37**, 58757–58791 (2024) 8, 11
3. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15619–15629 (2023) 4
4. Bar, A., Zhou, G., Tran, D., Darrell, T., LeCun, Y.: Navigation world models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15791–15801 (2025) 1, 2, 4, 8, 10, 11, 12
5. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023) 5
6. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., et al.: Video generation models as world simulators. OpenAI Blog **1**(8), 1 (2024) 4
7. Bruce, J., Dennis, M.D., Edwards, A., Parker-Holder, J., Shi, Y., Hughes, E., Lai, M., Mavalankar, A., Steigerwald, R., Apps, C., et al.: Genie: Generative interactive environments. In: Forty-first International Conference on Machine Learning (2024) 4
8. Chai, Y., Sapp, B., Bansal, M., Anguelov, D.: Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. arXiv preprint arXiv:1910.05449 (2019) 7
9. Chaplot, D.S., Gandhi, D., Gupta, S., Gupta, A., Salakhutdinov, R.: Learning to explore using active neural slam. arXiv preprint arXiv:2004.05155 (2020) 4
10. Chen, J., Hu, J., Bai, H., Luo, M., Xue, X., Ren, B., Bai, C., Xie, S., Chen, Z., Liu, F., et al.: Astranav-world: World model for foresight control and consistency. arXiv preprint arXiv:2512.21714 (2025) 2, 4
11. Chern, E., Su, J., Ma, Y., Liu, P.: Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. arXiv preprint arXiv:2407.06135 (2024) 8, 11, 12
12. Chi, C., Xu, Z., Feng, S., Cousineau, E., Du, Y., Burchfiel, B., Tedrake, R., Song, S.: Diffusion policy: Visuomotor policy learning via action diffusion. The International Journal of Robotics Research **44**(10-11), 1684–1704 (2025) 2, 3
13. Cui, H., Radosavljevic, V., Chou, F.C., Lin, T.H., Nguyen, T., Huang, T.K., Schneider, J., Djuric, N.: Multimodal trajectory predictions for autonomous driving using deep convolutional networks. In: 2019 international conference on robotics and automation (icra). pp. 2090–2096. IEEE (2019) 7
14. Ding, H., Xu, Z., Fang, Y., Wu, Y., Chen, Z., Shi, J., Huo, J., Zhang, Y., Gao, Y.: Lavira: Language-vision-robot actions translation for zero-shot vision language navigation in continuous environments. arXiv preprint arXiv:2510.19655 (2025) 4
15. Dong, Y., Wu, F., Chen, G., Kong, L., Zhu, X., Hu, Q., Zhou, Y., Sun, J., He, J.Y., Dai, Q., Hauptmann, A.G., Cheng, Z.Q.: Towards unified world models for visual navigation via memory-augmented planning and foresight (2026), `https://arxiv.org/abs/2510.08713` 2, 3, 4, 8, 10, 11, 12

16. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems **27** (2014) 6
17. Feng, W., Liu, J., Tu, P., Qi, T., Sun, M., Ma, T., Zhao, S., Zhou, S., He, Q.: I2vcontrol-camera: Precise video camera control with adjustable motion strength. arXiv preprint arXiv:2411.06525 (2024) 4
18. Frey, J., Mattamala, M., Chebrolu, N., Cadena, C., Fallon, M., Hutter, M.: Fast traversability estimation for wild visual navigation. arXiv preprint arXiv:2305.08510 (2023) 4
19. Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., Isola, P.: Dreamsim: Learning new dimensions of human visual similarity using synthetic data. arXiv preprint arXiv:2306.09344 (2023) 9, 10
20. Fu, Z., Kumar, A., Agarwal, A., Qi, H., Malik, J., Pathak, D.: Coupling vision and proprioception for navigation of legged robots. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17273–17283 (2022) 4
21. Gervet, T., Chintala, S., Batra, D., Malik, J., Chaplot, D.S.: Navigating to objects in the real world. Science Robotics **8**(79), eadf6991 (2023) 2, 4
22. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023) 5
23. Gu, J., Sun, C., Zhao, H.: Densetnt: End-to-end trajectory prediction from dense goal sets. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 15303–15312 (2021) 2, 6
24. Ha, D., Schmidhuber, J.: World models. arXiv preprint arXiv:1803.10122 **2**(3) (2018) 4
25. He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., Yang, C.: Cameractrl: Enabling camera control for text-to-video generation. arXiv preprint arXiv:2404.02101 (2024) 4
26. He, H., Yang, C., Lin, S., Xu, Y., Wei, M., Gui, L., Zhao, Q., Wetzstein, G., Jiang, L., Li, H.: Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13416–13426 (2025) 4
27. Hirose, N., Sadeghian, A., Vázquez, M., Goebel, P., Savarese, S.: Gonet: A semi-supervised deep learning approach for traversability estimation. In: 2018 IEEE/RSJ international conference on intelligent robots and systems (IROS). pp. 3044–3051. IEEE (2018) 3, 9
28. Hirose, N., Shah, D., Sridhar, A., Levine, S.: Sacson: Scalable autonomous control for social navigation. IEEE Robotics and Automation Letters **9**(1), 49–56 (2023) 3, 9
29. Karnan, H., Nair, A., Xiao, X., Warnell, G., Pirk, S., Toshev, A., Hart, J., Biswas, J., Stone, P.: Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation. IEEE Robotics and Automation Letters **7**(4), 11807–11814 (2022) 3, 9
30. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7482–7491 (2018) 9
31. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023) 6

32. Lin, B., Nie, Y., Wei, Z., Chen, J., Ma, S., Han, J., Xu, H., Chang, X., Liang, X.: Navcot: Boosting llm-based vision-and-language navigation via learning disentangled reasoning. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025) 2
33. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022) 8
34. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022) 8
35. Liu, X., Li, J., Jiang, Y., Sujay, N., Yang, Z., Zhang, J., Abanes, J., Zhang, J., Feng, C.: Citywalker: Learning embodied urban navigation from web-scale videos. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6875–6885 (2025) 1, 2, 4, 10
36. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 10
37. Mangalam, K., Girase, H., Agarwal, S., Lee, K.H., Adeli, E., Malik, J., Gaidon, A.: It is not the journey but the destination: Endpoint conditioned trajectory prediction. In: European conference on computer vision. pp. 759–776. Springer (2020) 7
38. Mei, Y., Yang, Y., Guo, L., Wang, Q., Yu, M.M., He, X., Wu, W., Liu, J.: Urbannav: Learning language-guided embodied urban navigation from web-scale human trajectories. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 18505–18513 (2026) 2, 4
39. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023) 8
40. Shah, D., Eysenbach, B., Kahn, G., Rhinehart, N., Levine, S.: Rapid exploration for open-world navigation with latent goal models. arXiv preprint arXiv:2104.05859 (2021) 3, 9
41. Shah, D., Sridhar, A., Bhorkar, A., Hirose, N., Levine, S.: Gnm: A general navigation model to drive any robot. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 7226–7233. IEEE (2023) 1, 2, 3, 4, 8
42. Shah, D., Sridhar, A., Dashora, N., Stachowicz, K., Black, K., Hirose, N., Levine, S.: Vint: A foundation model for visual navigation. arXiv preprint arXiv:2306.14846 (2023) 1, 2, 3, 4, 8, 11, 12
43. Sitzmann, V., Rezchikov, S., Freeman, B., Tenenbaum, J., Durand, F.: Light field networks: Neural scene representations with single-evaluation rendering. Advances in Neural Information Processing Systems **34**, 19313–19325 (2021) 8
44. Song, M., Deng, X., Zhou, Z., Wei, J., Guan, W., Nie, L.: A survey on diffusion policy for robotic manipulation: Taxonomy, analysis, and future directions. Authorea Preprints (2025) 2
45. Sridhar, A., Shah, D., Glossop, C., Levine, S.: Nomad: Goal masked diffusion policies for navigation and exploration. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 63–70. IEEE (2024) 1, 4, 8, 10, 11, 12
46. Sturm, J., Burgard, W., Cremers, D.: Evaluating egomotion and structure-from-motion approaches using the tum rgb-d benchmark. In: Proc. of the Workshop on Color-Depth Camera Fusion in Robotics at the IEEE/RSJ International Conference on Intelligent Robot Systems (IROS). vol. 13, p. 6 (2012) 10
47. Triest, S., Sivaprakasam, M., Wang, S.J., Wang, W., Johnson, A.M., Scherer, S.: Tartandrive: A large-scale dataset for learning off-road dynamics models. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 2546–2552. IEEE (2022) 3, 10

48. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004) 10
49. Xu, J., Sun, X., Zhang, Z., Zhao, G., Lin, J.: Understanding and improving layer normalization. Advances in neural information processing systems **32** (2019) 8
50. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. Advances in Neural Information Processing Systems **37**, 21875–21911 (2024) 6
51. Yao, X., Gao, J., Xu, C.: Navmorph: A self-evolving world model for vision-and-language navigation in continuous environments. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5536–5546 (2025) 4
52. Yin, H., Xu, X., Zhao, L., Wang, Z., Zhou, J., Lu, J.: Unigoal: Towards universal zero-shot goal-oriented navigation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19057–19066 (2025) 4
53. Yu, M., Zhu, F., Liu, W., Yang, Y., Wang, Q., Liu, J., et al.: C-nav: Towards self-evolving continual object navigation in open world. Advances in Neural Information Processing Systems **38**, 126181–126207 (2026) 4
54. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018) 9, 10
55. Zhao, H., Gao, J., Lan, T., Sun, C., Sapp, B., Varadarajan, B., Shen, Y., Shen, Y., Chai, Y., Schmid, C., et al.: Tnt: Target-driven trajectory prediction. In: Conference on robot learning. pp. 895–904. PMLR (2021) 2, 7